PRACE KÓŁ NAUKOWYCH Politechniki Rzeszowskiej w roku akademickim 2023/2024

Rzeszów 2024

Wydano za zgodą Rektora

R e d a k t o r n a c z e l n y
Wydawnictw Politechniki Rzeszowskiej
dr hab. inż. Lesław GNIEWEK, prof. PRz

Przewodniczący Rady Redakcyjnej
dr inż. Bartosz TRYBUS

Rada Redakcyjna
Prac Kół Naukowych 2023/2024
prof. dr hab. inż. Wiktor BUKOWSKI
dr Anna OSTROWSKA-DANKIEWICZ
dr inż. Dorota GŁOWACZ-CZERWONKA, prof. PRz
mgr. inż. Dawid KALANDYK, dr Agnieszka LEW
dr hab. inż. Grzegorz LEW, prof. PRz
mgr inż. Patryk ORGANIŚCIAK, dr inż. Bartosz PAWŁOWICZ
dr inż. Bartosz TRYBUS, mgr inż. Michał WANIC
mgr inż. Wiktoria WOJNAROWSKA

Wydruk z matryc dostarczonych przez Radę Redakcyjną Prac Kół Naukowych.
W procesie wydawniczym pominięto etap opracowania językowego.

segmentacja obrazu, sztuczna inteligencja, prognozowanie, bazy danych, broker ubezpieczeniowy, broker reasekuracyjny, włókna aramidowe, kompozyty, pianki poliuretanowe, uniepalniacze, antypireny,rolnictwo precyzyjne, technologie bezprzewodowe, drukarka 3D, blockchain, metody amortyzacji, konwolucja, CNN, sieć neuronowa, Next.js, React, framework, server components, client components, routing, kubernetes, web scraping, gradient błędu, bezpieczeństwo, cyberbezpieczeństwo, aplikacje chmurowe

p-ISBN 978-83-7934-743-8
e-ISBN 978-83-7934-744-5

Oficyna Wydawnicza Politechniki Rzeszowskiej
al. Powstańców Warszawy 12, 35-959 Rzeszów
https://oficyna.prz.edu.pl

Ark. wyd. 37,09. Ark. druk. 41,50. Wydrukowano we wrześniu 2024 r.
Drukarnia Oficyny Wydawniczej PRz, al. Powstańców Warszawy 12, 35-959 Rzeszów
Zam. nr 41/24

SPIS TREŚCI

KOŁO NAUKOWE SYSTEMÓW ZŁOŻONYCH

KOŁO NAUKOWE UBEZPIECZEŃ

KOŁO NAUKOWE STUDENTÓW CHEMII "ESPRIT"

KOŁO NAUKOWE INŻYNIERII MEDYCZNEJ X-MED

KOŁO NAUKOWE ELEKTRONIKI I TECHNOLOGII INFORMACYJNYCH

KOŁO NAUKOWE RACHUNKOWOŚCI „ASSETS”

KOŁO NAUKOWE INTERAKCJI CZŁOWIEK – KOMPUTER „GEST”

KOŁO NAUKOWE INFORMATYKÓW „KOD”

KOŁO

NAUKOWE

SYSTEMÓW

ZŁOŻONYCH

Jakub Kuźniar, Michał Kocik, Veronika Vanivska, Aldona Świrad
Koło Naukowe Systemów Złożonych

mgr inż. Patryk Organiściak
Opiekun Koła Naukowego

Asynchroniczny system segmentacji raka piersi

Streszczenie

Prezentowane badanie opisuje metodę tworzenia asynchronicznego systemu segmentacji raka piersi z użyciem biblioteki TiaToolbox. System analizuje obrazy w celu wykrywania raka piersi, wykorzystując TiaToolbox w Pythonie oraz FastAPI, co ułatwia wdrożenie do różnych projektów. Celery zapewnia spójność działania, umożliwiając kolejkowanie zadań bez blokowania głównego wątku aplikacji. Utworzono zapytania GET do kontrolowania stanu analizy. System dostępny jest jako dwa obrazy Docker'a, co upraszcza instalację. FastAPI zapewnia szybkie przetwarzanie zapytań, a Celery efektywne zarządzanie zadaniami, co zwiększa wydajność i niezawodność. Integracja z TiaToolbox zwiększa dokładność wykrywania raka piersi, a system można rozszerzać o dodatkowe moduły analityczne. Jest to wszechstronne narzędzie wspomagające diagnozowanie raka piersi, łatwe do wdrożenia w różnych środowiskach medycznych.

Słowa kluczowe: segmentacja obrazu, sztuczna inteligencja, uczenie maszynowe

1. Wprowadzenie

W ostatnich latach technologia medyczna znacząco ewoluowała, wprowadzając innowacyjne rozwiązania wspierające diagnozowanie i leczenie różnych schorzeń. Jednym z obszarów, który zyskał szczególną uwagę, jest diagnostyka raka piersi. Wczesne wykrywanie raka piersi jest kluczowe dla skutecznego leczenia i zwiększenia szans na przeżycie pacjentek. Tradycyjne metody diagnostyczne, takie jak mammografia, choć skuteczne, mają swoje ograniczenia, dlatego coraz większy nacisk kładzie się na rozwój zaawansowanych systemów wspomagających diagnozę.[1]

Celem niniejszego badania jest opracowanie asynchronicznego systemu segmentacji raka piersi, który pozwala na analizę obrazów medycznych w celu wykrywania zmian nowotworowych. System ten ma na celu zwiększenie dokładności i efektywności procesu diagnozowania raka piersi, wspierając pracę użytkowników w zakresie interpretacji wyników. Dzięki zastosowaniu zaawansowanych narzędzi do przetwarzania obrazów, system ten ma potencjał, aby stać się integralną częścią nowoczesnych metod diagnostycznych.

[1] J. V. Vardhan, G. S. Krishna, *Breast cancer segmentation using attention-based convolutional network and explainable AI*, arXiv.org. https://doi.org/10.48550/arXiv.2305.14389, 2023.

Obiektem badań jest system segmentacji raka piersi oparty na bibliotece TiaToolbox[2], która dostarcza narzędzi do zaawansowanego przetwarzania obrazów. System ten został zaprojektowany z myślą o łatwej integracji z istniejącymi rozwiązaniami oraz możliwością dostosowania do specyficznych wymagań różnych użytkowników. Wykorzystanie języka Python oraz frameworka FastAPI umożliwia szybkie i efektywne przetwarzanie danych, co jest kluczowe w kontekście analizy dużych ilości obrazów medycznych.

Metody badawcze obejmują zastosowanie asynchronicznego przetwarzania z użyciem biblioteki Celery, która umożliwia kolejkowanie zadań bez blokowania głównego wątku aplikacji. Celery[3] jest biblioteką Python, która jest szeroko stosowana do zarządzania zadaniami i przetwarzania asynchronicznego, co pozwala na efektywne wykorzystanie zasobów systemowych i zwiększenie wydajności całego procesu. Dzięki Celery, system może przetwarzać wiele zapytań jednocześnie, co jest szczególnie ważne w kontekście dużych placówek medycznych, gdzie codziennie przetwarzane są setki obrazów.

Dodatkowo, system został zaimplementowany jako dwa obrazy Docker'a, co znacząco upraszcza proces instalacji i konfiguracji. Docker jest narzędziem, które umożliwia tworzenie, uruchamianie i zarządzanie kontenerami, co zapewnia spójność środowiska uruchomieniowego i łatwość przenoszenia aplikacji między różnymi systemami. Dzięki Docker'owi, system może być szybko wdrożony na różnych platformach, minimalizując problemy związane z kompatybilnością i konfiguracją.

System wykorzystuje również zapytania GET do kontrolowania stanu analizy, co pozwala na bieżące monitorowanie procesu przez użytkowników. Dzięki temu użytkownicy mogą śledzić postęp analizy obrazów w czasie rzeczywistym, co zwiększa transparentność i kontrolę nad procesem diagnostycznym. Zapytania GET są standardowym elementem protokołu HTTP, co ułatwia integrację systemu z innymi aplikacjami i usługami.

Integracja z biblioteką TiaToolbox zapewnia dostęp do zaawansowanych algorytmów i narzędzi do przetwarzania obrazów, co zwiększa dokładność i efektywność systemu w wykrywaniu raka piersi. TiaToolbox oferuje szeroki zakres funkcji, które mogą być dostosowane do specyficznych potrzeb użytkowników, co pozwala na tworzenie bardziej precyzyjnych i

---

[2] C. Chen, M. Y. Lu, D. F. K. Williamson, T. Y. Chen, A. J. Schaumberg, F Mahmood, *TIAToolbox: An End-to-End Toolbox for Advanced Tissue Image Analytics,* https://www.researchgate.net/publication/357322919_TIAToolbox_An_End-to-End_Tooalbox_for_Advanced_Tissue_Image_Analytics, 2021.

[3] J. Doe, A. Smith, *Revolutionizing Concurrent Crawling: A Novel Approach to Enhance PHP-Python Integration using AMQP, Selenium, Celery, and RabbitMQ,* https://www.researchgate.net/publication/377684280_Revolutionizing_Concurrent_Crawling_A_Novel_Approach_to_Enhance_PHP-Python_Integration_using_AMQP_Selenium_Celery_and_RabbitMQ, 2023.

spersonalizowanych rozwiązań diagnostycznych. Dzięki temu system może być wykorzystywany w różnych kontekstach, od małych klinik po duże szpitale.

2. Opis systemu segmentacji raka piersi

System segmentacji raka piersi opracowany w ramach tego badania składa się z kilku kluczowych komponentów, które współpracują ze sobą w celu zapewnienia dokładnej i efektywnej analizy obrazów medycznych. Poniżej omówiono szczegółowo główne elementy systemu, przybliżając ich teoretyczne podstawy oraz praktyczne zastosowanie.

2.1. Biblioteka TiaToolbox

Głównym komponentem systemu jest biblioteka TiaToolbox, która dostarcza zaawansowane narzędzia do przetwarzania i analizy obrazów. TiaToolbox oferuje szeroki wachlarz funkcji, które umożliwiają zarówno podstawowe operacje na obrazach, jak i zaawansowane techniki segmentacji. Segmentacja obrazu jest kluczowym procesem w diagnostyce raka piersi, polegającym na wydzieleniu z obrazu obszarów podejrzanych o obecność zmian nowotworowych.

Biblioteka TiaToolbox została zaprojektowana specjalnie do pracy z dużymi obrazami mikroskopowymi, takimi jak pliki WSI[4] (Whole Slide Images). Pliki WSI są skanowanymi obrazami całych szkiełek mikroskopowych, które zawierają szczegółowe informacje na temat tkanki. Ze względu na swoją wysoką rozdzielczość i dużą ilość danych, obrazy te wymagają zaawansowanych narzędzi do przetwarzania i analizy, aby można było efektywnie identyfikować i segmentować zmiany nowotworowe. Kluczowym elementem TiaToolbox jest klasa WSIReader, która definiuje funkcje do odczytu danych pikselowych i metadanych z plików WSI.

Biblioteka TiaToolbox została zaprojektowana specjalnie do pracy z dużymi obrazami mikroskopowymi, takimi jak pliki WSI (Whole Slide Images). Pliki WSI są skanowanymi obrazami całych szkiełek mikroskopowych, które zawierają szczegółowe informacje na temat tkanki. Ze względu na swoją wysoką rozdzielczość i dużą ilość danych, obrazy te wymagają zaawansowanych narzędzi do przetwarzania i analizy, aby można było efektywnie identyfikować i segmentować zmiany nowotworowe. Kluczowym elementem TiaToolbox jest klasa WSIReader[5], która definiuje funkcje do odczytu danych pikselowych i metadanych z plików WSI.

---

[4] M. Khened, A. Kori, H. Rajkumar, G. Krishnamurthi, B. Srinivasan, *A generalized deep learning framework for whole-slide image segmentation and analysis*, Scientific Reports, 11, 90444. https://doi.org/10.1038/S41598-021-90444-8, 2021.

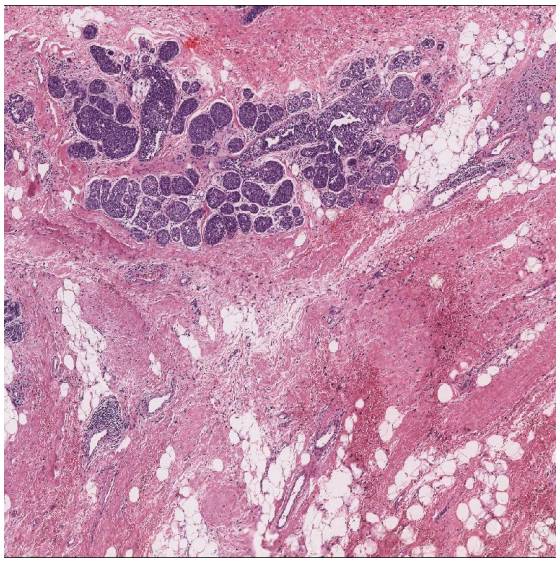
[5] https://tia-toolbox.readthedocs.io/en/latest/_autosummary/tiatoolbox.wsicore.wsireader.WSIReader.html.

Klasa `WSIReader` oferuje szereg funkcji, które są niezbędne do pracy z obrazami WSI:

- **WSIReader**: Jest to podstawowa klasa do pracy z obrazami WSI. Umożliwia ona ładowanie obrazów z różnych źródeł, takich jak ścieżki plików, tablice numpy, czy już istniejące obiekty `WSIReader`. Ta elastyczność jest kluczowa dla integracji z różnymi systemami i źródłami danych. `WSIReader` obsługuje różne formaty obrazów i zapewnia jednolity interfejs do pracy z nimi,
- **slide_thumbnail**: Ta funkcja generuje miniatury całych obrazów WSI, co ułatwia szybki podgląd i nawigację po obrazach o dużej rozdzielczości. Miniatury są szczególnie użyteczne do szybkiej oceny zawartości obrazu i identyfikacji obszarów zainteresowania przed przeprowadzeniem bardziej szczegółowej analizy. Funkcja przyjmuje parametry takie jak rozdzielczość (`resolution`) i jednostki (`units`), co pozwala na dostosowanie wielkości miniatury,
- **read_rect**: Funkcja ta pozwala na odczyt wybranego regionu obrazu na podstawie jego lokalizacji i rozmiaru. Jest to niezwykle przydatne w przypadku analizy specyficznych obszarów tkanki, umożliwiając precyzyjne wyodrębnienie i analizę małych fragmentów obrazu bez konieczności wczytywania całego pliku. Parametry tej funkcji obejmują lokalizację (`location`), rozmiar (`size`), rozdzielczość (`resolution`) oraz jednostki (`units`). Odczytany region obrazu jest zwracany jako tablica numpy, co umożliwia łatwą integrację z innymi narzędziami do analizy danych.

Klasa WSIReader pozwala również na dostęp do metadanych obrazu, które zawierają istotne informacje o całym slajdzie. Poniżej przedstawiono przykładowe metadane, które można uzyskać za pomocą funkcji info.as_dict():

Na przykładzie:



Rys. 1 Przykładowy obraz SVS
Źródło: Dokumentacja TiaToolbox [1]

Listing 1. Przykładowy opis parametrów obrazu SVS zwracany przez API TiaToolbox

```
{'axes': 'YXS',
 'file_path': PosixPath('svs_example.svs'),
 'level_count': 2,
 'level_dimensions': ((12000, 12000), (3000, 3000)),
 'level_downsamples': [1.0, 4.0],
 'mpp': (0.2505, 0.2505),
 'objective_power': 40.0,
 'slide_dimensions': (12000, 12000),
 'vendor': 'aperio'
}
```

Opis poszczególnych parametrów obrazu:

- axes: Określa układ osi obrazu. 'YXS' oznacza, że oś Y jest pierwsza, oś X jest druga, a oś S (kolory) jest trzecia. Jest to ważne dla zrozumienia, w jaki sposób dane obrazowe są zorganizowane w macierzy,
- file_path: Ścieżka do pliku WSI. W tym przykładzie jest to 'svs_example.svs'. Informacja ta jest kluczowa dla identyfikacji i odnalezienia konkretnego pliku obrazu na dysku,
- level_count: Liczba poziomów rozdzielczości dostępnych w obrazie WSI. W tym przypadku wartość to 2, co oznacza, że obraz ma dwa poziomy rozdzielczości. Każdy poziom może reprezentować obraz w innej rozdzielczości, co jest użyteczne dla różnych potrzeb analitycznych,
- level_dimensions: Wymiary obrazu na każdym poziomie rozdzielczości. Tutaj podano dwie wartości: ((12000, 12000), (3000, 3000)). Pierwsza para (12000, 12000) odnosi się do pełnej rozdzielczości obrazu (level 0), a druga (3000, 3000) do zredukowanej rozdzielczości (level 1),

- level_downsamples: Współczynniki zmniejszenia rozdzielczości dla każdego poziomu. Tutaj wartości to [1.0, 4.0], co oznacza, że pierwszy poziom ma pełną rozdzielczość, a drugi jest zmniejszony czterokrotnie. Ten parametr jest istotny, gdy potrzebujemy przetwarzać obraz w różnych skalach,
- mpp: Rozdzielczość w mikrometrach na piksel. W tym przykładzie wynosi (0.2505, 0.2505), co oznacza, że jeden piksel obrazu odpowiada 0.2505 mikrometra w rzeczywistości. Ta informacja jest kluczowa dla dokładnych pomiarów morfologicznych w analizach obrazów mikroskopowych,
- objective_power: Moc obiektywu używanego do skanowania obrazu. W tym przypadku wartość wynosi 40.0, co oznacza, że obraz został zeskanowany przy użyciu obiektywu o mocy 40-krotnego powiększenia. Wysoka moc obiektywu pozwala na uzyskanie szczegółowych obrazów wysokiej jakości,
- slide_dimensions: Wymiary całego slajdu w pikselach. Tutaj wynosi (12000, 12000), co wskazuje na wielkość obrazu na najwyższym poziomie rozdzielczości. Jest to przydatne dla zrozumienia pełnego zakresu obrazu, który może być bardzo duży w przypadku całych szkiełek mikroskopowych,
- vendor: Producent skanera, który wygenerował obraz. W tym przypadku jest to 'aperio'. Informacja o producencie może być istotna dla zrozumienia specyfiki formatu obrazu i kompatybilności z różnymi narzędziami do analizy.

2.2. Silnik segmentacji semantycznej w TiaToolbox

Silnik segmentacji semantycznej w TiaToolbox jest kluczowym komponentem umożliwiającym zaawansowaną analizę obrazów histologicznych. Segmentacja semantyczna polega na podziale obrazu na segmenty, gdzie każdy segment odpowiada określonej kategorii obiektu obecnego na obrazie. W przypadku obrazów histologicznych segmentacja semantyczna identyfikuje klasę każdego piksela na podstawie typu tkanki lub regionu, w którym się znajduje. Jest to niezwykle istotne w rozwoju algorytmów do diagnozowania i prognozowania raka, ponieważ umożliwia obiektywne i powtarzalne pomiary właściwości tkanek.

TiaToolbox udostępnia wcześniej wytrenowany model do segmentacji raka piersi, który w prosty sposób można zaimplementować oraz przykładowe obrazy (rys. 1). Model ten, w pełni konwolucyjny VGG-16, sieć FCN-8 (Fully Convolutional Network)[6], został wytrenowany do

[6] S. A. Kamran, A. S. Sabbir, *Efficient Yet Deep Convolutional Neural Networks for Semantic Segmentation*, https://www.researchgate.net/publication/318720999_Efficient_Yet_Deep_Convolutional_Neural_Networks_for_Semantic_Segmentation, 2017.

segmentacji obrazów histologicznych na pięć klas: guz, stroma, infiltraty zapalne, martwica i inne klasy[7]. Aby poprawić odporność modelu, zastosowano augmentację danych poprzez przesunięcia i przycinanie — szczegóły znajdują się w Metodach uzupełniających.

Silnik segmentacji semantycznej w TiaToolbox przewiduje wszystkie pięć klas, jednak dla uproszczenia wizualizacji, tylko jedna klasa – guz – jest wyświetlana. Klasa guz została oznaczona kolorem zielonym, co pozwala na łatwiejszą identyfikację i wizualizację wyników analizy.

Model ten był trenowany na 125 regionach zainteresowania (ROIs) z naciekającymi rakami przewodowymi (większość TNBC - Triple-Negative Breast Cancer[8]), stosując normalizację kolorów do obrazów RGB (Reinhard et al., 2001). Różne typy modeli zostały przeszkolone, aby ocenić różne aspekty crowdsourcingu[9].

Wykorzystując wstępnie wytrenowany model TiaToolbox, proces segmentacji semantycznej obrazów histologicznych[10] staje się prostszy i bardziej zautomatyzowany, co pozwala na efektywną analizę dużych zbiorów danych w badaniach nad rakiem.

W skrócie, segmentacja semantyczna w TiaToolbox umożliwia:

- precyzyjną klasyfikację pikseli: Każdy piksel obrazu jest klasyfikowany do jednej z pięciu kategorii, co pozwala na dokładną analizę struktury tkanki,
- wizualizację zmian nowotworowych: Dzięki wizualizacji tylko klasy guza, można łatwo zidentyfikować obszary podejrzane o obecność nowotworu,
- automatyzację procesu: Automatyzacja analizy pozwala na szybsze i bardziej spójne wyniki, co jest kluczowe w diagnostyce i badaniach nad rakiem piersi.

Poniżej przedstawiono przykłady anotacji z QuPath oraz wyniki analizy przeprowadzone przez system. Program QuPath umożliwia ręczne oznaczanie obszarów zainteresowania na obrazach histologicznych, które można następnie wyeksportować w formacie GeoJSON[11].

---

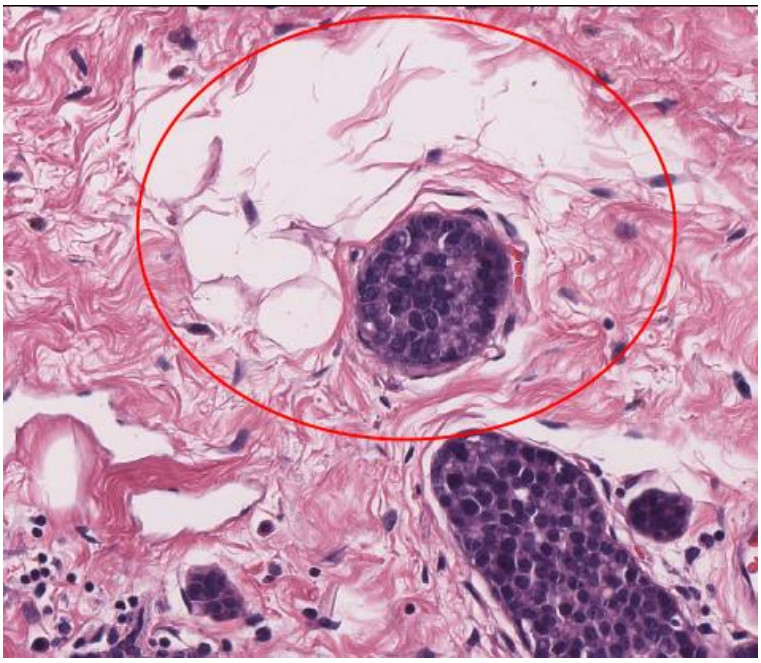
[7] J. Long, E. Shelhamer, T. Darrell, *Fully convolutional networks for semantic segmentation*, 10.1109/CVPR.2015.7298965. https://arxiv.org/pdf/1411.4038, 2015, s. 3431-3440.

[8] G. Bianchini, J. M. Balko, I. A. Mayer, M. E. Sanders, L. Gianni, *Triple-negative breast cancer: A distinct subtype of breast cancer with unique challenges*, Nature Reviews Clinical Oncology,. https://doi.org/10.1038/nrclinonc.2016.79, 13(11), 2016, s. 674-690

[9] A. Ghezzi, D. Gabelloni, A. Martini, A. Natalicchio, *Crowdsourcing: A Review and Suggestions for Future Research*, International Journal of Management Reviews, https://doi.org/10.1111/ijmr.12135, 20(2), 2017, s. 2-29.

[10] A. Seth, S. Sharma, *Semantic Segmentation: A Systematic Analysis From State-of-the-Art Techniques to Advance Deep Networks*, Journal of Information Technology Research, https://doi.org/10.4018/JITR.299388, 15(1), 2022, s.1-28.

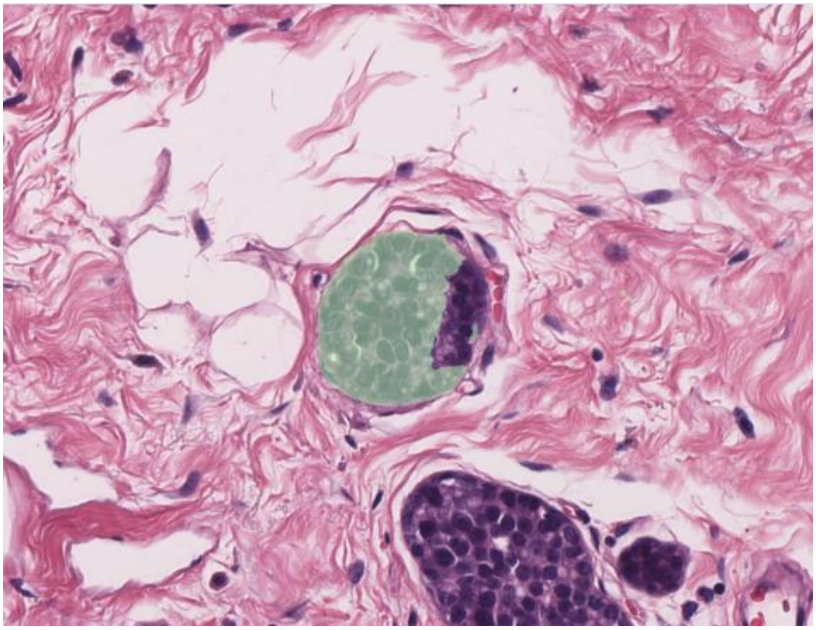
[11] https://geojson.org /



Rys. 2 Anotacja wykonana w programie QuPath
Źródło: Opracowanie własne

Listing 2. Geojson adnotacji z Qupath


```
{
  "type": "FeatureCollection",
  "features": [
    {
      "type": "Feature",
      "id": "26cada93-5f19-4709-aa2b-d445a78dfb2c",
      "geometry": {
        "type": "Polygon",
        "coordinates": [
          [
            [300, 400],
            [1000, 2500],
            [1200, 1835],
            [1045, 2264],
            [1394, 1304]
          ]
        ]
      },
      "properties": {
        "objectType": "annotation"
      }
    }
  ]
}
```



Rys. 3 Zaznaczony na zielono guz
Źródło: Opracowanie własne

Na zrzutach ekranu powyżej zostały przedstawione anotacja obszaru, która została poddana segmentacji, na zielono widać zaznaczony guz. W ten sposób, TiaToolbox dostarcza potężne narzędzie do analizy histologicznych obrazów mikroskopowych, wspierając badania i praktykę kliniczną w diagnostyce raka piersi.

2.3. REST API

REST API (Representational State Transfer Application Programming Interface[12]) jest jednym z kluczowych komponentów systemu segmentacji raka piersi, umożliwiającym komunikację między użytkownikami a serwerem. Dzięki REST API użytkownicy mogą przesyłać obrazy do analizy, monitorować status analiz oraz pobierać wyniki w sposób prosty i zautomatyzowany. W naszym systemie do implementacji REST API wykorzystano framework FastAPI, który jest znany ze swojej wysokiej wydajności i prostoty użycia.

REST API pozwala na elastyczną interakcję z systemem, umożliwiając łatwe przesyłanie danych i odbieranie wyników. Dzięki zastosowaniu standardowych metod HTTP, takich jak POST, GET, PUT i DELETE, można w sposób intuicyjny i skuteczny zarządzać zadaniami analizy obrazów. Kluczowe operacje REST API w naszym systemie obejmują:

- **POST /analyzeWSI**: Umożliwia użytkownikom przesyłanie obrazów do analizy. Po odebraniu żądania, serwer tworzy nowe zadanie analizy i zwraca unikalny identyfikator,

[12] G. Sharma, G. Lavania, D. Goyal, *Rest API: Data retrieval and applications*, AIP Conference Proceedings, https://doi.org/10.1063/5.0155589, 2023, s.2782.

który może być użyty do śledzenia postępu analizy. Użytkownik przesyła obraz, typ analizy oraz dodatkowe parametry, które są następnie przetwarzane przez system.

Listing 3. Przykład zapytania POST /analyzeWSI

```
{
    "svs_path": "/DATA/svs_example.svs",
    "analysis_type": 1,
    "analysis_parameters": {
      "analysis_region_json": "/DATA/2.json",
      "is_normalized": false
    }
}
```

Listing 4. Przykładowa odpowiedź

```
{
    "analysis_id": "3bc182f6-161f-45e4-8053-6bbef9e6cfb4"
}
```

- **GET /checkStatus/{analysis_id}**: Pozwala użytkownikom sprawdzić aktualny status analizy na podstawie jej identyfikatora. Statusy mogą obejmować: w kolejce, w trakcie przetwarzania, zakończone sukcesem lub zakończone niepowodzeniem. Dzięki temu użytkownicy mogą na bieżąco monitorować postęp swoich zadań.

Listing 5. Przykładowe zapytanie GET /checkStatus/{analysis_id}:

```
GET /checkStatus/3bc182f6-161f-45e4-8053-6bbef9e6cfb4
```

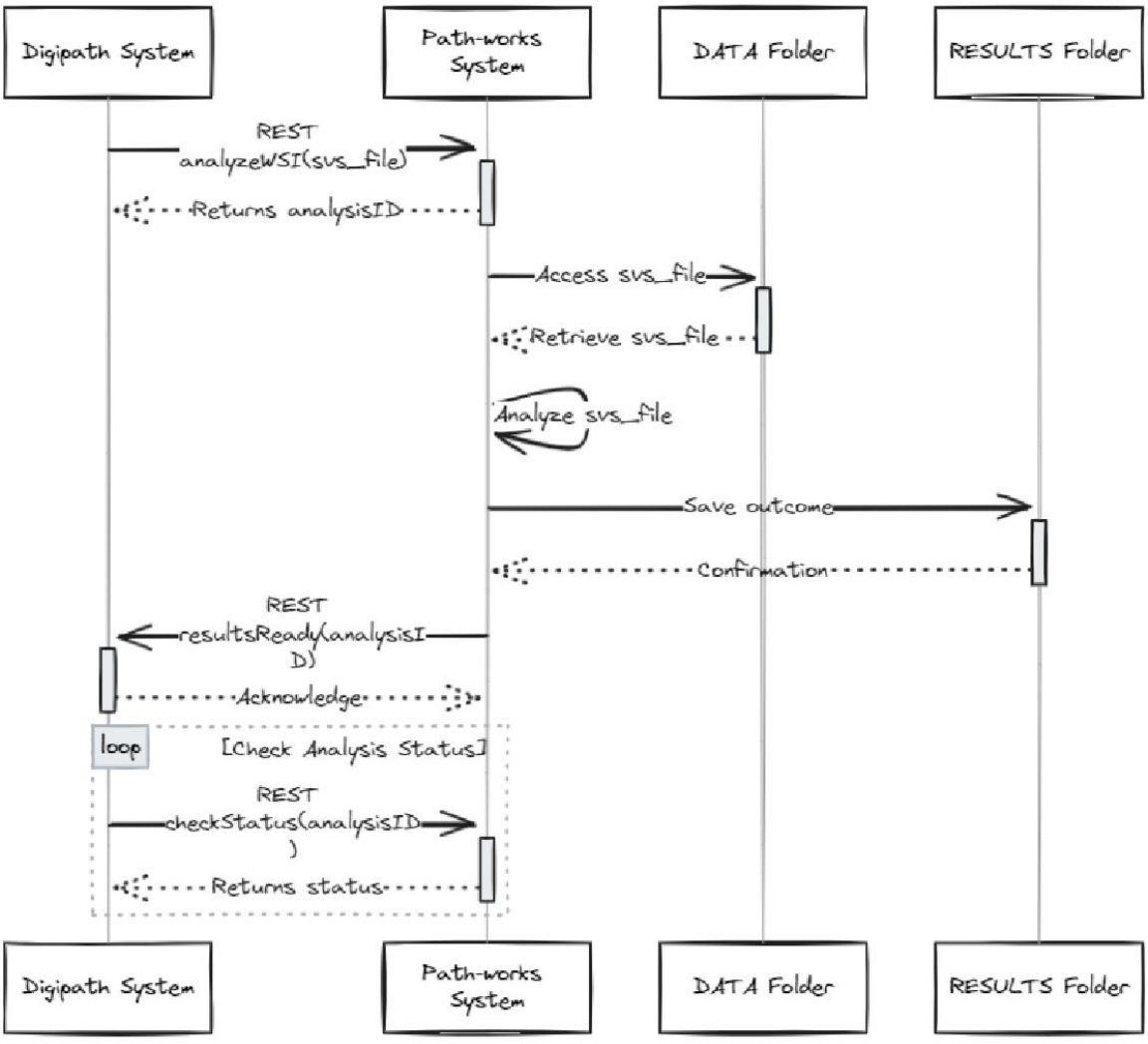
Listing 6. Przykładowa odpowiedź

```
{
    "analysis_id": "3b4312f1-806a-4105-a69f-8bd1f9f6f45b",
    "svs_path": "/DATA/svs_example.svs",
    "analysis_type": 1,
    "region_json_path": "/DATA/2.geojson",
    "is_normalized": false,
    "status": "finished",
    "result_json_path": "/RESULTS/3b4312f1-806a-4105-a69f-8bd1f9f6f45b/3b4312f1-
806a-4105-a69f-8bd1f9f6f45b.json",
    "prediction_time": 105.86610126495361,
    "overlay_time": 14.791182279586792,
    "result_image_path": "/RESULTS/3b4312f1-806a-4105-a69f-
8bd1f9f6f45b/result.tif"
}
```

- **Callback** to kolejna ważna funkcjonalność naszego systemu, umożliwiająca informowanie użytkowników o zakończeniu analizy. Po zakończeniu analizy system automatycznie wywołuje zdefiniowany endpoint, przesyłając dane o zakończonej analizie

FastAPI umożliwia łatwe definiowanie punktów końcowych i obsługę różnych typów żądań HTTP. Dzięki wbudowanej walidacji danych i automatycznemu generowaniu dokumentacji API, integracja z systemami zewnętrznymi jest szybka i bezproblemowa.

Poniżej przedstawiono przepływ danych i interakcje między różnymi komponentami systemu, w tym Digipath System, Path-works System, DATA Folder, oraz RESULTS Folder. Proces rozpoczyna się od przesłania żądania analizy obrazu, po czym Path-works System przetwarza obraz i zapisuje wyniki w odpowiednim folderze.



Rys. 4 Sekwencja komunikacji w asynchronicznym systemie segmentacji raka piersi
Źródło: opracowanie własne

Taki diagram pomaga lepiej zrozumieć operacyjny przepływ danych w systemie, ilustrując kolejne etapy przetwarzania i interakcji między komponentami.

2.4. Celery

Celery to potężna biblioteka do obsługi asynchronicznych kolejek zadań, która jest używana w naszym systemie do zarządzania zadaniami analizy obrazów. Celery umożliwia wykonywanie zadań w tle, co pozwala na efektywne przetwarzanie dużych ilości danych bez blokowania głównego wątku aplikacji.

Celery działa poprzez dystrybucję zadań do workerów, które wykonują je asynchronicznie. Dzięki temu, nawet przy dużej liczbie równoczesnych żądań, system pozostaje responsywny. Kluczowe komponenty konfiguracji Celery obejmują:

Celery App: Główna aplikacja Celery, skonfigurowana do korzystania z brokera wiadomości (np. Redis) do przechowywania zadań i wyników. Konfiguracja aplikacji obejmuje ustawienia takie jak serializacja zadań, akceptowalne formaty danych oraz ustawienia dotyczące strefy czasowej.

Zadania Celery: Definiowane funkcje, które są wykonywane asynchronicznie. W naszym systemie kluczowe zadanie to perform_analysis, które wykonuje analizę obrazu, generuje maski, przeprowadza predykcję oraz nakłada wyniki na obraz.

Celery jest niezwykle elastyczny i skalowalny, co umożliwia łatwe dostosowywanie systemu do rosnących potrzeb. Dzięki Celery, system może przetwarzać zadania w tle, co znacząco zwiększa jego wydajność i pozwala na równoczesne przetwarzanie wielu zadań.

Przykładowe zastosowanie Celery w analizie obrazów:

- Enqueue Task: Kiedy użytkownik przesyła żądanie analizy poprzez REST API, FastAPI tworzy zadanie Celery i dodaje je do kolejki.
- Process Task: Worker Celery odbiera zadanie z kolejki i rozpoczyna jego przetwarzanie. Zadanie może obejmować kroki takie jak wczytanie obrazu, segmentacja, analiza oraz generowanie wyników.
- Store Results: Po zakończeniu analizy, wyniki są zapisywane w określonym miejscu, a użytkownik jest powiadamiany o zakończeniu zadania.

Integracja FastAPI z Celery umożliwia użytkownikom przesyłanie zadań do wykonania oraz monitorowanie ich stanu za pomocą intuicyjnego interfejsu API. Po przesłaniu żądania analizy przez REST API, FastAPI tworzy zadanie Celery i zwraca użytkownikowi identyfikator analizy. Celery zarządza wykonaniem zadania w tle, a użytkownik może sprawdzać stan analizy i pobierać wyniki po jej zakończeniu.

W ten sposób, połączenie REST API i Celery zapewnia efektywne i skalowalne rozwiązanie do analizy obrazów histologicznych, wspierając zarówno badania naukowe, jak i praktykę

kliniczną w diagnostyce raka piersi. Dzięki zastosowaniu Celery, system jest w stanie obsługiwać dużą liczbę równoczesnych zadań, co znacznie zwiększa jego wydajność i elastyczność.

2.5. Docker i Połączenie Całego Systemu

Docker jest kluczowym narzędziem w naszym systemie segmentacji raka piersi, umożliwiającym wygodne wdrożenie, zarządzanie i skalowanie aplikacji w różnych środowiskach. Docker pozwala na pakowanie aplikacji i jej zależności w kontenery, co zapewnia spójność działania na różnych platformach. W naszym systemie Docker jest używany do uruchamiania wszystkich komponentów, takich jak REST API, Celery oraz moduł analizy obrazów.

Docker umożliwia połączenie wszystkich modułów systemu w jedną spójną aplikację. Dzięki Docker Compose, które pozwala na definiowanie i uruchamianie wielokontenerowych aplikacji, możemy łatwo skonfigurować środowisko i uruchomić wszystkie niezbędne usługi.

REST API[13] jest głównym punktem wejściowym do systemu, umożliwiającym użytkownikom przesyłanie obrazów do analizy, monitorowanie statusu analiz oraz pobieranie wyników. Docker uruchamia serwer REST API jako jeden z kontenerów, który nasłuchuje na określonym porcie, np. http://localhost:8000. Dzięki temu użytkownicy mogą w prosty sposób komunikować się z systemem za pomocą standardowych zapytań HTTP.

Celery jest używane do zarządzania zadaniami analizy obrazów w tle. Gdy użytkownik przesyła żądanie analizy za pomocą REST API, zadanie to jest dodawane do kolejki Celery. Kontenery Celery worker odbierają zadania z kolejki i przetwarzają je asynchronicznie, co pozwala na efektywne zarządzanie dużą ilością równoczesnych żądań. Celery worker jest również uruchamiany w osobnym kontenerze, co zapewnia izolację i niezależność od innych usług.

Kontenery Docker komunikują się ze sobą za pomocą wewnętrznej sieci Docker, co zapewnia bezproblemową współpracę wszystkich komponentów systemu. REST API przekazuje zadania do Celery, który następnie zarządza przetwarzaniem i generowaniem wyników. Wyniki są przechowywane w określonym miejscu, skąd mogą być pobierane przez użytkowników za pośrednictwem REST API.

Jedną z największych zalet użycia Docker jest skalowalność. W miarę wzrostu liczby użytkowników i zapotrzebowania na zasoby, możemy łatwo zwiększać liczbę kontenerów Celery

[13] W.M.C.J.T. Kithulwatta, K.P.N. Jayasena, B. T. G. S. Kumara, R.M. K. T. Rathnayaka, *Integration With Docker Container Technologies for Distributed and Microservices Applications: A State-of-the-Art Review*, International Journal of Systems and Service-Oriented Engineering, https://doi.org/10.4018/IJSSOE.297136, 2022.

worker, co pozwala na obsługę większej liczby zadań w tym samym czasie. Docker zapewnia elastyczność w zarządzaniu zasobami, umożliwiając dynamiczne dostosowywanie się do bieżących potrzeb.

Nasz system korzysta z obrazu bazowego nvidia/cuda:11.8.0-runtime-ubuntu20.04, który zapewnia wsparcie dla technologii CUDA firmy Nvidia. Obraz ten jest zoptymalizowany pod kątem wykorzystania GPU, co jest kluczowe dla przyspieszenia procesów analizy obrazów. Dzięki temu, nasz system może korzystać z zaawansowanych możliwości obliczeniowych GPU, co znacząco zwiększa jego wydajność i efektywność.

Konteneryzacja z Dockerem, w połączeniu z technologiami Nvidia, umożliwia stworzenie wydajnego, skalowalnego i łatwego w zarządzaniu systemu segmentacji raka piersi, który może być wdrażany w różnych środowiskach medycznych.

3. Podsumowanie

Prezentowane badanie opisuje metodę tworzenia asynchronicznego systemu segmentacji raka piersi z użyciem biblioteki TiaToolbox. System analizuje obrazy w celu wykrywania raka piersi, wykorzystując TiaToolbox w Pythonie oraz FastAPI, co ułatwia wdrożenie do różnych projektów. Celery zapewnia spójność działania, umożliwiając kolejkowanie zadań bez blokowania głównego wątku aplikacji. Utworzono zapytania GET do kontrolowania stanu analizy. System dostępny jest jako dwa obrazy Docker'a, co upraszcza instalację. FastAPI zapewnia szybkie przetwarzanie zapytań, a Celery efektywne zarządzanie zadaniami, co zwiększa wydajność i niezawodność. Integracja z TiaToolbox zwiększa dokładność wykrywania raka piersi, a system można rozszerzać o dodatkowe moduły analityczne. Jest to wszechstronne narzędzie wspomagające diagnozowanie raka piersi, łatwe do wdrożenia w różnych środowiskach medycznych.

Literatura

1. Bianchini, G., Balko, J. M., Mayer, I. A., Sanders, M. E., & Gianni, L., *Triple-negative breast cancer: A distinct subtype of breast cancer with unique challenges*, Nature Reviews Clinical Oncology, https://doi.org/10.1038/nrclinonc.2016.79, 13(11), s. 674-690, 2016.
2. Chen, C., Lu, M. Y., Williamson, D. F. K., Chen, T. Y., Schaumberg, A. J., & Mahmood, F., *TIAToolbox: An End-to-End Toolbox for Advanced Tissue Image Analytics*, https://www.researchgate.net/publication/357322919_TIAToolbox_An_End-to-End_Toolbox_for_Advanced_Tissue_Image_Analytics, 2021.

3. Doe, J., & Smith, A., *Revolutionizing Concurrent Crawling: A Novel Approach to Enhance PHP-Python Integration using AMQP, Selenium, Celery, and RabbitMQ*, https://www.researchgate.net/publication/377684280_Revolutionizing_Concurrent_Crawling_A_Novel_Approach_to_Enhance_PHP-Python_Integration_using_AMQP_Selenium_Celery_and_RabbitMQ, 2023.
4. Ghezzi, A., Gabelloni, D., Martini, A., & Natalicchio, A., *Crowdsourcing: A Review and Suggestions for Future Research*, International Journal of Management Reviews, https://doi.org/10.1111/ijmr.12135, 20(2), s. 2-29, 2017.
5. Kamran, S. A., & Sabbir, A. S., *Efficient Yet Deep Convolutional Neural Networks for Semantic Segmentation*, https://www.researchgate.net/publication/318720999_Efficient_Yet_Deep_Convolutional_Neural_Networks_for_Semantic_Segmentation, 2017.
6. Khened, M., Kori, A., Rajkumar, H., Krishnamurthi, G., & Srinivasan, B., *A generalized deep learning framework for whole-slide image segmentation and analysis*, Scientific Reports, https://doi.org/10.1038/S41598-021-90444-8,11, s. 90444, 2021.
7. Kithulwatta, W. M. C. J. T., Jayasena, K. P. N., Kumara, B. T. G. S., & Rathnayaka, R. M. K. T., *Integration With Docker Container Technologies for Distributed and Micro-services Applications: A State-of-the-Art Review*, International Journal of Systems and Service-Oriented Engineering. https://doi.org/10.4018/IJSSOE.297136, 2022.
8. Long J., Shelhamer E., Darrell T., *Fully convolutional networks for semantic segmentation*, https://doi.org/10.1109/CVPR.2015.7298965, s.3431-3440, 2015.
9. Seth, A., & Sharma, S., *Semantic Segmentation: A Systematic Analysis From State-of-the-Art Techniques to Advance Deep Networks*, Journal of Information Technology Research, https://doi.org/10.4018/JITR.299388, 15(1), s. 1-28, 2022.
10. Sharma, G., Lavania, G., Goyal, D., *Rest API: Data retrieval and applications*, AIP Conference Proceedings, https://doi.org/10.1063/5.0155589, s. 2782, 2023.
11. Vardhan, J. V., Krishna, G. S., *Breast cancer segmentation using attention-based convolutional network and explainable AI. arXiv.org*. https://doi.org/10.48550/arXiv.2305.14389, 2023.

Źródła internetowe

1. https://tia-tool-box.readthedocs.io/en/latest/_autosummary/tiatoolbox.wsicore.wsireader.WSIReader.html
2. https://tia-tool-box.dcs.warwick.ac.uk/sample_wsis/wsi4_12k_12k.svs

Michał Kocik, Jakub Kuźniar, Veronika Vanivska, Aldona Świrad
Koło Naukowe Systemów Złożonych

mgr inż. Patryk Organiściak
Opiekun Koła Naukowego

Analiza szeregów czasowych z wykorzystaniem modeli opartych na drzewach decyzyjnych

Streszczenie

Prezentowane badanie omawia wykorzystanie modeli opartych na drzewach decyzyjnych, takich jak lasy losowe (Random Forest) i wzmocnienie gradientowe (XGBoost), w prognozowaniu szeregów czasowych z wyraźnym trendem. Podkreśla się znaczenie różnicowania danych w procesie prognozowania, szczególnie w kontekście szeregów czasowych z trendem. Prezentowane są kroki przetwarzania danych, analizy struktury szeregu czasowego, jak również implementacja prognozatorów oraz ocena ich wydajności.

Słowa kluczowe: szeregi czasowe, prognozowanie, Random Forest, XGBoost, TimeSeriesDifferentiator

1. Wprowadzenie

Modele oparte na drzewach decyzyjnych, takie jak drzewa decyzyjne, lasy losowe (Random Forest) i gradient boosting machines (GBMs), są znane ze swojej skuteczności i szerokiego zastosowania w różnych dziedzinach uczenia maszynowego [1]. Ich popularność wynika z kilku zalet, takich jak łatwość interpretacji, zdolność do pracy z danymi o różnej naturze (ciągłe, kategoryczne) oraz możliwość obsługi brakujących wartości. Mimo tych zalet, modele te mają istotne ograniczenia, zwłaszcza gdy chodzi o ekstrapolację, czyli prognozowanie wartości poza zakresem danych obserwowanych podczas treningu [2].

Ograniczenia te stają się szczególnie istotne przy prognozowaniu szeregów czasowych, które charakteryzują się wyraźnym trendem [3]. Tradycyjne modele oparte na drzewach decyzyjnych nie są w stanie efektywnie przewidywać wartości, które wykraczają poza zakres danych treningowych. Jest to problematyczne, ponieważ prognozy szeregów czasowych często wymagają przewidywania przyszłych wartości, które mogą znacznie odbiegać od wartości historycznych. W takich przypadkach prognozy generowane przez modele oparte na drzewach mogą znacznie odbiegać od rzeczywistego trendu, co prowadzi do niskiej dokładności prognoz.

Aby poradzić sobie z tym wyzwaniem, można skorzystać z kilku strategii. Jedną z najczęściej stosowanych technik jest różnicowanie, które polega na obliczaniu różnic między kolejnymi obserwacjami w szeregu czasowym. Zamiast modelować bezwzględne wartości szeregu czasowego, różnicowanie koncentruje się na modelowaniu względnych zmian między kolejnymi punktami danych. Ta transformacja może pomóc modelom lepiej uchwycić wzorce w danych, zwłaszcza w obecności trendów [4].

Po zastosowaniu różnicowania, szereg czasowy jest mniej podatny na wpływ trendów, co pozwala modelom lepiej przewidywać przyszłe wartości [5]. Po wygenerowaniu prognoz, transformacja różnicowania jest odwracana, aby odzyskać wartości w pierwotnej skali. W ten sposób uzyskuje się prognozy, które są bardziej zgodne z rzeczywistym trendem w danych [6].

Biblioteka skforecast, od wersji 0.10.0 lub wyższej [12], wprowadza nowy parametr różnicowania w swoich klasach prognozujących. Parametr ten wskazuje, że proces różnicowania musi być zastosowany przed treningiem modelu. Implementacja tego procesu jest realizowana za pomocą transformatora o nazwie skforecast.preprocessing. TimeSeriesDifferentiator. Transformator ten automatycznie różnicuje dane przed treningiem modelu i odwraca różnicowanie podczas fazy prognozowania, zapewniając, że wartości prognoz są w tej samej skali co oryginalne dane.

Pozwolenie na zarządzanie wszystkimi transformacjami wewnętrznie ma kilka istotnych zalet. Po pierwsze, zapewnia, że te same transformacje są stosowane zarówno podczas treningu modelu, jak i podczas prognozowania na nowych danych [7]. Jest to szczególnie ważne, gdy nowe dane nie następują bezpośrednio po danych treningowych, na przykład, gdy model nie jest ponownie trenowany dla każdej fazy prognozowania. W takich przypadkach automatycznie dostosowuje się rozmiar ostatniego okna potrzebnego do generowania predyktorów, stosuje różnicowanie do nowych danych i odwraca je w końcowych prognozach [8].

Te transformacje są nietrywialne i bardzo podatne na błędy, dlatego skforecast stara się unikać nadmiernego komplikowania już i tak trudnego zadania prognozowania szeregów czasowych [9]. Wprowadzenie automatycznego różnicowania wewnętrznego w bibliotece skforecast ułatwia proces modelowania szeregów czasowych z trendami, zwiększając dokładność i niezawodność prognoz.

Niniejszy artykuł pokazuje, jak różnicowanie może być wykorzystane do modelowania szeregów czasowych z pozytywnym trendem, przy użyciu modeli opartych na drzewach decyzyjnych, takich jak random forest i gradient boosting (xgboost). Prezentuje on krok po kroku, jak zastosować te techniki, aby uzyskać lepsze prognozy i jak unikać typowych błędów związanych z różnicowaniem [10].

2. Wybór danych

Chcąc wybrać szereg z mocnym trendem, posłużono się portalem fred.stlouisfed.org. FRED (Federal Reserve Economic Data) to internetowa baza danych zarządzana przez Federal Reserve Bank of St. Louis. Portal ten dostarcza szeroki zakres danych ekonomicznych i finansowych z całego świata. Jego głównym celem jest udostępnianie danych, które mogą być używane przez badaczy, analityków oraz ogół społeczeństwa do analizy i interpretacji zjawisk ekonomicznych. Szereg czasowy, który posłuży do demonstracji podanych modeli i metod analizy, to "Production, Sales, Work Started and Orders: Production Volume: Economic Activity: Manufacturing for Poland". Jest to szereg czasowy, który reprezentuje wolumen produkcji w sektorze produkcyjnym w Polsce.

2.1. Analiza struktury szeregu czasowego

- Dane w tym szeregu czasowym były zbierane miesięcznie, przez to analizując jeden rok jest dostęp do 12 próbek zebranych w tym czasie. Pierwsza próbka datowana jest na styczeń 1985 r., a ostatnia na marzec 2024r.

- Szereg czasowy cechuje się bardzo wyraźnym trendem wzrostowym. Oznacza to, że wartości obserwowane w tym szeregu mają tendencję do systematycznego wzrostu w czasie, co jest typowe dla rozwijającej się gospodarki oraz sektora produkcyjnego, który rozszerza swoje moce produkcyjne.

- Możliwe, lecz nie tak wyraźne są sezonowe wzorce, które mogą występować w regularnych odstępach czasu (np. miesięcznych lub rocznych), co jest typowe dla produkcji przemysłowej, gdzie cykle produkcyjne mogą być związane z sezonowym popytem.

- Szereg czasowy wyraźnie pokazuje moment wystąpienia pandemii COVID-19, co wpłynęło na wartości obserwowane. Pandemia miała znaczący wpływ na sektor produkcyjny, powodując zakłócenia w łańcuchach dostaw, zamknięcia zakładów produkcyjnych i zmniejszenie popytu na

niektóre produkty. Te zmiany są wyraźnie widoczne w danych jako gwałtowne spadki i fluktuacje w określonych okresach.

- Szereg czasowy przed dalszą analizą został edytowany w ten sposób, że wyodrębnione zostało 10 lat (120 miesięcy). Zaczynają się one od marca 2014 do marca 2024.

3. Przetwarzanie szeregu czasowego

Środowisko (IDE) wykorzystywane w tej analizie to Visual Studio Code Microsoft. Jest ono bardzo popularne i przyjazne dla użytkowników oraz deweloperów. Język programowania używany w tym zadaniu to Python, jeden z popularniejszych i szeroko wykorzystywanych narzędzi. Python ma szerokie zastosowanie zwłaszcza w dziedzinach Data Science, dlatego został wybrany ze względu na naturę badanego zagadnienia.

3.1. Opis potrzebnych bibliotek

W celu przeprowadzenia analizy szeregów czasowych użyto kilku istotnych bibliotek i modułów w Pythonie:

Listing 1 Wykorzystywane biblioteki

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from xgboost import XGBRegressor
from sklearn.ensemble import RandomForestRegressor
from skforecast.ForecasterAutoreg import ForecasterAutoreg
from skforecast.model_selection import backtesting_forecaster
from skforecast.preprocessing import TimeSeriesDifferentiator
from sklearn.metrics import mean_absolute_error
```

Poniżej znajduje się krótki opis każdej z nich:

3. NumPy - fundamentalna biblioteka dla obliczeń numerycznych w Pythonie. Umożliwia przeprowadzanie operacji na tablicach wielowymiarowych oraz zapewnia bogaty zbiór funkcji matematycznych,
4. Pandas - biblioteka służąca do manipulacji i analizy danych. Umożliwia łatwe wczytywanie, przekształcanie oraz analizowanie danych tabelarycznych za pomocą struktur danych takich jak DataFrame,
5. Matplotlib - biblioteka do tworzenia statycznych, animowanych i interaktywnych wizualizacji w Pythonie. Pyplot to jej podmoduł, który umożliwia tworzenie wykresów w sposób zbliżony do MATLABa,

6. XGBoost - wydajna i elastyczna biblioteka do uczenia gradientowego boostingu, stosowana w wielu konkursach machine learningowych. XGBRegressor to jej implementacja dla zadań regresyjnych.
7. RandomForestRegressor - moduł z biblioteki scikit-learn, który implementuje algorytm lasów losowych dla zadań regresyjnych. Lasy losowe to zbiór drzew decyzyjnych, które wspólnie podejmują decyzję,
8. ForecasterAutoreg - moduł z biblioteki skforecast, który umożliwia tworzenie modeli autoregresyjnych dla szeregów czasowych. Wspiera różne algorytmy regresji, w tym drzewa decyzyjne i XGBoost,
9. Backtesting_forecaster - narzędzie z biblioteki skforecast, które pozwala na przeprowadzanie testów historycznych (backtesting) modeli prognozujących. Umożliwia ocenę skuteczności modeli na podstawie danych historycznych,
10. TimeSeriesDifferentiator - narzędzie z biblioteki skforecast, które umożliwia różnicowanie szeregów czasowych. Proces różnicowania pomaga w stabilizowaniu trendów i sezonowości przed trenowaniem modeli,
11. Mean_absolute_error - funkcja z biblioteki scikit-learn, która oblicza średni błąd bezwzględny (MAE) między przewidywanymi a rzeczywistymi wartościami.

3.2. Wprowadzenie danych do środowiska

Dane pobrano w formacie CSV, a następnie plik nazwano ts1.csv.

Listing 2 Wprowadzenie danych

```
# Wczytanie danych i konwersja na DataFrame
ts1 = pd.read_csv('ts1.csv')
# Wyświetlenie pięciu pierwszych wierszy
print(ts1.head())
```

Tabela 1 Pierwsze wiersze danych szeregu czasowego.

Parametr ‘DATE’	Parametr ‘POLPROMANMISMEI’
2014-03-01	93.81107
2014-04-01	95.06747
2014-05-01	93.39227
2014-06-01	94.64867
2014-07-01	95.90508

Do dalszych działań dodano znacznik czasu, którym będzie data. Nadano kolumnie ‘DATE’ format daty.

Listing 3 Konwersja kolumny 'DATE' na typ daty

```
ts1['DATE'] = pd.to_datetime(ts1['DATE'], format='%Y-%m')
```

Ten fragment kodu konwertuje kolumnę 'DATE' na typ daty, używając formatu %Y-%m, gdzie %Y reprezentuje rok, a %m miesiąc.

3.3. Przetwarzanie danych

Przetwarzanie danych to zestaw operacji mających na celu przekształcenie surowych danych w użyteczną formę, która może być analizowana i interpretowana.

Listing 4 Przetwarzanie danych

```
# Ustawienie kolumny 'DATE' jako indeks
ts1 = ts1.set_index('DATE')
# Ustawienie częstotliwości danych na miesięczną (start miesiąca)
ts1 = ts1.asfreq('MS')
# Wybór kolumny 'POLPROMANMISMEI'
ts1 = ts1['POLPROMANMISMEI']
# Sortowanie indeksu
ts1 = ts1.sort_index()
ts1.head(4)
```

Ustawiono kolumnę 'DATE' jako indeks DataFrame, co pozwala na łatwiejsze operacje na szeregach czasowych. Nadano częstotliwość indeksu na miesięczną, gdzie 'MS' oznacza początek miesiąca.

Tabela 2 Wygląd fragmentu danych po przetworzeniu danych

Parametr ‘DATE'	
2014-03-01	93.81107
2014-04-01	95.06747
2014-05-01	93.39227
2014-06-01	94.64867

4. Różnicowanie

W kontekście analizy szeregów czasowych, "Data differentiated" odnosi się do procesu różnicowania danych (differencing), który jest techniką stosowaną w celu uczynienia szeregu czasowego stacjonarnym. Szereg czasowy jest stacjonarny, jeśli jego statystyki, takie jak średnia i wariancja, są stałe w czasie. Różnicowanie jest szczególnie przydatne, gdy dane mają trend lub sezonowość.

Różnicowanie polega na odejmowaniu wartości poprzedniego okresu od wartości bieżącej, co formalnie zapisuje się jako:

Równanie 1 Różnicowanie

$$Y_t' = Y_t - Y_{t-1}$$

Gdzie:
Y_t – to wartości szeregu czasowego o okresie t
Y_{t-1} - to wartości szeregu czasowego o okresie t-1
Y_t' - to zróżnicowana wartość szeregu czasowego o okresie t

4.1. Działanie różnicowania

Różnicowanie pomaga usunąć trend i sezonowość z szeregu czasowego, co czyni go bardziej stacjonarnym. Jest to ważny krok w analizie szeregów czasowych. Używam do tego funkcji TimeSeriesDifferentiator.

Utworzono instancję klasy TimeSeriesDifferentiator, z parametrem order=1. Oznacza on, że dokonano pierwsze różnicowanie, tj. różnicę pomiędzy każdą wartością a jej poprzednikiem. TimeSeriesDifferentiator jest częścią biblioteki specyficznej dla obróbki szeregów czasowych.

Listing 5 Instancja klasy TimeSeriesDifferenitiator

```
diferenciator = TimeSeriesDifferentiator(order=1)
```

4.2. Konwersja danych na dwa różne formaty

Przekonwertowanie danych na numpy.ndarray jest wymagane, ponieważ niektóre metody z biblioteki skforecast mogą nie obsługiwać bezpośrednio obiektów pandas.Series czy DataFrame. Do tego celu stworzone zostały dwie zmienne:

12. ts1 – pozostanie ona w formacie pandas.core.series.Series
13. ts11 – przekonwertuje ją na zmienną typu numpy.ndarray

Listing 6 Zmienne ts1 i ts11 w różnych formatach

```
type(ts1)
pandas.core.series.Series

# Konwersja pandas.Series na numpy.ndarray
ts11 = ts1.values

type(ts11)
numpy.ndarray
```

4.3. Zastosowanie różnicowania na danych

Metoda fit_transform(data) najpierw "dopasowuje" różnicowanie do danych, a następnie przekształca dane, zwracając różnicowane wartości.

Listing 7 Zastosowanie różnicowania na danych

```
ts1_diff = diferenciator.fit_transform(ts11)
ts1_diff.head(4)
```

Tabela 3 Fragment danych po różnicowaniu

Parametr 'DATE'	Wartość danych po zróżnicowaniu
2014-04-01	1.25640
2014-05-01	-1.67520
2014-06-01	1.25640
2014-07-01	1.25641

5. Podział danych na dane testowe i treningowe

Podział danych na dane testowe i treningowe jest kluczowym krokiem w uczeniu maszynowym. Zapewnia on to, że model jest oceniany na danych, których nigdy wcześniej nie widział, co pozwala na sprawdzenie jego zdolności do generalizacji. Dokonano podziału danych w proporcji 2:1. Dane treningowe zawierają informacje o okresie pandemii.

Przeniesienie danych z okresu pandemii do danych treningowych może wpłynąć na sposób, w jaki model uczy się reprezentować ten szczególny czas. Jednak, jak można zauważyć w późniejszym etapie, może to nie mieć aż tak wielkiego wpływu, jakby się wydawało. Sugeruje to, że model potrafi efektywnie generalizować poza ten okres. Jest to ciekawy aspekt analizy danych, który może prowadzić do głębszego zrozumienia zachowań modelu w obliczu nieprzewidywalnych okoliczności, takich jak pandemia.

Ważne jest również, aby upewnić się, że dane testowe rozpoczynają się od momentu, który jest zaraz po okresie danych treningowych. To zapewnia, że model jest testowany na danych, które są rzeczywiście nowe dla niego i które powinny odzwierciedlać warunki poza okresem, na którym był trenowany.

Listing 8 Podział danych na dane treningowe i testowe

```
end_train = '2021-03-01'
print(
```

```
    f"Train dates : {ts1.index.min()} --- {ts1.loc[:end_train].index.max()}  "
    f"(n={len(ts1.loc[:end_train])})")
print(
    f"Test dates  : {ts1.loc[end_train:].index.min()} --- {ts1.index.max()}  "
    f"(n={len(ts1.loc[end_train:])})")
```

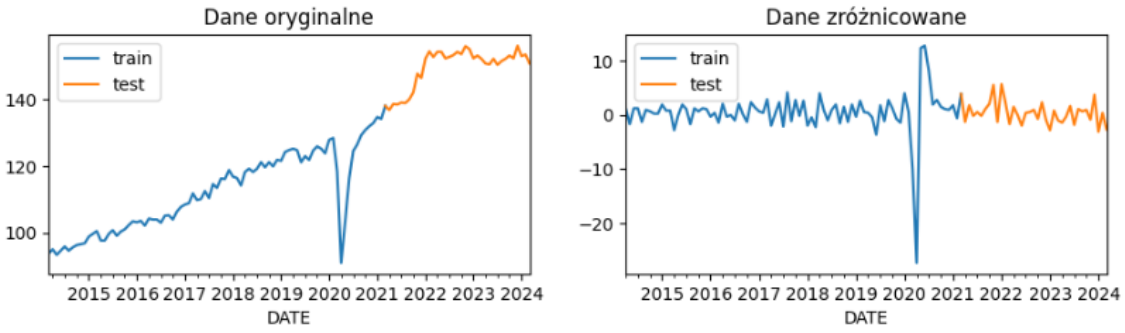
Tabela 4 Dane po podziale na część treningową i testową

Dane treningowe	2014-03-01 00:00:00 --- 2021-03-01 00:00:00 (n=85)
Dane testowe	2021-03-01 00:00:00 --- 2024-03-01 00:00:00 (n=37)

Aby zilustrować ten podział możemy stworzyć wykresy przedstawiające dane testowe i dane treningowe.

Listing 9 Kod generujący wykresy danych testowych i danych treningowych

```
fig, axs = plt.subplots(1, 2, figsize=(11, 2.5))
axs = axs.ravel()
ts1.loc[:end_train].plot(ax=axs[0], label='train')
ts1.loc[end_train:].plot(ax=axs[0], label='test')
axs[0].legend()
axs[0].set_title('Dane oryginalne')

ts1_diff.loc[:end_train].plot(ax=axs[1], label='train')
ts1_diff.loc[end_train:].plot(ax=axs[1], label='test')
axs[1].legend()
axs[1].set_title('Dane zróżnicowane')
```



Rysunek 1 Wykresy podziałów danych oryginalnych i danych zróżnicowanych ze względu na dane testowe i treningowe.

6. Prognozowanie używając drzew losowych i wzmocnienia gradientowego

Utworzono dwa autoregresyjne prognozatory. Jeden za pomocą RandomForestRegressor z biblioteki scikit-learn, a drugi z użyciem XGBoost.

RandomForestRegressor i XGBRegressor są dwoma popularnymi modelami używanymi do problemów regresji w uczeniu maszynowym. Oto krótkie opisy obu modeli:

14. RandomForestRegressor - jest modelem zespołowym opartym na drzewach decyzyjnych. Tworzy on wiele drzew decyzyjnych podczas treningu i łączy ich przewidywania w celu uzyskania bardziej stabilnych i dokładnych prognoz. Każde drzewo decyzyjne w lesie losowym jest trenowane na losowo wybranym podzbiorze danych treningowych i losowo wybranych funkcjach, co pomaga uniknąć przetrenowania i zwiększyć różnorodność drzew w lesie. Podczas prognozowania, wyniki z poszczególnych drzew są uśredniane, co pomaga zredukować wariancję modelu,
15. XGBRegressor - implementacja algorytmu Gradient Boosting Machine, która wykorzystuje bibliotekę XGBoost. Jest to silny model zespołowy, który buduje sekwencyjnie drzewa decyzyjne, a następnie poprawia błędy poprzednich drzew poprzez dopasowywanie nowych drzew do reszt (gradientu) poprzednich przewidywań. XGBoost wykorzystuje wiele technik optymalizacyjnych, takich jak regularyzacja, próbkowanie wierszy i kolumn, aby zapobiec przetrenowaniu i zwiększyć wydajność modelu. Ponadto, XGBoost umożliwia elastyczne dostosowywanie hiperparametrów, co pozwala na optymalizację modelu pod kątem konkretnych danych i problemu.

Oba modele, RandomForestRegressor i XGBRegressor, są potężnymi narzędziami w dziedzinie uczenia maszynowego i znajdują zastosowanie w szerokim zakresie problemów regresji, od przewidywania cen nieruchomości po prognozowanie przychodów biznesowych. Oba zostały wytrenowane na danych od 2014-03-01 do 2024-03-01 i generują prognozy na kolejne 36 miesięcy (3 lata).

6.1. Prognozowanie bez różnicowania

Prognozowanie bez uwzględniania różnicowania w danych może prowadzić do kilku istotnych konsekwencji. Po pierwsze, model, zwłaszcza z wyraźnym trendem, może nie być w stanie uchwycić subtelnych różnic między różnymi grupami lub warunkami, co może prowadzić do przeszacowania lub niedoszacowania pewnych zjawisk. Na przykład, w przypadku analizy danych związanych z pandemią, brak różnicowania może sprawić, że model nie będzie w stanie uwzględnić różnych strategii zarządzania pandemią, które mogą mieć wpływ na dane. Ponadto, prognozowanie bez różnicowania może prowadzić do uproszczeń i generalizacji, które mogą nie odzwierciedlać rzeczywistości w pełni. Aby to udowodnić wykonam prognozy utworzone bez wcześniejszego różnicowania.

Listing 10 Przygotowanie do tworzenia prognoz bez różnicowania

```
steps = len(ts1.loc[end_train:])
forecaster_rf = ForecasterAutoreg(
                    regressor = RandomForestRegressor(random_state=963),
                    lags      = 12
                   # brak różnicowania
                )
forecaster_gb = ForecasterAutoreg(
                    regressor = XGBRegressor(random_state=963),
                    lags      = 12
                   # brak różnicowania
```

Listing 11 Trenowanie danych bez różnicowania

```
# Train
forecaster_rf.fit(ts1.loc[:end_train])
forecaster_gb.fit(ts1.loc[:end_train])
```

Listing 12 Predykcja danych bez różnicowania

```
# Predict
predictions_rf = forecaster_rf.predict(steps=steps)
predictions_gb = forecaster_gb.predict(steps=steps)
```

Błąd MEA (Mean Absolute Error) jest jednym z powszechnie stosowanych wskaźników oceny wydajności modelu w uczeniu maszynowym i analizie danych. Jest to miara średniej wartości bezwzględnej różnicy między wartościami przewidywanymi przez model a rzeczywistymi wartościami.

Aby obliczyć błąd MEA, dla każdej obserwacji obliczana jest różnica między wartością przewidywaną przez model a rzeczywistą wartością, a następnie wszystkie te różnice są uśredniane, ignorując ich kierunek (czyli wartości bezwzględne).

Błąd MEA jest wyrażony w tych samych jednostkach co oryginalne dane, co ułatwia interpretację. Im niższa wartość błędu MEA, tym lepiej dopasowany jest model do danych. Jest to użyteczne narzędzie zarówno podczas tworzenia modelu, jak i jego oceny, ponieważ umożliwia porównanie wydajności różnych modeli lub ocenę, czy model osiąga satysfakcjonujące rezultaty.

Listing 13 Obliczenie błędu MAE danych bez różnicowania

```
# Error
error_rf = mean_absolute_error(ts1.loc[end_train:], predictions_rf)
error_gb = mean_absolute_error(ts1.loc[end_train:], predictions_gb)
print(f"Error (MAE) Random Forest: {error_rf:.2f}")
print(f"Error (MAE) Gradient Boosting: {error_gb:.2f}")
```

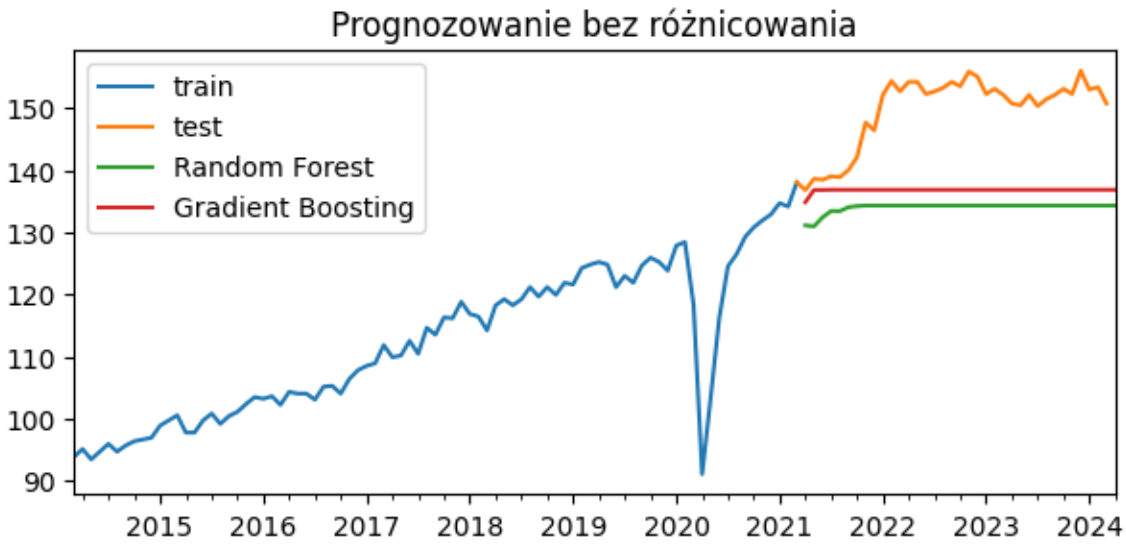
Tabela 5 Błędy MEA dla danych bez różnicowania

Nazwa algorytmu	Błąd MAE
Random Forest	15.56
Gradient Boosting	12.78

Otrzymane błędy nie są duże, lecz nadal są większe niż gdyby przeprowadzić wcześniej różnicowanie. Chcąc sprawdzić jak dane modele poradziły sobie w prognozowaniu można narysować odpowiedni wykres.

Listing 14 Kod generujący wykres prognozowania bez różnicowania

```
fig, ax = plt.subplots(figsize=(7, 3), sharex=True, sharey=True)
ts1.loc[:end_train].plot(ax=ax, label='train')
ts1.loc[end_train:].plot(ax=ax, label='test')
predictions_rf.plot(ax=ax, label='Random Forest')
predictions_gb.plot(ax=ax, label='Gradient Boosting')
ax.set_title(f'Prognozowanie bez różnicowania')
ax.set_xlabel('')
ax.legend()
```



Rysunek 2 Wykres prognozowania bez różnicowania

Wykres pokazuje, że żaden z modeli nie jest w stanie dokładnie przewidzieć trendu. Po kilku krokach prognozy stają się niemal stałe, zbliżone do maksymalnych wartości obserwowanych w danych treningowych.

Następnie dwa nowe prognozatory są trenowane przy użyciu tej samej konfiguracji, ale z argumentem differentiation = 1. Aktywuje to wewnętrzny proces różnicowania szeregu czasowego przed trenowaniem modelu i odwraca różnicowanie, znane również jako całkowanie, dla wartości prognozowanych.

6.2. Prognozowanie z różnicowaniem

Prognozowanie z różnicowaniem dla danych z wyraźnym trendem ma na celu uwzględnienie zmian w trendzie w różnych grupach lub warunkach w danych w celu uzyskania bardziej precyzyjnych prognoz. Gdy dane wykazują wyraźny trend, różnice między grupami mogą wpływać na dynamikę tego trendu, dlatego ważne jest uwzględnienie tych różnic w procesie prognozowania.

Listing 15 Przygotowanie do tworzenia prognoz z różnicowaniem

```
steps = len(ts1.loc[end_train:])
forecaster_rf = ForecasterAutoreg(
                    regressor       = RandomForestRegressor(random_state=910),
                    lags            = 12,
                    differentiation = 1 #dodajemy różnicowanie
                )
forecaster_gb = ForecasterAutoreg(
                    regressor       = XGBRegressor(random_state=910),
                    lags            = 12,
                    differentiation = 1 #dodajemy różnicowanie
                )
```

Listing 16 Trenowanie danych z różnicowaniem

```
# Train
forecaster_rf.fit(ts1.loc[:end_train])
forecaster_gb.fit(ts1.loc[:end_train])
```

Listing 17 Predykcja danych z różnicowaniem

```
# Predict
predictions_rf = forecaster_rf.predict(steps=steps)
predictions_gb = forecaster_gb.predict(steps=steps)
```

Listing 18 Obliczenie błędu MEA danych z różnicowaniem

```
# Error
error_rf = mean_absolute_error(ts1.loc[end_train:], predictions_rf)
error_gb = mean_absolute_error(ts1.loc[end_train:], predictions_gb)
print(f"Error (MAE) Random Forest: {error_rf:.2f}")
print(f"Error (MAE) Gradient Boosting: {error_gb:.2f}")
```

Tabela 6 Błędy MAE dla danych z różnicowaniem

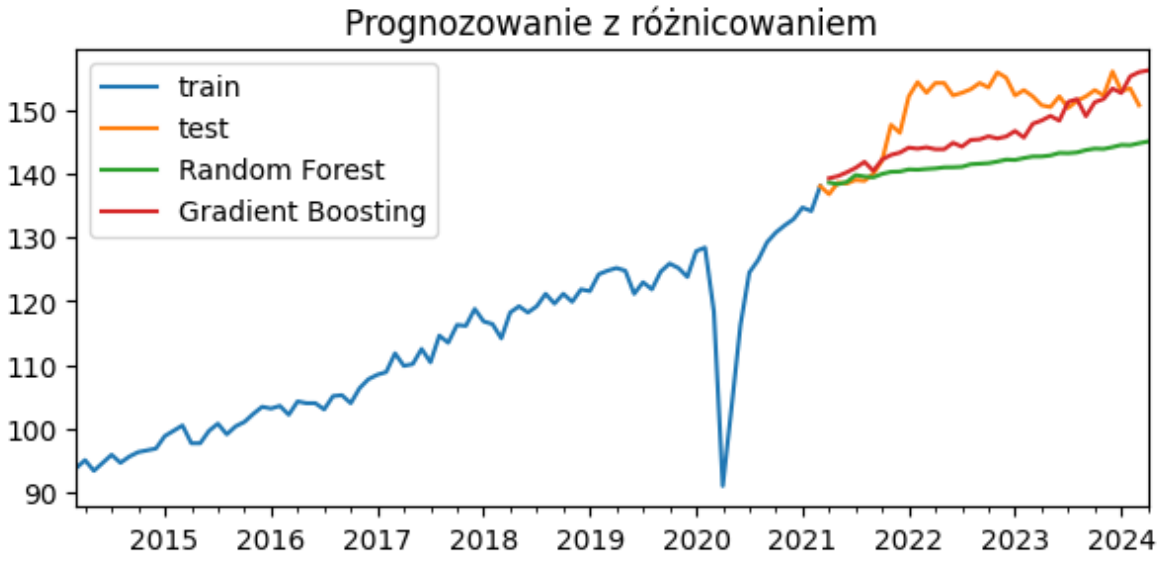
Nazwa algorytmu	Błąd MAE
Random Forest	8.01
Gradient Boosting	4.56

Można zauważyć, że błędy MAE są zdecydowanie mniejsze niż gdy nie wykonałam różnicowania. Zmiany można również zbadać na wykresie. Dane nie są idealnie dopasowane, co

jest poprawne, lecz mogłyby być lepiej dopasowane, lecz występująca pandemia może mieć wpływ na działanie algorytmów. Niemniej jednak, dane dalej są prognozowane w zadawalający sposób. Zwłaszcza dużo lepiej spisuje się Gradient Boosting.

Listing 19 Kod generujący wykres prognozowania z różnicowaniem

```
fig, ax = plt.subplots(figsize=(7, 3), sharex=True, sharey=True)
ts1.loc[:end_train].plot(ax=ax, label='train')
ts1.loc[end_train:].plot(ax=ax, label='test')
predictions_rf.plot(ax=ax, label='Random Forest')
predictions_gb.plot(ax=ax, label='Gradient Boosting')
ax.set_title(f'Prognozowanie z różnicowaniem')
ax.set_xlabel('')
ax.legend()
```



Rysunek 3 Wykres prognozowania z różnicowaniem

Poprzedni przykład pokazał, jak łatwo jest wprowadzić różnicowanie do procesu prognozowania dzięki funkcjonalnościom dostępnym w skforecast. Jednakże, aby osiągnąć płynną interakcję, konieczne jest zastosowanie kilku nietrywialnych transformacji.

7. Transformator TimeSeriesDifferentiator

TimeSeriesDifferentiator to niestandardowy transformator, który został zaprojektowany do przetwarzania danych szeregów czasowych w kontekście uczenia maszynowego, a szczególnie jako część przetwarzania wstępnego w bibliotece Scikit-learn (sklearn). Jako transformator

Scikit-learna, TimeSeriesDifferentiator implementuje metody fit, transform, fit_transform oraz inverse_transform, które są standardowymi metodami interfejsu API Scikit-learn.

Proces odwrotnej transformacji, inverse_transform, można zastosować tylko do tego samego szeregu czasowego, który wcześniej był różnicowany za pomocą tego samego obiektu TimeSeriesDifferentiator. To ograniczenie wynika z konieczności użycia początkowych n wartości szeregu czasowego, gdzie n równa się rzędowi różnicowania, do pomyślnego odwrócenia różnicowania. Te wartości są przechowywane podczas wykonywania metody fit.

W transformatorze TimeSeriesDifferentiator dostępna jest dodatkowa metoda inverse_transform_next_window. Metoda ta została zaprojektowana do użycia wewnątrz prognozatorów (Forecasters) w celu odwrócenia różnicowania wartości prognozowanych. Jeśli regressor prognozatora jest trenowany na różnicowanym szeregu czasowym, to prognozowane wartości również będą różnicowane. Metoda inverse_transform_next_window pozwala na przywrócenie prognoz do oryginalnej skali, zakładając, że zaczynają się one bezpośrednio po ostatnich obserwowanych wartościach (last_window).

7.1. Różnicowanie wewnętrzne, a przetwarzanie wstępne

Prognozatory zarządzają procesem różnicowania wewnętrznie, więc nie ma potrzeby dodatkowego przetwarzania wstępnego szeregu czasowego ani postprocesowania prognoz. Porównane zostaną wyniki obu podejść.

Listing 20 Różnicowanie szeregu czasowego za pomocą TimeSeriesDifferentiator

```
diferenciator = TimeSeriesDifferentiator(order=1)
data_diff = diferenciator.fit_transform(ts11)
data_diff = pd.Series(data_diff, index=ts1.index).dropna()

forecaster = ForecasterAutoreg(
                 regressor = RandomForestRegressor(random_state=963),
                 lags      = 15
             )
forecaster.fit(y=data_diff.loc[:end_train])
predictions_diff = forecaster.predict(steps=steps)

# Revert differentiation to obtain final predictions
last_value_train = ts1.loc[:end_train].iloc[[-1]]
predictions_1 = pd.concat([last_value_train, predictions_diff]).cumsum()[1:]
predictions_1 = predictions_1.asfreq('MS')
predictions_1.name = 'pred'
predictions_1.head(5)
```

Tabela 7 Fragment danych po zróżnicowaniu szeregu czasowego za pomocą TimeSeriesDifferentiator

Data	Prognoza
2021-04-01	139.062341
2021-05-01	139.041399
2021-06-01	138.315471
2021-07-01	138.961815
2021-08-01	138.351763

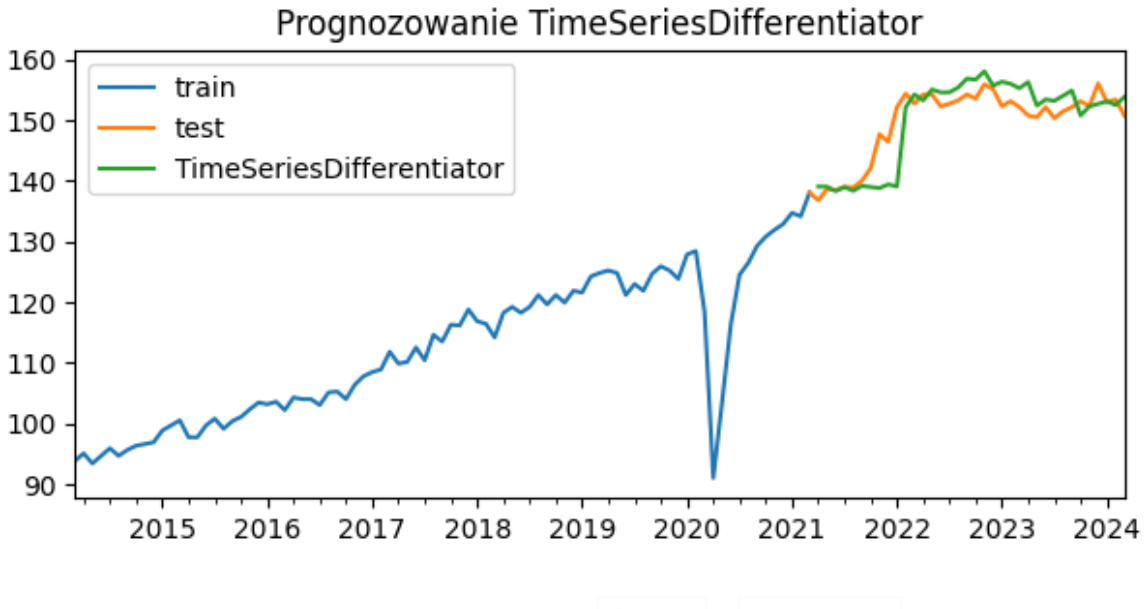
Listing 21 Błędy MAE danych po uzyciu TimeSeriesDifferentiator

```
# Error
error_1 = mean_absolute_error(ts1.loc[end_train:], predictions_1)
error_1
8.622837859459462
```

Można zauważyć, że błąd MAE jest średniej wielkości, raczej interpretowany jako nieduży. Dobrze zwiastuje to na naszą prognozę.

Listing 22 Kod generujący wykres prognozowania TimeSeriesDifferentiator

```
fig, ax = plt.subplots(figsize=(7, 3), sharex=True, sharey=True)
ts1.loc[:end_train].plot(ax=ax, label='train')
ts1.loc[end_train:].plot(ax=ax, label='test')
predictions_1['pred'].plot(ax=ax, label='TimeSeriesDifferentiator')

ax.set_title(f'Prognozowanie TimeSeriesDifferentiator')
ax.set_xlabel('')
ax.legend()
```



Rysunek 4 Wykres prognozowania po użyciu TimeSeriesDifferentiator

Jak można zauważyć, że ten sposób prognozowania wydaje się najbardziej trafny i można uznać go za najlepszy spośród badanych.

Podsumowanie

Wprowadzenie różnicowania do modelowania szeregów czasowych z wyraźnym trendem okazuje się kluczowym krokiem, który pozwala na lepsze przewidywanie przyszłych wartości. Ten artykuł naukowy eksploruje skuteczność modeli opartych na drzewach decyzyjnych, takich jak random forest i gradient boosting, w połączeniu z techniką różnicowania, aby zaradzić problemom ekstrapolacji w prognozowaniu szeregów czasowych. Przy użyciu biblioteki skforecast, wprowadzono nowy parametr różnicowania, który automatyzuje proces różnicowania wewnętrznego. Analiza danych pokazuje wyraźny trend wzrostowy, który jest typowy dla sektora produkcyjnego. Dzięki różnicowaniu, modele są w stanie lepiej przewidywać przyszłe wartości, zwłaszcza w obecności trendów. Eksperymenty z modelami Random Forest i XGBoost zarówno z, jak i bez różnicowania, wykazały, że modele z różnicowaniem osiągają niższe błędy MAE, co sugeruje ich większą dokładność w przewidywaniu. Różnicowanie pozwala uniknąć przeszacowania i niedoszacowania, co prowadzi do lepszej generalizacji modeli w przypadku szeregów czasowych z trendami. Dodatkowo, narzędzie TimeSeriesDifferentiator biblioteki skforecast umożliwia łatwe stosowanie różnicowania w analizie szeregów czasowych, co zwiększa wygodę i efektywność procesu modelowania.

Literatura

[1] A. Nielsen. (2019). Practical Time Series Analysis: Prediction with Statistics and Machine Learning (pp. 243-250). Helion

[2] M. Muller (2007). Dynamic Time Wrapping (pp. 69-84). Springer Berlin Heidelberg

[3] Joan Serr`a, Josep Ll. Arcos (2018). An Empirical Evaluation of Similarity Measures for Time Series Classification. o Knowledge-based Systems

[4] Ch. A. Ratanamahatana , J. Lin , D. Gunopulos , E. Keogh , M. Vlachos , G. Das. (2012). Mining Time Series Data (pp. 1049-1073). Springer

[5] S.Makridakis, E. Spiliotis ,V. Assimakopoulos (2018). The M4 Competition: Results, findings, conclusion and way forward. Elsevier

[6] A. Géron (2020). Uczenie maszynowe z użyciem Scikit-Learn i TensorFlow. Wydanie II, (rozdział 15.), Helion

[7] L.Zhu. N. Laptev (2017). Deep and Confident Prediction for Time Series at Uber, ICDWM, New Orleans

[8] Z. Che, S. Purushotham, K. Cho, D. Sontag , Y. Liu (2018) Recurrent Neural Networks for Multivariate Time Series with Missing Values., Nature

[9] P. Esling, C. Agon (2012). Time-series data mining. ACM Computing Surveys

[10] Davis, H. T., & Nelson, W. F. C. (1935). The analysis of time series. In H. T. Davis & W. F. C. Nelson, Elements of statistics with applications to economic data (pp. 119–145). Principia Press.

Źródła internetowe

[11] https://fred.stlouisfed.org/series/POLPROMANMISMEI (dostęp: 03.05.2024)
[12] https:// skforecast.org (dostęp: 03.05.2024).

Michał Kocik, Jakub Kuźniar, Veronika Vanivska, Aldona Świrad
Koło Naukowe Systemów Złożonych

mgr inż. Patryk Organiściak
Opiekun Koła Naukowego

Inżynieria wsteczna relacyjnej bazy danych z głównego urzędu statystycznego (GUA) w celu optymalizacji zapytań aktywnego formularza z danymi adresowymi

Streszczenie

Prezentowane badanie opisuje metodę tworzenia bazy danych adresów z arkuszy dostarczonych przez Główny Urząd Statystyczny. Omówiono proces odtworzenia tej bazy danych oraz jej wykorzystanie w dynamicznym wyborze adresu przez użytkownika. Rozwiązanie to pozwala uniknąć konieczności korzystania z interfejsu programistycznego (API), zapewnia spójność danych oraz zwiększa szybkość systemu co ma pozytywny wpływ na odbiór przez użytkownika (User Experience). Przedstawiono szczegółowy opis procesu konstrukcji bazy danych oraz jej praktycznej implementacji. Celem tego projektu jest zwiększenie zrozumienia sposobów wykorzystania danych statystycznych do tworzenia funkcjonalnych narzędzi dla użytkowników końcowych, zwłaszcza w kontekście ograniczeń związanych z dostępem do zewnętrznych zasobów.

Słowa kluczowe: gus, bazy danych, dane adresowe.

1. Wprowadzenie

Inżynieria odwrotna baz danych ma na celu pozyskanie abstrakcyjnych i koncepcyjnych opisów lub schematów danych istniejącego systemu informacyjnego, czyli jego bazy danych, bez względu na to, czy jest ona zrealizowana jako zbiór plików, czy też za pośrednictwem rzeczywistego systemu zarządzania bazą danych [1]. Zbiór danych w formie arkusza XLSX lub plików CSV, jeżeli spełnia określone wymogi formatu spójności danych może zostać przetworzony do postaci relacyjnej bazy danych.

Relacyjna baza danych to system zarządzania danymi oparty na modelu relacyjnym, który organizuje dane w tabele składające się z wierszy i kolumn. Każda tabela reprezentuje zbiór rekordów (wierszy), z których każdy składa się z atrybutów (kolumn) [2]. Relacje między tabelami są definiowane za pomocą kluczy obcych, co umożliwia powiązanie danych z różnych tabel. Relacyjne bazy danych używają języka SQL (Structured Query Language) do zarządzania i manipulowania danymi, umożliwiając efektywne wyszukiwanie, aktualizację i zarządzanie informacjami [3].

Celem projektu jest otworzenie bazy danych na podstawie zbioru plików CSV udostępnionych przez GUS z danymi adresowymi. Utworzona baza docelowo jest elementem systemu, w którym użytkownik dynamicznie wybiera adres na podstawie predefiniowanych składowych, takich jak miejscowość czy województwo. Wybrane podejście, w porównaniu do wykorzystania API, zapewnia znacznie szybsze generowanie dynamicznych list z polami wyboru elementów adresów, co przekłada się na pozytywny odbiór systemu przez użytkownika (UX) [4] oraz zapewnia inwariantność danych.

System wraz z przepływem danych mają pozwolić na zachowanie spójności danych podczas ich importu do z system przeznaczonego do raportowania uczestników projektów państwowych. Integralność danych jest krytycznym elementem zarządzania cyklem życia danych [5]. Błędy w systemach internetowych mogą powodować podatności na ataki [6]. Ataki te mogą powodować wycieki danych, które są bardzo kosztowne dla firm lub organizacji [7]. Aby wyeliminować ewentualne błędy, stosuje się testy systemów przepływu danych [8].

W badanym problemie, zbiór danych pochodzi z bazy danych Głównego Urzędu Statystycznego [11]. Są to otwarte dane wytworzone przez urząd administracji publicznej (lub na jego zlecenie). Otwarte dane stanowią ogromną szansę na innowacje i rozwój, umożliwiając swobodny dostęp do informacji, co wspiera przejrzystość, współpracę oraz tworzenie nowych usług i produktów [10]. Sprawdzono, czy wygenerowane pliki z danymi z tej bazy są poprawne, m.in spójne, syntaktycznie poprawne oraz kompletne w kontekście relacyjności.

Zastosowane metody badawcze obejmują analizę, implementację, testowanie i ocenę procesu odtwarzania bazy danych na podstawie plików CSV pochodzących z GUS, przy uwzględnieniu zarówno technicznych, jak i jakościowych aspektów.

2. Analiza danych i implementacja

Proces analizy i implementacji bazy danych opierał się na przeanalizowaniu struktury danych, wykorzystaniu biblioteki Django do tworzenia modeli danych oraz skonfigurowaniu funkcjonalności importu danych z plików CSV do bazy danych w systemie zarządzania bazą danych SQLite. Django jest gotową biblioteką do tworzenia aplikacji internetowych w języku Python [8]. SQLite jest prostym system baz danych do celów deweloperskich. Dzięki takiemu podejściu uzyskano efektywną i wygodną bazę danych, która spełniała wymagania aplikacji.

2.1. Analiza struktury danych w plikach CSV

Proces analizy i implementacji bazy danych rozpoczęto od szczegółowej analizy struktury danych zawartych w plikach CSV. Podczas tego procesu zbadano kolumny, relacje między danymi oraz ustalono schematy danych. Następnie, zamiast bezpośredniej implementacji w aplikacjach bazodanowych, zdecydowano się na wykorzystanie bibliotek Django.

Framework Django został wybrany ze względu na jego wygodne narzędzia do pracy z bazą danych oraz możliwość szybkiego tworzenia modeli danych. W ramach tego frameworka utworzono modele odzwierciedlające strukturę danych z plików CSV.

Baza danych została utworzona w SQLite, wykorzystując możliwości tego systemu zarządzania bazą danych, takie jak elastyczność i łatwość w konfiguracji. Modele danych w Django zostały skonfigurowane zgodnie z analizą struktury danych z plików CSV (Tabela 1.), a także zostały uwzględnione relacje między nimi.

Następnie, zaimplementowano funkcjonalność importu danych z plików CSV do bazy danych przy użyciu Django. Dzięki temu użytkownik mógł łatwo zaimportować dane bezpośrednio z plików, co zapewniło elastyczność i wygodę w zarządzaniu danymi. W przypadku aktualizacji danych przez GUS, istnieje możliwość szybkiej aktualizacji w opracowywanym zbiorze.

W projekcie stosowano dane z plików TERC (Tabela 1.) oraz SIMC (Tabela 2.). Pliki te zawierają zestawy danych na temat polskich adresów. Zawartość tych plików przedstawiono na rysunkach Rysunek 1. oraz Rysunek 2.

Tabela 1. Spis analizowanych pól z pliku TERC pochodzącego z GUS

Plik TERC	
Pole	Opis
WOJ	Pole to zawiera ID województwa, wartości przechodzą co 2
POW	Pole to zawiera ID powiatu. Wartości mogą się powtarzać dla różnych województw
GMI	Pole to zawiera ID gminy. Wartości mogą się powtarzać dla różnych powiatów

RODZ	Pole to zawiera ID rodzaju gminy. Wartości te nie są wczytywane z pliku lecz manualnie wstawione w bazę danych.
NAZWA	Pole to zawiera nazwę danej jednostki podziału administracyjnego
NAZWA_DOD	Pole to zawiera poziom danej jednostki podziału administracyjnego
STAN_NA	Pole to zawiera datę ostatniej aktualizacji danych

Źródło: opracowanie własne.

Tabela 2. Spis analizowanych pól z pliku SIMC pochodzącego z GUS

Plik SIMC	
Pole	Opis
WOJ	Pole to zawiera ID województwa, wartości przechodzą co 2
POW	Pole to zawiera ID powiatu. Wartości mogą się powtarzać dla różnych województw
GMI	Pole to zawiera ID gminy. Wartości mogą się powtarzać dla różnych powiatów
RODZ	Pole to zawiera ID rodzaju gminy. Wartości te nie są wczytywane z pliku lecz manualnie wstawione w bazę danych.
RM	Pole to zawiera ID rodzaju miejscowości
MZ	Pole to zawiera wartości 1(NIE) lub 0(TAK) oznaczające to czy nazwa miejscowości jest zwyczajowa
NAZWA	Pole to zawiera nazwę danej miejscowości
SYM	Pole to zawiera ID miejscowości
SYMPOD	Pole to zawiera ID miejscowości podstawowej
STAN_NA	Pole to zawiera datę ostatniej aktualizacji danych

Źródło: opracowanie własne.

	A	B	C	D	E	F	G	H
1	WOJ	POW	GMI	RODZ	NAZWA	NAZWA_DOD	STAN_NA	
2	2				DOLNOŚLĄSKIE	województwo	2024-01-01	
3	2	1			bolesławiecki	powiat	2024-01-01	
4	2	1	1	1	Bolesławiec	gmina miejska	2024-01-01	
5	2	1	2	2	Bolesławiec	gmina wiejska	2024-01-01	
6	2	1	3	2	Gromadka	gmina wiejska	2024-01-01	
7	2	1	4	3	Nowogrodziec	gmina miejsko-wiejska	2024-01-01	
8	2	1	4	4	Nowogrodziec	miasto	2024-01-01	
9	2	1	4	5	Nowogrodziec	obszar wiejski	2024-01-01	
10	2	1	5	2	Osiecznica	gmina wiejska	2024-01-01	
11	2	1	6	2	Warta Bolesławiecka	gmina wiejska	2024-01-01	
12	2	2			dzierżoniowski	powiat	2024-01-01	
13	2	2	1	1	Bielawa	gmina miejska	2024-01-01	
14	2	2	2	1	Dzierżoniów	gmina miejska	2024-01-01	
15	2	2	3	3	Pieszyce	gmina miejsko-wiejska	2024-01-01	
16	2	2	3	4	Pieszyce	miasto	2024-01-01	
17	2	2	3	5	Pieszyce	obszar wiejski	2024-01-01	
18	2	2	4	1	Piława Górna	gmina miejska	2024-01-01	
19	2	2	5	2	Dzierżoniów	gmina wiejska	2024-01-01	
20	2	2	6	2	Łagiewniki	gmina wiejska	2024-01-01	
21	2	2	7	3	Niemcza	gmina miejsko-wiejska	2024-01-01	
22	2	2	7	4	Niemcza	miasto	2024-01-01	
23	2	2	7	5	Niemcza	obszar wiejski	2024-01-01	

Rysunek 1. Zawartość pliku TERC.
Źródło: opracowanie własne.

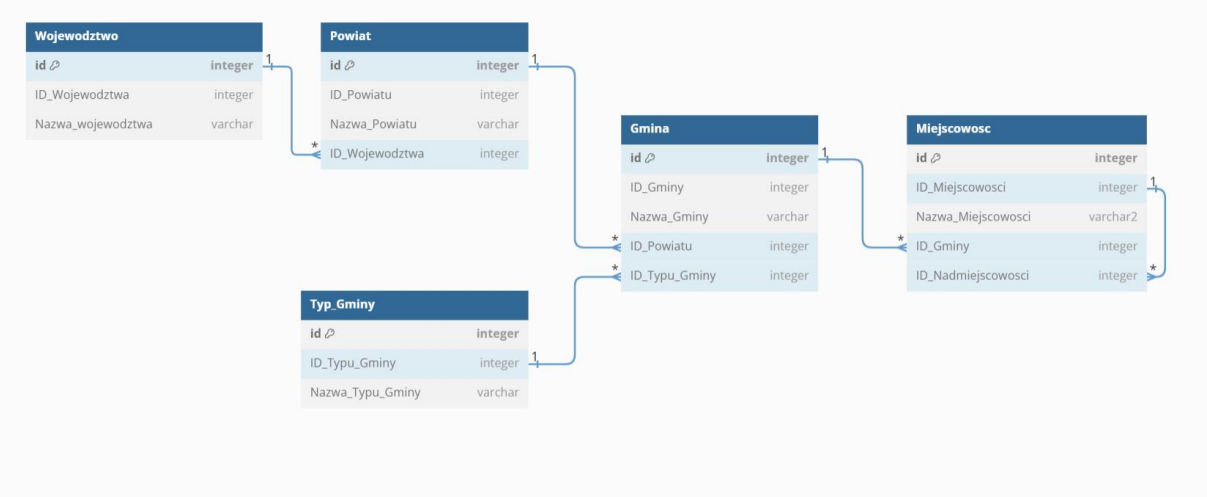
	A	B	C	D	E	F	G	H	I	J	K
1	WOJ	POW	GMI	RODZ_GMI	RM	MZ	NAZWA	SYM	SYMPOD	STAN_NA	
2	2	17	1	2	1	1	Brzezica	872792	872792	2024-01-01	
3	2	17	1	2	1	1	Kępino	872852	872852	2024-01-01	
4	2	18	5	2	1	1	Lasek	367918	367918	2024-01-01	
5	2	18	5	2	1	1	Łagiewniki Średzkie	367930	367930	2024-01-01	
6	2	18	5	2	1	1	Pielaszkowice	367960	367960	2024-01-01	
7	2	18	5	2	1	1	Sokolniki	367999	367999	2024-01-01	
8	2	19	3	2	1	1	Kłaczyna	852097	852097	2024-01-01	
9	2	8	9	2	1	1	Gołaczów	853257	853257	2024-01-01	
10	2	16	5	5	1	1	Krępa	366830	366830	2024-01-01	
11	2	16	5	5	1	1	Łężce	366853	366853	2024-01-01	
12	2	16	6	2	1	1	Sieroszowice	367090	367090	2024-01-01	
13	2	17	1	2	1	1	Borek Strzeliński	872770	872770	2024-01-01	
14	2	8	12	5	1	1	Karłów	854937	854937	2024-01-01	
15	2	4	4	5	1	1	Dochowa	377147	377147	2024-01-01	
16	2	4	4	5	1	1	Ługi	377302	377302	2024-01-01	
17	2	4	4	5	1	1	Pobiel	377348	377348	2024-01-01	
18	2	4	4	5	1	1	Wiklina	377443	377443	2024-01-01	
19	2	4	4	5	1	1	Wrząca Śląska	377466	377466	2024-01-01	
20	2	5	2	5	1	1	Jastrowiec	189380	189380	2024-01-01	
21	2	14	5	5	1	1	Ligota Rybińska	203654	203654	2024-01-01	
22	2	20	6	5	1	1	Kaszyce Milickie	884170	884170	2024-01-01	
23	2	20	6	5	1	1	Książęca Wieś	884223	884223	2024-01-01	
24	2	19	4	5	1	1	Bolesławice	852499	852499	2024-01-01	
25	2	19	4	5	1	1	Pasieczna	852542	852542	2024-01-01	

Rysunek 2. Zawartość pliku SIMC.

Źródło: opracowanie własne

2.2. Projektowanie bazy danych

Schemat bazy danych został wykonany przy pomocy przeglądarkowego narzędzia dbdiagram.io. Jest to aplikacja przeznaczona do projektowania schematów [10]. Utworzony schemat jest zoptymalizowany do formy, w której poszczególne wartości w tabelach nie są zbędnie powtarzane. Podczas realizacji uwzględniono normalizację, indeksowanie i optymalizację.



Rysunek 3. Schemat utworzonej bazy danych

Źródło: opracowanie własne

2.3. Implementację procesu importu danych

Opracowano i wdrożono proces importu danych z plików CSV do utworzonej bazy danych, uwzględniając konwersję formatów danych, walidację i mapowanie pól. Pseudkod zawarty w Listing 1 – 6 zawiera funkcje odwzorowania danych.

Pierwszym etapem jest załadowanie modelu zawierającego województwa (Listing 1). Następnie wgrywany jest model zawierający powiaty, który odwołuje się do modelu województw za pomocą kluczy obcych (Listing 2). Ostatecznym elementem wczytywanym z pliku TERC jest model zawierający gminy. Po wczytaniu następuje także weryfikacja całej operacji w celu sprawdzenia poprawności wczytania danych (Listing 3).

Listing 1. Wczytywanie pliku TERC oraz wgranie danych do modelu Wojewodztwo

```
    Wczytaj dane z pliku CSV
    Wybierz z pliku CSV tylko te rekordy gdzie kolumna 'POW' jest pusta a następnie zapisz do zmiennej przefiltrowane_wiersze
    Jeśli wiersz nie jest pusty:
        Utwórz pustą listę wojewodztwa_do_utworzenia
        Dla wszystkich wierszy:
                Utwórz obiekt zawierający:
                ID_Wojewodztwa = wartości kolumny 'WOJ'
                Nazwa_Wojewodztwa = wartości kolumny 'NAZWA'
                Dodaj obiekt do listy wojewodztwa_do_utworzenia
    Dodaj elementy listy wojewodztwa_do_utworzenia do modelu Wojewodztwo
```

Listing 2. Wgranie danych do modelu Powiat

```
    Wybierz z pliku CSV tylko te rekordy w których kolumna 'NAZWA_DOD' zawiera "powiat" a następnie nadpisz tym zmienną przefiltrowane_wiersze
    Jeśli wiersz nie jest pusty:
        Utwórz pustą listę powiaty_do_utworzenia
        Dla wszystkich wierszy:
                Utwórz obiekt zawierający:
                ID_Powiatu = wartości kolumny 'POW'
                Nazwa_Powiatu = wartości kolumny 'NAZWA'
                ID_Wojewodztwa = obiekt modelu Wojewodztwo gdzie pole ID_Wojewodztwa odpowiadają kolumnie 'WOJ'
                Dodaj obiekt do listy powiaty_do_utworzenia
    Dodaj elementy listy powiaty_do_utworzenia do modelu Powiat
```

Listing 3. Wczytanie danych do modelu Gmina

```
    Wybierz z pliku CSV tylko te rekordy w których kolumna 'NAZWA_DOD' zawiera "gmina" i gdzie 'GMI' nie jest nullem a następnie nadpisz tym zmienną przefiltrowane_wiersze
    Jeśli wiersz nie jest pusty:
        Utwórz pustą listę gminy_do_utworzenia
        Dla wszystkich wierszy:
                Utwórz obiekt zawierający:
                ID_Gminy = wartości kolumny 'GMI'
                Nazwa_Gminy = wartości kolumny 'NAZWA'
                ID_Powiatu = obiekt modelu Powiat gdzie pole ID_Powiatu odpowiada kolumnie 'POW' i pole ID_Wojewodztwa odpowiada kolumnie 'WOJ'
                ID_Typu_Powiatu = obiekt modelu Powiat odpowiadający kolumnie 'RODZ'
                Dodaj obiekt do listy gminy_do_utworzenia
    Dodaj elementy listy gminy_do_utworzenia do modelu Gmina
```

Listing 4. Przygotowanie danych pliku SIMC

```
Wczytaj dane z pliku CSV
Wybierz z pliku CSV tylko te rekordy w których 'RODZ_GMI' różny od 8 i 9 a następnie zapisz do zmiennej przefiltrowane_wiersze
Wybierz z przefiltrowane_wiersze wszystkie wiersze gdzie kolumna 'SYM' = 'SYMPOD' i zapisz je do zmiennej przefiltrowane_wiersze_1
Wybierz z przefiltrowane_wiersze wszystkie wiersze gdzie kolumna 'SYM' != 'SYMPOD' i zapisz je do zmiennej przefiltrowane_wiersze_2
```

Listing 5. Wgranie podstawowych miejscowości do modelu Miejscowosc

```
Jeśli wiersz z przefiltrowane_wiersze_1 nie jest pusty:
    Utwórz pustą listę miejscowosci_do_utworzenia
    Dla wszystkich wierszy:
        Utwórz obiekt zawierający
        ID_Miejscowosci = wartości kolumny 'SYM'
        Nazwa_Miejscowosci = wartości kolumny 'NAZWA'
        ID_Gminy = obiekt modelu powiat gdzie ID_Gminy odpowiada kolumnie 'GMI' i pole ID_Powiatu odpowiada kolumnie 'POW' i pole ID_Wojewodztwa odpowiada kolumnie 'WOJ'
        ID_Podmiejscowosci = null
    Dodaj obiekt do listy miejscowosci_do_utworzenia
Dodaj elementy listy miejscowosci_do_utworzenia do modelu Miejscowosc
```

Listing 6. Wgranie miejscowosci nie podstawowych do modelu Miejscowosc

```
Jeśli wiersz z przefiltrowane_wiersze_2 nie jest pusty:
    Utwórz pustą listę miejscowosci_do_utworzenia_2
    Dla wszystkich wierszy:
        Utwórz obiekt zawierający
        ID_Miejscowosci = wartości kolumny 'SYM'
        Nazwa_Miejscowosci = wartości kolumny 'NAZWA'
        ID_Gminy = obiekt modelu powiat gdzie ID_Gminy odpowiada kolumnie 'GMI' i pole ID_Powiatu odpowiada kolumnie 'POW' i pole ID_Wojewodztwa odpowiada kolumnie 'WOJ'
        ID_Podmiejscowosci = obiekt modelu Miejscowosc odpowiadający kolumnie 'SYMPOD'
    Dodaj obiekt do listy miejscowosci_do_utworzenia_2
Dodaj elementy listy miejscowosci_do_utworzenia_2 do modelu Miejscowosc
```

Jeśli wczytywanie danych się powiodło:

Zwróć informację czy wczytywanie danych się powiodło

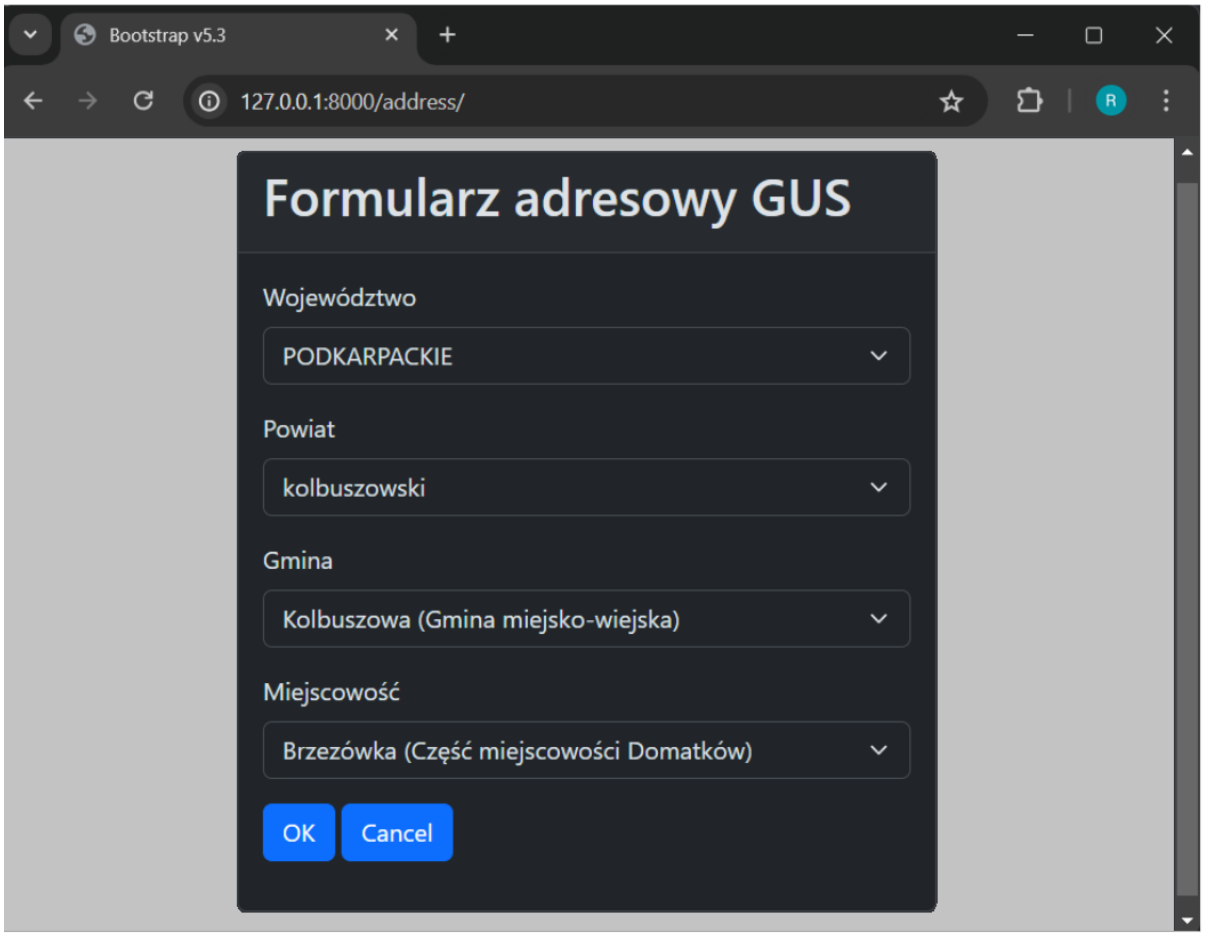
W przeciwnym razie

Zwróć błąd

3. Testowanie i weryfikacja

Testowanie i weryfikację osiągniętych rezultatów sprawdzono poprzez utworzenie aktywnego formularza, w którym użytkownik może wybierać adresy w sposób dynamiczny i zhierarchizowany. Formularz z polami wyboru został napisany w języku HTML z wykorzystaniem biblioteki Django.

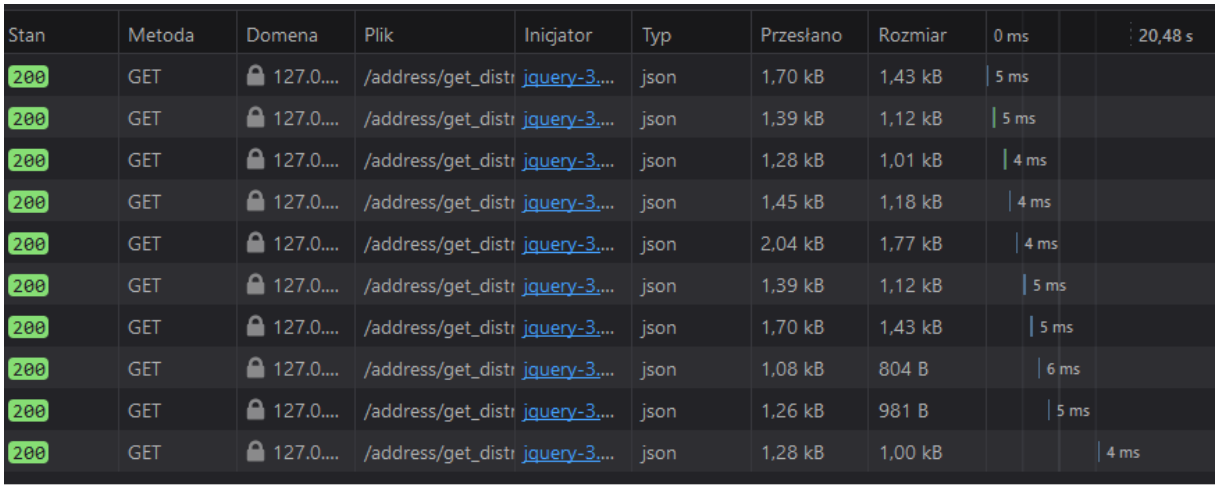
Przeprowadzono testy manualne i zweryfikowano poprawności importu oraz spójności danych w odtworzonej bazie danych. Sprawdzono również użyteczność oraz stabilność rozwiązania.



Rysunek 4. Utworzony formularz testowy do dynamicznej selekcji elementów składowych adresu fizycznego
Źródło: opracowanie własne.

3.2. Ocena wydajności

Analiza wydajności operacji odczytu w odtworzonej bazie danych podczas wybierania zadanych pól została dokonana za pomocą konsoli przeglądarki (Rysunek 5.). W Tabeli 3 zawarto średnie wyniki. Średnie wartości czasów zapytań wyniosły od 4.7 ms do 6.2 ms w zależności od rodzaju pobieranych obiektów.



Rysunek 5. Czas odpowiedzi serwera na żądanie pobrania listy powiatów dla danego województwa.
Źródło: opracowanie własne.

Tabela 3. Średnie czasy żądań HTTP dla zadanych list obiektów

Rodzaj obiektów	Min [ms]	Max [ms]	Średnia [ms]
Powiat	4	6	4.7
Gmina	4	11	5.3
Miejscowość	4	11	6.2

Źródło: opracowanie własne.

3.3. Ustalenie dokładności

Określenie dokładności odtworzonej bazy danych poprzez porównanie wyników z bazą danych źródłową dokonano poprzez sprawdzenie liczby danych ze względu na ich rodzaj: województwo, miasto, gmina.

Pakiet Django pozwala na zliczanie elementów w konkretnych tabelach co pozwoliło na łatwą walidację dokładności odwzorowania pod względem ilościowym. Ustalono, że baza została zaimportowana w pełni i bez błędów pod względem ilościowym.

Dodatkowo, w celu weryfikacji poprawności powiązań elementów adresów zastosowano manualne sprawdzenie losowo wybranych 20 próbek danych. Próba ta polegała na szczegółowej analizie każdego elementu w kontekście jego zgodności z innymi powiązanymi elementami. Przeprowadzone sprawdzenie nie wykazało żadnych błędów ani niespójności, co potwierdza poprawność powiązań w badanym zbiorze danych.

Podsumowanie

Celem projektu było utworzenie relacyjnej bazy danych na podstawie danych adresowych udostępnionych przez Główny Urząd Statystyczny (GUS) w formie plików CSV. Proces obejmował inżynierię odwrotną, która umożliwiła pozyskanie koncepcyjnych schematów danych i ich przetworzenie do postaci relacyjnej bazy danych. Utworzona baza danych wspiera dynamiczne wybieranie adresów przez użytkowników na podstawie predefiniowanych kryteriów, co zapewnia szybkie generowanie list wyboru i pozytywny odbiór systemu (UX). System został zaprojektowany z naciskiem na integralność i spójność danych, co jest kluczowe w kontekście ich późniejszego importu do systemu raportowania projektów państwowych. Testowanie systemu przepływu danych pozwoliło na wyeliminowanie błędów i zapewnienie wysokiej jakości danych. Badanie potwierdziło, że pliki CSV z GUS są syntaktycznie poprawne, spójne i kompletne w kontekście relacyjności, co umożliwia efektywne ich przetwarzanie i wykorzystanie. Otwarte dane stanowią dużą szansę na innowacje, wspierając przejrzystość oraz tworzenie nowych usług i produktów.

Literatura

[1] Hainaut, J. L., Henrard, J., Roland, D., Hick, J. M., & Englebert, V. (2009). Database reverse engineering. In Handbook of Research on Innovations in Database Technologies and Applications: Current and Future Trends (pp. 181-189). IGI Global

[2] Paredaens, J., De Bra, P., Gyssens, M., & Van Gucht, D. (2012). The structure of the relational database model (Vol. 17). Springer Science & Business Media.

[3] Date, C. J. (1989). A Guide to the SQL Standard. Addison-Wesley Longman Publishing Co., Inc..

[4] Hartson, R., & Pyla, P. S. (2012). The UX Book: Process and guidelines for ensuring a quality user experience. Elsevier.

[5] Duggineni, S. (2023). Data Integrity and Risk. Open Journal of Optimization, 12(2), 25-33.

[6] Dul, M., GUGAŁA, Ł., & Kamil, Ł. A. B. A. (2023). Protecting web applications from authentication attacks. Advances in Web Development Journal, 1(1).

[7] Cerrudo, C., & Fayo, E. M. (2007). Hacking databases for owning your data.

[8] Rubio, D. (2017). Beginning Django. Apress.

[9] Su, T., Wu, K., Miao, W., Pu, G., He, J., Chen, Y., & Su, Z. (2017). A survey on data-flow testing. ACM computing surveys (CSUR), 50(1), 1-35.

[10] Martin, S., Foulonneau, M., Turki, S., & Ihadjadene, M. (2013, June). Open data: Barriers, risks and opportunities. In Proceedings of the 13th European Conference on eGovernment: ECEG (pp. 301-309).

Źródła internetowe

[11] https://www.dbdiagram.io/ (dostęp: 09.05.2024)

[12] https://stat.gov.pl/ (dostęp: 09.03.2024).

KOŁO

NAUKOWE

SYSTEMÓW

ZŁOŻONYCH

Justyna Mazur
Koło Naukowe Ubezpieczeń

Dr Anna Ostrowska-Dankiewicz
Opiekun Koła Naukowego Ubezpieczeń

Analiza polskiego rynku brokerów ubezpieczeniowych i reasekuracyjnych

Streszczenie

Celem artykułu jest analiza rynku brokerskiego w Polsce. W części teoretycznej wyjaśniono pojęcia broker ubezpieczeniowy oraz broker reasekuracyjny. Przedstawiono zadania wykonywane przez brokerów. Wskazano również kto nie może prowadzić działalności brokerskiej. Wyjaśniono kto może uzyskać wpis w rejestrze brokerów, a także wskazano, kiedy broker może zostać wykreślony z rejestru brokerów. W części badawczej dokonano analizy rynku brokerów, obejmującą liczbę brokerów ubezpieczeniowych reasekuracyjnych, wydawane zezwolenia, liczbę wykreśleń z rejestru brokerów oraz średnie roczne przychody uzyskiwane przez brokerów z tytułu prowizji od zakładów ubezpieczeń. W analizie wykorzystano raporty o stanie rynku brokerskiego z KNF w latach 2015-2022. Brokerzy prowadzący działalność reasekuracyjną stanowią ułamek osób w stosunku do brokerów ubezpieczeniowych. Pomimo dużej różnicy pod względem liczebności brokerów, to brokerzy reasekuracyjni osiągają wyższe przychody liczone w mln zł.

Słowa kluczowe: broker ubezpieczeniowy, broker reasekuracyjny, KNF.

1. Wprowadzenie

Na rynku ubezpieczeniowym w ramach wykonywania czynności na rzecz podmiotów poszukujących ochrony ubezpieczeniowej wyróżnia się agentów oraz brokerów. Należy zaznaczyć, że agenci ubezpieczeniowi współpracują z konkretnymi zakładami ubezpieczeń, co może mieć swoje odzwierciedlenie w rekomendacjach dotyczących ofert ubezpieczenia. Inaczej wygląda kwestia w przypadku brokerów. Są oni niezależnymi pośrednikami ubezpieczeniowymi, którzy działają na rzecz osób potrzebujących i szukających ochrony ubezpieczeniowej. Przepisy prawa wykluczają możliwość działania na rzecz zakładów ubezpieczeń. Aspektem, który może budzić kontrowersje jest natomiast otrzymywanie przez brokerów kurtażu, jako formy wynagrodzenia od zakładów ubezpieczeń w przypadku zawarcia umowy między klientem, a danym ubezpieczycielem.

Celem artykułu jest dokonanie analizy rynku brokerów ubezpieczeniowych i reasekuracyjnych. Został on zrealizowany poprzez przedstawienie najważniejszych informacji o brokerach ubezpieczeniowych oraz brokerach reasekuracyjnych, obejmujących analizę zmiany liczby brokerów w Polsce, liczby wydawanych zezwoleń dla brokerów, liczby

wykreśleń z rejestru brokerów, a także przychodów uzyskiwanych w ramach premii od zakładów ubezpieczeń dla brokerów w Polsce w latach. Badanie zostało przeprowadzone na podstawie raportów Komisji Nadzoru Finansowego w latach 2015-2022.

2. Istota oraz zadania brokera ubezpieczeniowego i reasekuracyjnego

Broker to pośrednik ubezpieczeniowy, który dokonuje czynności w imieniu lub na rzecz podmiotu szukającego ochrony ubezpieczeniowej[1].

Broker ubezpieczeniowy jest niezależnym doradcą pośredniczącym pomiędzy klientem a towarzystwem ubezpieczeniowym. Jego zadaniem jest wspieranie klienta poprzez rekomendację najlepiej dopasowanego ubezpieczenia, które zapewni przy najmniejszym wkładzie jak najwięcej korzyści[2]. Brokerem ubezpieczeniowym może zostać osoba fizyczna bądź osoba prawna, znajdująca się w rejestrze brokerów oraz mająca zezwolenie wydane przez organ nadzoru dotyczące wykonywania działalności w zakresie ubezpieczeń[3].

Broker reasekuracyjny pośredniczy pomiędzy zakładem ubezpieczeń i zakładem reasekuracji w procesie związanym z podpisaniem umowy reasekuracyjnej. Głównym założeniem umowy jest wymiana oraz podział ryzyka, ale również eliminacja ryzyka powstania strat w zakładach ubezpieczeń[4]. Brokerem reasekuracyjnym może zostać osoba fizyczna lub osoba prawna, która została wpisana do rejestru brokerów, a także posiada zezwolenie wydane przez organ nadzoru dotyczące wykonywania działalności brokerskiej w zakresie reasekuracji[5].

Do podstawowych zadań brokerów ubezpieczeniowych zalicza się między innymi[6]:

- analizę zagrożeń i ich optymalizację,
- pozyskiwanie potencjalnych klientów,
- porównywanie oraz rekomendowanie ofert klientom,
- pomoc klientom w uzyskaniu odszkodowań w przypadku zajścia zdarzenia objętego ubezpieczeniem,
- zawieranie, opracowywanie oraz przekazywanie umów ubezpieczenia klientom.

[1] Raport o stanie rynku brokerskiego w 2022 roku, UKNF, Warszawa 2023, s. 4.
[2] https://businessinsider.com.pl/poradnik-finansowy/ubezpieczenia/kim-jest-broker-ubezpieczeniowy/fwldnd1 (10.06.2024 r.).
[3] Ustawa z dnia 15 grudnia 2017 r. o dystrybucji ubezpieczeń (Dz. U. z 2023 r., poz. 1111, z późn. zm.).
[4] https://www.praca.pl/poradniki/rynek-pracy/broker-reasekuracyjny-kompetencje,zakres-pracy,wynagrodzenie_pr-5301.html (dostęp: 08.06.2024).
[5] Ustawa z dnia 15 grudnia 2017 r. o dystrybucji ubezpieczeń (Dz. U. z 2023 r., poz. 1111, z późn. zm.).
[6] *Informacja o zawodzie – Broker ubezpieczeniowy* (332104), Wydawnictwo Naukowe Instytutu Technologii Eksploatacji – PIB, Warszawa 2018, s. 5.

Przekrój zadań brokerów ubezpieczeniowych jest szeroki, rozpoczyna się od momentu pozyskiwania klientów, poprzez rekomendowanie i zawieranie umowy po wsparcie klientów w uzyskaniu odszkodowania.

Działalność brokerów reasekuracyjnych związana jest między innymi z[7]:

- oceną ryzyka oraz analizą zagrożeń,
- prowadzeniem negocjacji warunków umowy reasekuracyjnej pomiędzy ubezpieczycielem i reasekuratorem,
- działalnością informacyjną odnośnie dostępnych form i rodzajów reasekuracji,
- przygotowaniem oraz obsługiwaniem umów reasekuracji,
- współpracą z instytucjami nadzoru finansowego.

Zadania brokerów reasekuracyjnych dotyczą różnorodnych zagadnień. Od dokonywania oceny ryzyka i analizy zagrożeń, poprzez negocjacje warunków umowy reasekuracyjnej jako pośrednik między ubezpieczycielem a reasekuratorem, po czynności przygotowania i obsługiwanie zawartej umowy reasekuracji. Istotnym aspektem jest również współpraca brokerów z instytucjami nadzoru finansowego.

Decyzja o prowadzeniu działalności brokerskiej wiąże się z pewnymi ograniczeniami. Do wyłączeń zalicza się przede wszystkim: pozostawanie w stałym stosunku z zakładem ubezpieczeń oraz zakładem reasekuracji (wyłączenie nie dotyczy umowy ubezpieczenia poprzez, którą broker ubezpieczeniowy jest ubezpieczony lub ubezpieczającym); wykonywanie działalności agencyjnej. Dodatkowo wyłączenie obejmuje wykonywanie czynności agencyjnych; bycie posiadaczem akcji, udziałów zakładów ubezpieczeń oraz zakładów reasekuracji, a także akcji i udziałów (wyjątek od reguły stanowią akcje, które zostały dopuszczone do obrotu na rynku regulowanym), a także uczestnictwa w organach nadzorczych albo zarządzających zakładu ubezpieczeń, zakładu reasekuracji[8].

Wpisanie do rejestru brokerów wiąże się z koniecznością zdania egzaminu przeprowadzanego przez Komisję Egzaminacyjną dla Brokerów Ubezpieczeniowych i Reasekuracyjnych. Przeprowadzany jest on średnio trzy razy do roku[9]. Egzamin regulowany jest przepisami rozporządzenia Ministra Finansów z dnia 23 kwietnia 2019 r. w sprawie

[7] *Informacja o zawodzie – Broker reasekuracyjny* (332103), Wydawnictwo Naukowe Instytutu Technologii Eksploatacji – PIB, Warszawa 2018, s. 5.
[8] Ustawa z dnia 15 grudnia 2017 r. o dystrybucji ubezpieczeń (Dz. U. z 2023 r., poz. 1111, z późn. zm.).
[9] https://www.knf.gov.pl/dla_rynku/egzaminy/brokerzy (dostęp:09.05.2024).

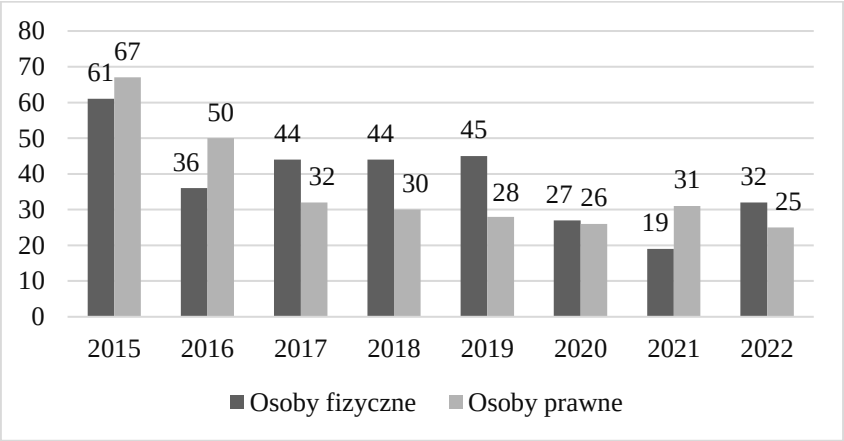
egzaminu dla brokerów ubezpieczeniowych i reasekuracyjnych oraz Komisji Egzaminacyjnej dla Brokerów Ubezpieczeniowych i Reasekuracyjnych[10].

Wykreślenie z rejestru brokerów może być spowodowane: wygaśnięciem zezwolenia, wnioskiem brokera o wykreślenie z rejestru, cofnięciem zezwolenia na skutek niespełnienia wymogów prowadzenia działalności lub naruszenia przepisów prawa[11].

3. Metodyka i wyniki badań

Celem przeprowadzonego badania była analiza rynku brokerskiego obejmującego brokerów ubezpieczeniowych oraz brokerów reasekuracyjnych w latach 2015-2022. Analiza dotyczyła liczby brokerów w Polsce, liczby wydawanych zezwoleń dla brokerów, liczy wykreśleń z rejestru brokerów, a także przychodów uzyskiwanych w ramach premii od zakładów ubezpieczeń dla brokerów.

W 2024 roku zaplanowano trzy terminy egzaminów dla brokerów ubezpieczeniowych i reasekuracyjnych. Wydawane zezwolenia dla brokerów ubezpieczeniowych z podziałem na osoby fizyczne i prawne w latach 2015-2022 zaprezentowano na wykresie 1.



Wykres 1. Wydane zezwolenia dla brokerów ubezpieczeniowych w latach 2015-2022

Źródło: opracowanie na podstawie danych KNF: Raport o stanie rynku brokerskiego w 2015 roku, UKNF, Warszawa 2016, Raport o stanie rynku brokerskiego w 2016 roku, UKNF, Warszawa 2017, Raport o stanie rynku brokerskiego w 2017 roku, UKNF, Warszawa 2018, Raport o stanie rynku brokerskiego w 2018 roku, UKNF, Warszawa 2019, Raport o stanie rynku brokerskiego w 2019 roku, UKNF, Warszawa 2020, Raport o stanie rynku brokerskiego w 2020 roku, UKNF, Warszawa 2021, Raport o stanie rynku brokerskiego w 2021 roku, UKNF, Warszawa 2022, Raport o stanie rynku brokerskiego w 2022 roku, UKNF, Warszawa 2023.

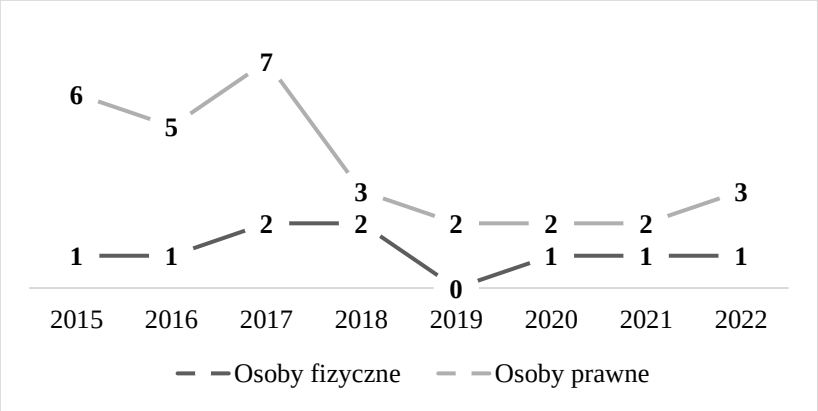
W analizowanym okresie największą liczbę zezwoleń dla brokerów wydano w 2015 roku, która wynosiła 128 z czego zgodę otrzymało 61 osób fizycznych i 67 osób prawnych. Od 2016 roku widoczna była tendencja spadkowa wydawanych zgód dla osób prawnych. Trend

[10] Rozporządzenie Ministra Finansów z dnia 23 kwietnia 2019 r. w sprawie egzaminu dla brokerów ubezpieczeniowych i reasekuracyjnych oraz Komisji Egzaminacyjnej dla Brokerów Ubezpieczeniowych i Reasekuracyjnych (Dz. U. z 2019 r., poz. 879).

[11] Raport o stanie rynku brokerskiego w 2017 roku, UKNF, Warszawa 2018, s. 8.

utrzymywał się do 2020 roku. Porównując dane z 2015 i 2020 roku liczba zezwoleń zmniejszyła się o 61%. W przypadku osób fizycznych liczba zgód wydanych w 2016 roku znacznie spadła w stosunku do roku 2015 (obniżenie o 41%). W latach 2017-2019 liczba zezwoleń utrzymywała się na podobnym poziomie (44-45). Sumarycznie najmniej zezwoleń dla brokerów wydano w 2021 roku, i było to odpowiednio 19 dla osób fizycznych i 31 dla osób prawnych.

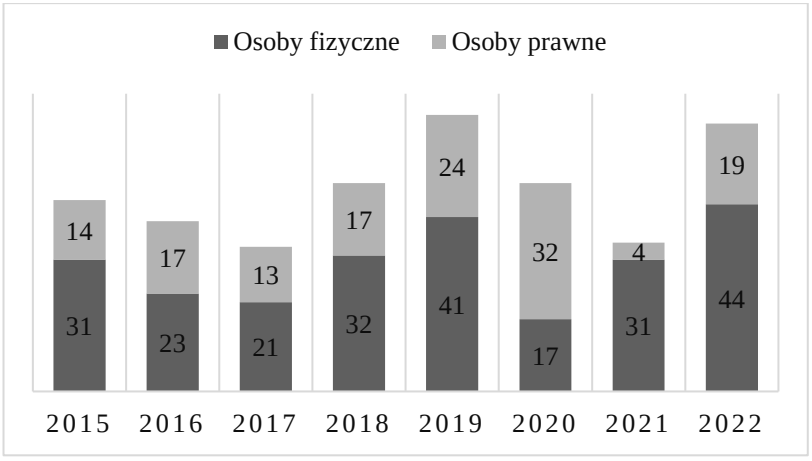
Wydawane zezwolenia dla brokerów reasekuracyjnych z podziałem na osoby fizyczne i prawne w latach 2015-2022 zaprezentowano na wykresie 2.



Wykres 2. Wydane zezwolenia dla brokerów reasekuracyjnych w latach 2015-2022
Źródło: opracowanie na podstawie danych KNF…, op. cit.

Liczba brokerów reasekuracyjnych uzyskująca pozwolenie na wykonywanie działalności brokerskiej w badanym okresie była stosunkowo niewielka. Osobom prawnym w porównaniu do osób fizycznych wydawano w badanym okresie więcej zezwoleń. W 2015 roku wydano osobom prawnym 6 zezwoleń, w roku kolejnym liczba ta zmniejszyła się o 1. Najwięcej takich zezwoleń wydano w 2017 roku, w liczbie 7. W roku kolejnym zanotowano spadek liczby o 57% w stosunku do roku ubiegłego. W latach 2019-2021 wydawano po dwa zezwolenia na prowadzenie działalności brokerskiej. W 2022 roku liczba zezwoleń wróciła do poziomu z 2018 roku. Osoby fizyczne otrzymały 1-2 zezwolenia na prowadzenie działalności brokerskiej w badanym okresie. Wyjątek stanowi 2019 rok, gdzie nie wydano żadnego zezwolenia.

Brokerzy mogą zostać wykreśleni z rejestru ze względu na niespełnienie wymogów prawnych, wygaśnięcie pozwolenia, a także mogą również złożyć wniosek o wykreślenie. Wykreślenie osób fizycznych i prawnych z rejestru brokerów ubezpieczeniowych i reasekuracyjnych w latach 2015-2022 zaprezentowano na wykresie 3.



Wykres 3. Wykreślenia brokerów ubezpieczeniowych i reasekuracyjnych w latach 2015-2022.
Źródło: opracowanie na podstawie danych KNF…, op. cit.

W analizowanym okresie liczba wykreśleń osób fizycznych i osób prawnych wahała się. W latach 2017-2020 widoczna była tendencja wzrostowa wykreślanych osób prawnych z rejestru brokerów (wzrost o 146% w skrajnych latach). W 2020 roku wykreślono największą liczbę osób prawnych w wysokości 32 osób. Natomiast w 2021 roku wykreślono zaledwie 4 brokerów, którzy byli osobami prawnymi. Zanotowano tym samym spadek o 87,5% w stosunku do roku 2020. Największą liczbę osób fizycznych wykreślono z rejestru brokerów w roku 2022 i roku 2019, pod względem liczbowym było to odpowiednio 44 przypadki i 41 przypadków. W roku 2020 liczba wykreśleń zmniejszyła się z poziomu 41 do 17 osób prawnych, co oznacza obniżenie o 58,5% decyzji.

Prowadzenie działalności brokerskiej wiąże się z koniecznością zadania egzaminu. Liczbę brokerów ubezpieczeniowych oraz reasekuracyjnych oraz zmiany procentowe liczebności brokerów jakie zachodziły na przestrzeni lat 2015-2022 zaprezentowano w tabeli 1.

Tabela 8.Liczba brokerów oraz zmiany jakie zaszły na przestrzeni lat 2015-2022

	2015	2016	2017	2018	2019	2020	2021	2022
brokerzy ubezpieczeniowi	1276	1323	1367	1392	1405	1411	1426	1420
zmiany		3,68%	3,33%	1,83%	0,93%	0,43%	1,06%	-0,42%
brokerzy reasekuracyjni	39	42	49	54	53	54	57	58
zmiany		7,69%	16,67%	10,20%	-1,85%	1,89%	5,56%	1,75%
razem	1315	1365	1416	1446	1458	1465	1483	1478
zmiany		3,80%	3,74%	2,12%	0,83%	0,48%	1,23%	-0,34%

Źródło: opracowanie na podstawie danych KNF…, op. cit.

W latach 2015-2021 zauważalny był trend wzrostowy liczby brokerów ubezpieczeniowych i reasekuracyjnych. W 2015 roku zarejestrowanych było 1276 brokerów ubezpieczeniowych. W 2016 roku zanotowano spory przyrost liczby brokerów ubezpieczeniowych wynoszący 3,68% w stosunku do roku ubiegłego. Nieco mniejszy wzrost zanotowano natomiast w 2017

roku wynoszący 3,33%. W 2020 roku liczba brokerów ubezpieczeniowych mogących prowadzić działalność brokerską wzrosła zaledwie o 6 osób (0,43%) w stosunku do roku 2019. Po raz pierwszy spadek zanotowano w roku 2022 liczba brokerów ubezpieczeniowych spadła o 0,42% w stosunku do 2021 roku. Brokerzy zarejestrowani w rejestrze brokerów stanowili znacznie mniejszą grupę w porównaniu do brokerów ubezpieczeniowych. Dlatego też zmiany procentowe zachodzące na przestrzeni lat były bardziej widoczne. W 2016 roku zanotowano wzrost o 7,39% w stosunku do 2015 roku, natomiast w 2017 udział ten wzrósł o 16,67% w stosunku do roku ubiegłego (7 osób). W 2022 roku liczba brokerów reasekuracyjnych wzrosła o 48,7% w stosunku do 2015 roku. Sumarycznie w 2022 roku uprawnienia brokerskie posiadało 1478 osób co w stosunku do początkowego oznacza wzrost o 12,78%.

Brokerzy uzyskują prowizję od zakładów ubezpieczeń. Średnią roczną wysokość przychodów uzyskanych z tytułu prowizji od zakładów ubezpieczeniowych przez brokerów w latach 2015-2022 zaprezentowano w tabeli 2.

Tabela 9. Średnie roczne przychody z tytułu prowizji od zakładów ubezpieczeń przypadające na brokera w latach 2015-2022

	2015	**2016**	**2017**	**2018**	**2019**	**2020**	**2021**	**2022**
Osoby fizyczne	219955	233785	250722	337385	293659	327999	335166	399497
Zmiany		6,29%	7,24%	34,57%	-12,96%	11,69%	2,19%	19,19%
Osoby prawne	2402753	2447933	2457345	2343377	2287358	2444816	2791682	3273429
Zmiany		1,88%	0,38%	-4,64%	-2,39%	6,88%	14,19%	17,26%

Źródło: opracowanie na podstawie danych KNF…, op. cit.

W latach 2015-2018 średnie roczne przychody uzyskiwane przez osoby fizyczne z tytułu prowizji od zakładów ubezpieczeń wzrastały rok do roku. W 2016 roku zanotowano 1,06-krotny wzrost w stosunku do roku 2015. Rok 2018 okazał się rekordowy pod względem przychodów uzyskanych z prowizji w stosunku do 2017 roku. Osoby fizyczne osiągnęły średnio 337 385 zł przychodu w przeliczeniu na osobę, co oznacza wzrost na poziomie 34,57%. W 2019 roku zanotowano spadek przychodów wynoszący 12,96% (obniżenie o 43 726 zł). Od 2020 roku średnie przychody brokerów będących osobami fizycznymi wzrastają. W 2015 roku osoby prawne generowały średnie przychody wyższe o 11 razy w stosunku do osób fizycznych. W 2016 i 2017 roku wzrost rok do roku był stosunkowo niewielki i wynosił odpowiednio 1,88% oraz 0,38%. Spadek przychodów z tytułu prowizji zanotowano w latach 2017-2018. Porównujące średnie roczne przychody do 2017 roku spadły one w 2018 roku o 4,64%, zaś w 2019 roku o 6,92%. W 2022 roku roczne przychody uzyskiwane przez brokerów były

najwyższe w badanym okresie. Osoby fizyczne zarabiały średnio 399 tys. zł, z kolei osoby prawne 3,273 mln zł.

4. Podsumowanie

Działalność gospodarcza w zakresie ubezpieczeń i reasekuracji prowadzona przez brokerów jest bardzo istotna w gospodarce. Ich zadaniem, jako niezależnych pośredników finansowych jest w sposób obiektywny doradzić oraz pomóc wybrać najbardziej dopasowaną ofertę do potrzeb klienta.

Działalność brokerów regulowana jest przepisami prawa. Brokerami mogą zostać zarówno osoby fizyczne jak i prawne, by uzyskać licencję na prowadzenie działalności ubezpieczeniowej bądź reasekuracyjnej konieczne jest zdanie egzaminu organizowanego przez KNF. Nabycie uprawnień i wpisanie do rejestru brokerów nie jest dożywotnie, między innymi w przypadku nieprzestrzegania przepisów prawa brokerzy mogą zostać z niego wykreśleni.

W Polsce w latach 2015-2022 widoczna jest duża dysproporcja pomiędzy brokerami ubezpieczeniowymi i reasekuracyjnymi, zarówno pod względem liczby osób, jak i pod względem osiąganych przychodów. W 2022 roku zaledwie 3,92% osób fizycznych i prawnych stanowili brokerzy reasekuracyjni. Porównując ich udział w 2015 roku, brokerzy reasekuracyjni stanowili jedynie 2,97% brokerów ogółem, co stanowi różnicę na poziomie 0,95 pp.

Pomimo większej liczby brokerów ubezpieczeniowych średnie roczne przychody uzyskiwane przez brokerów reasekuracyjnych przyjmują znacznie wyższe wartości liczone w mln złotych. W przypadku brokerów ubezpieczeniowych mowa o przychodach pomiędzy 200 tys. zł, a 400 tys. złotych.

Egzaminy na brokerów przeprowadzane są kilka razy do roku. Na przestrzeni lat widoczny jest trend spadkowy liczby wydawanych zezwoleń dla brokerów ubezpieczeniowych przez KNF. W 2015 roku wydano ich 128 natomiast w 2021 roku zezwolenia otrzymało zaledwie 50 osób (obniżenie o 60,93%). Brokerzy reasekuracyjni w badanym okresie otrzymywali znacznie mniej zezwoleń. Najwięcej wydano w 2017 roku, gdy 9 brokerów otrzymało zezwolenia z czego 2 dotyczyły osób fizycznych natomiast 7 osób prawnych. W 2019 roku żadna osoba fizyczna nie uzyskała wpisu do rejestru brokerów. W 2022 roku wykreślono więcej brokerów niż zostało przyznanych licencji.

Rynek brokerów ubezpieczeniowych i reasekuracyjnych ma potencjał rozwoju, co widać ~~go~~ przede wszystkim w przychodach osiąganych przez brokerów, w szczególności brokerów reasekuracyjnych.

Literatura

1. *Informacja o zawodzie – Broker reasekuracyjny* (332103), Wydawnictwo Naukowe Instytutu Technologii Eksploatacji – PIB, Warszawa 2018.
2. *Informacja o zawodzie – Broker ubezpieczeniowy* (332104), Wydawnictwo Naukowe Instytutu Technologii Eksploatacji – PIB, Warszawa 2018.
3. Raport o stanie rynku brokerskiego w 2015 roku, UKNF, Warszawa 2016.
4. Raport o stanie rynku brokerskiego w 2016 roku, UKNF, Warszawa 2017.
5. Raport o stanie rynku brokerskiego w 2017 roku, UKNF, Warszawa 2018.
6. Raport o stanie rynku brokerskiego w 2018 roku, UKNF, Warszawa 2019.
7. Raport o stanie rynku brokerskiego w 2019 roku, UKNF, Warszawa 2020.
8. Raport o stanie rynku brokerskiego w 2020 roku, UKNF, Warszawa 2021.
9. Raport o stanie rynku brokerskiego w 2021 roku, UKNF, Warszawa 2022.
10. Raport o stanie rynku brokerskiego w 2022 roku, UKNF, Warszawa 2023.

Akty normatywne

1. Rozporządzenie Ministra Finansów z dnia 23 kwietnia 2019 r. w sprawie egzaminu dla brokerów ubezpieczeniowych i reasekuracyjnych oraz Komisji Egzaminacyjnej dla Brokerów Ubezpieczeniowych i Reasekuracyjnych (Dz. U. z 2019 r., poz. 879).
2. Ustawa z dnia 15 grudnia 2017 r. o dystrybucji ubezpieczeń (Dz. U. z 2023 r., poz. 1111, z późn. zm.).

Źródła internetowe

1. https://businessinsider.com.pl/poradnik-finansowy/ubezpieczenia/kim-jest-broker-ubezpieczeniowy/fwldnd1 (10.06.2024).
2. https://www.praca.pl/poradniki/rynek-pracy/broker-reasekuracyjny-kompetencje,zakres-pracy,wynagrodzenie_pr-5301.html (dostęp: 08.06.2024).
3. https://www.knf.gov.pl/dla_rynku/egzaminy/brokerzy (dostęp:09.05.2024).

KOŁO

NAUKOWE

STUDENTÓW

CHEMII

„Esprit”

inż. Anna Rybka
Koło Naukowe Studentów Chemii ESPRIT

Prof. dr hab. inż. Wiktor Bukowski
Opiekun Koła Naukowego

Kompozyty aramidowe odporne na płomień

Streszczenie

W poniższym artykule zaprezentowano możliwości wykorzystania włókien aramidowych w materiałach i kompozytach, które są odporne na działanie płomieni i mają bardzo dobre właściwości w wysokich temperaturach. Opisano charakterystykę aramidów i metody otrzymywania kompozytów aramidowych. Wytypowano modyfikatory ognioodporności, zwiększające właściwości wytrzymałościowe materiału oraz omówiono kierunki zastosowań tego typu kompozytów w przemyśle lotniczym, samochodowym oraz w produkcji kombinezonów przeciwpożarowych.

Słowa kluczowe: włókna aramidowe, kompozyty, ognioodporność, Nomex, Kevlar

1. Wprowadzenie

W obecnych czasach, w których pożary i ekstremalne temperatury stanowią stałe zagrożenie, kompozyty aramidowe pełnią kluczową funkcję w zapewnianiu bezpieczeństwa i ochrony życia ludzkiego. Ich wyjątkowe cechy czynią je nowatorskim rozwiązaniem w dziedzinie prewencji i zwalczania pożarów.

Główną cechą, która wyróżnia kompozyty aramidowe jest ich wyjątkowa odporność na wysoką temperaturę. Włókna aramidowe ze względu na swoją wytrzymałość termiczną są idealnym materiałem do zastosowań wymagających ochrony przed płomieniem i wysoką temperaturą. Dodatkowo, kompozyty aramidowe są odporne na wiele substancji chemicznych. W związku z tym są szeroko wykorzystywane w przemyśle ochrony przeciwpożarowej, przemyśle lotniczym, motoryzacyjnym, a także wielu innych dziedzinach.

Celem artykułu jest przedstawienie właściwości mechanicznych i technologicznych kompozytów aramidowych, metod otrzymywania oraz wytypowanie modyfikatorów ognioodporności zwiększających właściwości użytkowe materiału.

2. Charakterystyka aramidów

Termin "aramid" odnosi się do włókien z rodzaju poliamidu aromatycznego, w których co najmniej 85% wiązań amidowych (-CO-NH-) jest bezpośrednio przyłączonych do dwóch pierścieni aromatycznych. Zamiana alifatycznego szkieletu węglowego na grupy aromatyczne powoduje znaczące zmiany w właściwościach tych włókien[1].

[1]Ertekin M., *Aramid fibers,* Fiber Technology for Fiber-Reinforced Composites, Woodhead Publishing Series in Composites Science and Engineering, 2017 p. 2-12

Pierwszym opracowanym włóknem tej klasy był Nomex, wprowadzony przez firmę DuPont w latach 60. XX wieku. Chociaż ma ono umiarkowaną wytrzymałość na rozciąganie, to charakteryzuje się niepalnością i jest szeroko stosowane w produkcji odzieży ognioodpornej. Kilka lat później firma DuPont wprowadziła na rynek włókna aramidowe Kevlar, które zawierają łańcuchy z *p*-dipodstawionymi pierścieniami benzenowymi. Oprócz doskonałej stabilności termicznej, te włókna wykazują również wybitne właściwości mechaniczne. Ich wyjątkowy potencjał wynika przede wszystkim z anizotropii ich struktury, która zawiera cechy włókniste, krystaliczne i rdzeniowe[2].

Główne cechy włókien aramidowych obejmują doskonałą odporność na wysoką temperaturę, wytrzymałość na ścieranie i odporność na działanie substancji chemicznych. Te włókna nie przewodzą prądu elektrycznego, nie topią się, ale ulegają degradacji w temperaturze przekraczającej 500°C. W szczególności tzw. meta-aramidy (Nomex), wyróżniają się wysoką wytrzymałością na rozciąganie, umożliwiając wytrzymywanie naprężeń rozciągających, ścierania i substancji chemicznych, nawet pod wpływem ognia i temperatur sięgających 400°C. Włókna meta-aramidowe mogą zatrzymać wilgoć w ilości wynoszącej 5%, a ich wydłużenie przy zerwaniu wynosi 15%. Z kolei tzw. para-aramidy charakteryzują się znaczną wytrzymałością na rozciąganie oraz wysokim modułem elastyczności. Moduł ten opisuje zdolność materiału do odkształcenia sprężystego pod wpływem przyłożonej siły. Włókna para-aramidowe są dostępne w różnych odmianach, w tym o niskim, wysokim i bardzo wysokim module elastyczności. Właściwości niektórych z tych odmian przedstawiono w poniższej Tabeli 1[3].

Włókna aramidowe posiadają również pewne wady, gdyż są to materiały zbudowane z silnie ustalonych domen krystalicznych, które często zawierają defekty i puste przestrzenie. Ze względu na obecność dużej benzenowej struktury pierścieniowej w łańcuchu molekularnym, grupa amidowa jest mało reaktywna w stosunku do innych atomów lub grup funkcyjnych. Materiał ten wykazuje niską kompatybilność z matrycą z powodu swojej obojętności chemicznej. Dlatego konieczne jest dokonanie modyfikacji włókna aramidowego przed jego zastosowaniem w celu utworzenia kompozytu[4].

Tabela 10 Właściwości fizyczne wybranych włókien para-aramidowych. Źródło: Opracowanie własne na podstawie [3].

Włókna	Wytrzymałość na rozciąganie [N/tex]	Moduł sprężystośći [N/tex]	Wydłużenie przy zerwaniu [%]
Kevlar 29	2	490	3,6

[2]Kirona M. I., Aramid Fibers: Types, Properties, Manufacturing Process and Applications, University of Management And Technology Lahore, 2020

[3]Deopura B.L., Padaki N.V., Chapter 5 – Synthetic Textile Fibres: Polyamide, Polyester and Aramid Fibres, Textiles and Fashion. Materials, Design and Technology, Woodhead Publishing Series in Textiles, 2015, p. 12

[4] Wu K., Wang X., Xu Y., Guo W., Flame retardant efficiency of modified para-aramid fiber synergizing with ammonium polyphosphate on PP/EPDM, Polymer Degradation and Stability, 2020, p. 1-12

Kevlar 49	2,1	780	2,8
Kevlar 149	2,1	1000	2

Włókno aramidowe występuje w formie nici, które można prząść w różnego rodzaju tkaniny techniczne. Może być dostępne również jako włóknina, czyli materiał nietkany, używany do filtracji oraz izolacji termicznej[5]. Zdjęcie przykładowej tkaniny aramidowej przestawiono na Rysunku 1.

Rysunek 5 Tkanina aramidowa. Źródło [5].

3. Otrzymywanie kompozytów aramidowych

Kompozyt to materiał złożony z co najmniej dwóch różnych składników. Tworzy się go poprzez połączenie dwóch lub więcej materiałów, które posiadają znacznie odmienne właściwości. Te różne materiały współpracują ze sobą, ale nie ulegają wzajemnemu rozpuszczeniu ani mieszaniu, co nadaje kompozytowi wyjątkowe właściwości. Kompozyty są tworzone przez selektywne łączenie odpowiednich wzmacniaczy, takich jak na przykład włókna.

3.1 Metoda prepregu

Prepeg to prosta forma kompozytu, w której matryca i włókno są impregnowane razem, tworząc rolki. Proces wytwarzania rozpoczyna się od włókien wychodzących z różnych szpul i osadzanych na cienkich warstwach materiału matrycowego w taki sposób, aby uzyskać pożądaną frakcję włókien i poziom przyczepności poprzez różne cykle zagęszczania i ogrzewania[6].

[5] https://holtex.pl/blog/tkanina-aramidowa-zastosowanie-i-wlasciwosci (dostęp 5.06.2024)
[6] Hasan Z., Chapter 2 – Composite materials, Tooling for Composite Aerospace Structures, Manufacturing and Applications, 2020, p. 7-9

3.2 Formowanie przetłoczne żywicy wspomagane próżniowo

Żywica poliestrowa (epoksydowa, fenolowa, poliuretanowa) ze względu na niskie ciśnienie generowane przez pompę próżniową o niskiej lepkości mieszana jest z włóknami szklanymi. W wyniku tego procesu uzyskuje się kompozyty, w których udział objętościowy włókien wynosi od 40-50%[7].

3.3 Nawijanie włókna

Technika ta wykorzystuje pasma włókien, które są impregnowane w sposób ciągły lub przerywany i przepuszczaniu ich przez pojemnik wypełniony roztworem żywicy. Włókna nawijane są na obracający się trzpień[8].

3.4 Pultruzja

To proces, w którym pakiet włókien wzmocnionych i impregnowanych żywicą jest przeciągany przez podgrzany blok matrycowy, na którym zachodzi polimeryzacja. Technika jest bardzo szybka w porównaniu z innymi wymienionymi metodami[9].

3.5 Metoda kontaktowa (ręczna)

Metoda kontaktowa polega na układaniu zbrojenia w postaci maty lub tkaniny w formie i nasączaniu jej mieszaniną żywicy i utwardzacza z wykorzystaniem gumowego wałka. Po nasączeniu jednej warstwy mieszaniną żywicy i utwardzacza, nakłada się kolejne, aż do pożądanej ilości zbrojenia[9].

4. Modyfikatory ognioodporności kompozytów aramidowych

W celu opóźnienia lub zahamowania procesu spalania polimeru do matrycy polimerowej dodawane środki zmniejszające palność, zarówno działające na zasadzie fizycznej (rozrzedzanie paliwo, chłodzenie lub karbonizacja), jak i chemiczne. Halogenowe środki zmniejszające palność skutecznie redukują poziom wolnych rodników podczas spalania. Jednak ze względu na emisję toksycznych gazów i dymów podczas spalania, ich stosowanie jest ograniczone z uwagi na negatywny wpływ na zdrowie ludzi i środowisko. Mineralne środki zmniejszające palność stanowią dobrą alternatywę dla halogenowych środków. Działają one na zasadzie rozkładu podczas spalania i wydzielania gazów obojętnych, takich jak para wodna i dwutlenek węgla. Te gazy rozcieńczają palną mieszankę węglowodorów i tlenu oraz spowalniają proces spalania. Dodatkowo, w wyniku rozkładu powstają

[7] Dharmavarapu P., Reddy S., Aramid fibre as potential reinforcement for polymer matrix composites: a review, emergent mater. 5, 2021, p. 1-18,
[8] Prakash M. Gore and Balasubramanian Kandasubramanian, Functionalized Aramid Fi-bers and Composites for Protective Applications: A Review, Industrial and Engineering Chemistry Research 57, 2018, p. 50-60
[9] Boczkowska A., Krzesiński G., Kompozyty i techniki ich wytwarzania, Oficyna Wydaw-nicza Politechniki Warszawskiej, 2016, p. 75-76

tlenki metali, które zmniejszają lepkość matrycy polimerowej i zapobiegają tworzeniu się kropelek, co ogranicza rozprzestrzenianie się ognia[10].

Jako mineralne środki zmniejszające palność stosuje się na przykład wodorotlenki magnezu i wodorotlenek glinu. Stosowane są w układach polimerowych ze względu na właściwość endotermicznego odwodnienia pod wpływem ciepła i ognia. Dodatek 35-55% wagowych wodorotlenku magnezu znacznie zmniejsza palność nienasyconej żywicy poliestrowej. Jednakże wysoki poziom obciążenia wodorotlenkami metali w celu osiągnięcia odpowiedniej ognioodporności może niekorzystnie zmniejszyć właściwości mechaniczne osnów polimerowych[11].

W przypadku modyfikatorów stosowanych obecnie występują ograniczenia w zastosowaniu halogenowych substancji zmniejszających palność, ze względu na wydzielanie silnie toksycznego dymu i związane z tym zanieczyszczenie środowiska podczas procesu spalania. Dlatego coraz częściej sięgamy po nowe substancje opóźniające proces spalania, które nie zawierają halogenków. W szczególności, substancje te obejmują bezhalogenowe modyfikatory zmniejszające palność, takie jak związki fosforu i krzemu. Ich popularność wynika z łatwego dostępu oraz prostych procesów syntetycznych, jakie oferują.

Istnieje także możliwość wykorzystania polifosforanu amonu (APP), dwuwarstwowych wodorotlenków glinu lub magnezu, grafitu ekspandowanego oraz związków fosforoorganicznych jako środków zmniejszających palność. Polifosforan amonu jest używany w przypadku nienasyconych żywic poliestrowych. APP, będący związkiem zawierającym zarówno fosfor, jak i azot, może zwiększać odporność ogniową w procesie spalania i zwiększać wartość LOI (z ang. Limiting Oxygen Index) przy stosunkowo niewielkich ilościach[12].

Możliwe jest dokonanie modyfikacji włókien aramidowych poprzez wykorzystanie substancji pęczniejących zmniejszających łatwopalność, znanych jako IFR (ang. Intumescent Flame Retardants). Te substancje charakteryzują się ekologicznym podejściem oraz wykazują doskonałe cechy, takie jak brak zawartości halogenków, niska toksyczność oraz zdolność do zapobiegania topnieniu. Zazwyczaj, system IFR składa się z trzech elementów: źródła kwasu, źródła węgla i źródła gazu. W tradycyjnym systemie IFR, źródłem kwasu jest polifosforan amonu, źródłem węgla jest pentaerytrytol, a źródłem gazu jest melamina. Mechanizm działania tych substancji opiera się na tworzeniu warstwy węgla. Zwęglina jest substancją stałą o gęstej strukturze porowatej, składającej się z aromatycznych

---

[10] Gunes O., Gomek R., Tamar A., Kandemir O. A., Karaorman A., Albayrak A. Z., Com-parative Study on Flame Retardancy, Thermal, and Mechanical Properties of Glass Fiber Reinforced Polyester Composites with Ammonium Polyphosphate, Expandable Graphite, and Aluminum Tri-hydroxide, Arabian Journal for Science and Engineering 43, 2018, p. 1-3

[11] Rajaei M., Wang D., Bhattacharyya D., Combined effects of ammonium polyphosphate and talc on the fire and mechanical properties of epoxy/glass fabric composites, Compo-sites Part B: Engineering, 2017, p. 1

[12] Chen Z., Jiang M., Zhang Q., Yuan Yu, Sun G. i Jianga J., Synergistic Effect of Combined Dimethyl Methyiphosphonate with Aluminum Hydroxide or Ammonium Polyphosphate Retardants Systems on the Flame Retardancy and Thermal Properties of Unsaturated Polyester Resin, International Journal of Polymer Analysis and Characterization, 2017 p. 1-3

węglowodorów policyklicznych. Stopniowe uporządkowanie tych węglowodorów wzrasta wraz ze wzrostem temperatury. Proces tworzenia zwęgliny wydaje się mieć kluczowe znaczenie dla procesu spalania. Warstwa zwęgliny pełni rolę bariery, ochraniając materiał przed wpływem zewnętrznego promieniowania cieplnego i stabilizując warstwę węgla, co zapobiega jej utlenianiu do tlenku lub ditlenku węgla[13].

Modyfikacja kompozytów jest również możliwa przez zastosowanie soli melaminy, które dodawane są do matrycy polimerowej, wpływając na właściwości kompozytu polimerowego. Istotna jest powierzchnia kontaktu fazy rozproszonej oraz charakter oddziaływań pomiędzy fazą ciągłą i rozproszoną. Gdy wzrasta wskaźnik kształtu wypełniacza i zmniejsza się jego wymiar poprzeczny, zwiększa się jego powierzchnia właściwa, skutkuje to większymi całkowitymi siłami interakcji pomiędzy matrycą polimerową a cząstkami wypełniacza[14].

5. Zastosowanie kompozytów aramidowych

Współcześnie, materiały polimerowe wzmocnione włóknami znajdują zastosowanie w produkcji części samolotów, struktur morskich i pojazdów samochodowych, mając na celu zmniejszenie całkowitej masy, redukcję emisji CO_2 oraz zwiększenie efektywności energetycznej.

Kompozyty aramidowe stosowane są również do zastosowań wzmacniających, takich jak kordy opon lub elementy samochodowe takie jak okładziny, sprzęgła, uszczelki oraz inne. Tkanina wykonana z Nomexu może być stosowana jako odzież ochronna dla strażaków lub jako izolacja elektryczna. Kevlar, który jest wykorzystywany do otrzymywania kompozytów aramidowych ma również zastosowanie takie jak produkcja odpornych na przecięcia rękawic ochronnych oraz warstw pokrowców na siedzenia. Ze względu na dobre właściwości dielektryczne, odporność na korozję oraz odporność cieplną nadają się do produkcji lin, kabli oraz lin cumowniczych[15].

6. Podsumowanie

Kompozyty aramidowe, dzięki swojej wyjątkowej odporności na wysokie temperatury i działanie substancji chemicznych, odgrywają kluczową rolę w zapewnianiu bezpieczeństwa i ochrony życia ludzkiego w obliczu pożarów i ekstremalnych temperatur. Aramidy, takie jak Nomex i Kevlar, wykazują doskonałe właściwości mechaniczne i termiczne, które sprawiają, że są idealne do zastosowań

[13] Mizera K., Celiński M., Kozikowski P., Borucka M., Przybysz J, Gajek A., Wpływ antypi-renów fosforowych na palność pianek poliizocyjanurowych, Centralny Instytut Ochrony Pracy- Państwowy Instytut Badawczy, 2023, p. 2

[14] Kużdżał E., Cichy B., Kicko-Walczak B, Rymarz G, Rheological and fire properties of a composite of unsaturated polyester resin and halogen-free flame retardants, Journal of Applied Polymer Science, 2016, p. 1-4

[15] Akato K., Bhat G., 10 - High performance fibers from aramid polymers, Structure and Properties of High-Performance Fibers, Woodhead Publishing, 2017, p. 1-18

w odzieży ochronnej, przemyśle lotniczym, motoryzacyjnym oraz wielu innych dziedzinach. Włókna aramidowe charakteryzują się doskonałą wytrzymałością na ścieranie, odpornością na działanie chemikaliów, nie przewodzą prądu elektrycznego i nie topią się, ale ulegają degradacji powyżej 500°C. Meta-aramidy (np. Nomex) są odporne na rozciąganie i wysokie temperatury, a para-aramidy (np. Kevlar) mają wysoką wytrzymałość na rozciąganie i moduł elastyczności.

Otrzymywanie kompozytów aramidowych odbywa się poprzez różne metody, takie jak metoda prepregu, formowanie przetłoczne żywicy wspomagane próżniowo, nawijanie włókna, pultruzja oraz metoda kontaktowa. Każda z tych metod umożliwia tworzenie kompozytów o specyficznych właściwościach, odpowiednich do różnych zastosowań. Aby zwiększyć ognioodporność kompozytów aramidowych, stosuje się modyfikatory ognioodporności, takie jak halogenowe i mineralne środki zmniejszające palność.

Z uwagi na toksyczność halogenów, coraz częściej używa się bezhalogenowych modyfikatorów, takich jak związki fosforu i krzemu, polifosforan amonu, IFR (intumescent flame retardants) oraz sole melaminy. Kompozyty aramidowe znajdują szerokie zastosowanie w produkcji części samolotów, struktur morskich, pojazdów samochodowych, odzieży ochronnej, lin, kabli oraz innych elementów, które wymagają wysokiej wytrzymałości, odporności na ścieranie, działanie chemikaliów i wysokie temperatury.

Literatura

1. Akato K., Bhat G., *10 - High performance fibers from aramid polymers,* Structure and Properties of High- Performance Fibers, Woodhead Publishing, 2017, p. 1-18
2. Boczkowska A., Krzesiński G., *Kompozyty i techniki ich wytwarzania*, Oficyna Wydawnicza Politechniki Warszawskiej, 2016, p. 75-76
3. Chen Z., Jiang M., Zhang Q., Yuan Yu, Sun G. i Jianga J., *Synergistic Effect of Combined Dimethyl Methyiphosphonate with Aluminum Hydroxide or Ammonium Polyphosphate Retardants Systems on the Flame Retardancy and Thermal Properties of Unsaturated Polyester Resin,* International Journal of Polymer Analysis and Characterization, 2017 p. 1-3
4. Deopura B.L., Padaki N.V., Chapter 5 – *Synthetic Textile Fibres: Polyamide, Polyester and Aramid Fibres, Textiles and Fashion. Materials, Design and Technology, Woodhead Publishing Series in Textiles*, 2015, p. 12
5. Dharmavarapu P., Reddy S., *Aramid fibre as potential reinforcement for polymer matrix composites: a review*, emergent mater. 5, 2021, p. 1-18,
6. Ertekin M., *Aramid fibers*, Fiber Technology for Fiber-Reinforced Composites, Woodhead Publishing Series in Composites Science and Engineering, 2017 p. 2-12,
7. Gunes O., Gomek R., Tamar A., Kandemir O. A., Karaorman A., Albayrak A. Z., *Comparative Study on Flame Retardancy, Thermal, and Mechanical Properties of Glass Fiber Reinforced Polyester*

Composites with Ammonium Polyphosphate, Expandable Graphite, and Aluminum Tri-hydroxide, Arabian Journal for Science and Engineering 43, 2018, p. 1-3

8. Hasan Z., Chapter 2 – *Composite materials*, Tooling for Composite Aerospace Structures, Manufacturing and Applications, 2020, p. 7-9
9. Kużdżał E., Cichy B., Kicko-Walczak B, Rymarz G, *Rheological and fire properties of a composite of unsaturated polyester resin and halogen-free flame retardants, Journal of Applied Polymer Science*, 2016, p. 1-4
10. Kirona M. I., *Aramid Fibers: Types, Properties, Manufacturing Process and Applications*, University of Management And Technology Lahore, 2020
11. Łunarski J., *Zarzadzanie jakością. Standardy i zasady*, Wydawnictwo Naukowo-Techniczne, 2008
12. Mizera K., Celiński M., Kozikowski P., Borucka M., Przybysz J, Gajek A., *Wpływ antypirenów fosforowych na palność pianek poliizocyjanurowych*, Centralny Instytut Ochrony Pracy- Państwowy Instytut Badawczy, 2023, p. 2
13. Prakash M. Gore and Balasubramanian Kandasubramanian, *Functionalized Aramid Fibers and Composites for Protective Applications: A Review*, Industrial and Engineering Chemistry Research 57, 2018, p. 50-60
14. Rajaei M., Wang D., Bhattacharyya D., *Combined effects of ammonium polyphosphate and talc on the fire and mechanical properties of epoxy/glass fabric composites*, Composites Part B: Engineering, 2017, p. 1
15. Wu K., Wang X., Xu Y., Guo W., *Flame retardant efficiency of modified para-aramid fiber synergizing with ammonium polyphosphate on PP/EPDM*, Polymer Degradation and Stability, 2020, p. 1-12

Źródła internetowe

1. https://holtex.pl/blog/tkanina-aramidowa-zastosowanie-i-wlasciwosci (dostęp 5.06.2024)

Wiktoria Słąba
Koło Naukowe Esprit

dr inż. Dorota Głowacz-Czerwonka, prof. PRz [1]

Metody uniepalniania sztywnych pianek poliuretanowych

Streszczenie

Celem artykułu był przegląd wiedzy na temat sposobów uniepalniania sztywnych pianek poliuretanowych poprzez zastosowanie antypirenów reaktywnych i addytywnych.

Słowa kluczowe: Sztywne pianki poliuretanowe, uniepalniacze, antypireny

1. Wprowadzenie

Wytwarzane pianki poliuretanowe pomimo niskich współczynników przenikania ciepła (co potwierdza ich przydatność jako materiałów termoizolacyjnych), nie spełniają wymagań testów ognioodporności, ze względu na swoją rozwiniętą powierzchnię i organiczny charakter matrycy polimerowej. co sprawia że, są palne.[2] Zwiększające się z każdym rokiem wymagania przepisów dotyczących bezpieczeństwa pożarowego materiałów budowlanych skłaniają do ciągłych modyfikacji pianek poliuretanowych. Zwiększenie odporności na płomień pianek poliuretanowych ma na celu zapewnienie bezpieczeństwa użytkowników.

Efekt taki jest możliwy do uzyskania poprzez: wprowadzanie reaktywnych lub addytywnych uniepalniaczy, zmianę struktury komórek lub pokrywanie powierzchni pianki odpowiednimi retardantami. Istotą uniepalniania jest zmniejszenie palności, przy zachowaniu parametrów kompozycji referencyjnej, tj. uzyskanej bez udziału antypirenu.

W niniejszym opracowaniu skupiono się głównie na metodzie addytywnego uniepalniania.

2. Sposoby uniepalniania

Spośród wielu prób uniepalniania materiałów poliuretanowych najczęściej stosuje się antypireny. Istotnym zagadnieniem jest wykorzystywanie związków, które nie zawierają fluorowców (halogenów) w swojej budowie, ponieważ podczas spalania wydzielają się silnie toksyczne gazy i dymy zagrażające życiu ludzi.

[1]Politechnika Rzeszowska im. Ignacego Łukasiewicza, Katedra Chemii Organicznej.
[2] Prociak A., Rokicki G., Ryszkowska J., Materiały poliuretanowe, PWN, Warszawa 2016, s 11-15; 162-184.

Zazwyczaj w przypadku pianek poliuretanowych (PPUR) ograniczenia palności uzyskuje się poprzez synergiczne działanie uniepalniaczy [3]. Odporność PPUR na ogień uzyskuje nowatorsko na dwa sposoby. Prostszym jest połączenie konwencjonalnych uniepalniaczy zmniejszających palność sztywnych pianek poliuretanowych (SPPUR) na drodze mieszania fizycznego (antypireny addytywne). Drugą jest syntezowanie reaktywnych uniepalniaczy wbudowanych w strukturę poliolu, posiadających różne grupy funkcyjne (antypireny reaktywne) [4].

2.1.Melamina

Tanim środkiem zmniejszających palność jest (chętnie stosowana) melamina, którą charakteryzuje wysoka zawartość azotu i nietoksyczność. Jej wadą jest słaba dyspersja w matrycy PU, co powoduje trudności z spełnieniem wymagań wielkości cząsteczek i dystrybucji. Konwencjonalne środki fizyczne powodują dużą agregację i szybkie opadanie (sedymentację) proszku uniepalniacza prowadząc do niezadowalającej ognioodporności i pogorszenia właściwości mechanicznych PPUR. Poliol pochodzący z melaminy (MADP) wykazuje kompatybilność, którą można zastąpić częściowo polieter, gdyż może kopolimeryzować z izocyjanianem tworząc na bazie melaminy SPPUR. Otrzymana pianka wystawiona na działanie ciepła będzie się rozkładać na N_2, NH_3, rozcieńczając palne gazy.

Tak przygotowaną piankę można połączyć z 9,10-dihydro-9-oksa-10-fosfafenantren-10-tlenek (DOPO), jest to związek zawierający reaktywne wiązania P-H, co pozwala na kowalencyjne związanie tego środku ze strukturą grup –NCO SPPUR. Nadaje to wysoko skutecznej ognioodporności, stabilności i małego prawdopodobieństwa wymywania z matrycy polimerowej. Powoduje to podczas spalania pojawienie się kwasu polifosforowego, który katalizuje powstanie ochronnej warstwy zwęglania.

Poliol pochodzący z melaminy połączony z DOPO poprawia kompatybilność i właściwości termiczne SPPUR. Dodatkowo tego typu pianki charakteryzują się niewielką wagą, dobrym właściwością mechanicznym i doskonałym termiczno-izolacyjnych [5].

---

[3] Czech-Polak J., Oliwa R., Oleksy M., Budzik G., Sztywne *pianki poliuretanowe o zwiększonej odporności na płomień*, Wydawnictwo Sieć Badawcza Łukasiewicz – Instytut Chemii Przemysłowej, T: 63 nr 2,Warszawa 2018, s. 115-124.

[4] Yao Yuan, Chao Ma, Yongqian Shi, Lei Song, Yuan Hu, Weizhao Hu, *Highly-efficient reinforcement and flame retardancy of rigid polyurethane foam with phosphorus-containing additive andnitrogen-containing compound*, Materials Chemistry and Physics V: 211, s.44, ELSEVIER 2018.

[5] Dz. Cyt. Yao Yuan *Highly-efficient reinforcement...* s.44.

2.2.Chitosan

Można również zastosować powlekanie wyrobów niepalnymi powłokami izolacyjnymi. Materiały biomasowe znalazły w ostatnim czasie duże zainteresowanie, wynikające z ich biodegradowalności , nietoksycznej natury i biokompatybilności. Takim biopolimerem jest Chitosan (CS) uzyskiwany z chityny znajdującej się w muszlach skorupiaków. Zmniejsza palność, łatwo się odwodnia i pod wpływem kwasów tworzy się na jego powierzchni warstwa węglowa. Taka warstwa ogranicza uwalnianie ciepła i dymu podczas palenia. Dodatkowo podczas spalania wydzielają się amoniak i woda, jako para wodna. Powoduje to rozrzedzenie stężenia gazów palnych i zmniejszenie temperatury spalania [6].

2.3.Fosfor czarny

W ostatnich latach na rynku wyłonił się fosfor czarny (BP) jako antypiren. Jest to unikalna odmiana alotropowa fosforu z warstwową strukturą krystaliczną, która wykazuje intrygujące właściwości. Między innymi wysoka mobilność przenośników, silna anizotropia czy dostrajana przerwa energetyczna. Wykazuje dużą powierzchnię właściwą i wybitne właściwości katalityczne. Ponadto jest mało toksyczny, przyjazny dla środowiska i wykazujewysoką efektywność zmniejszania palenia. Działa synergicznie z innymi antypirenami, co może bardziej poprawiać efektywność. Odkryto, że należy dodać niewielką ilość nanocząsteczek BP do polimeru aby osiągnąć odpowiedni standard ognioodporności bez znacznych wpływów na właściwości mechaniczne.

Występują natomiast pewne problemy podczas używania BP jako pojedynczy środek uniepalniający. Są nimi słaba stabilność środowiskowa, ograniczona poprawa właściwości mechanicznych i ognioodpornych. Dlatego podjęto się modyfikacji w celu funkcjonalizacji BP, poprawiając jego odporność na degradację i zwiększając skuteczność uniepalniania, jak również poprawiając równomierne zdyspergowanie [7].

2.4.Polifosforan amonu

Uniepalniacze zawierające fosfor i azot cechują się dużą skutecznością zmniejszania palności, dobrą zdolnością zwęglania i małą toksycznością, dlatego wykorzystywane są przy produkcji polimerów na szeroką skalę. Podczas spalania związki fosforu katalizują procesy zwęglania lub działają jako zmiatacze rodników. Natomiast cząsteczki zawierające azot uwalniają obojętne produkty, które rozcieńczają gazy podczas pirolizy.

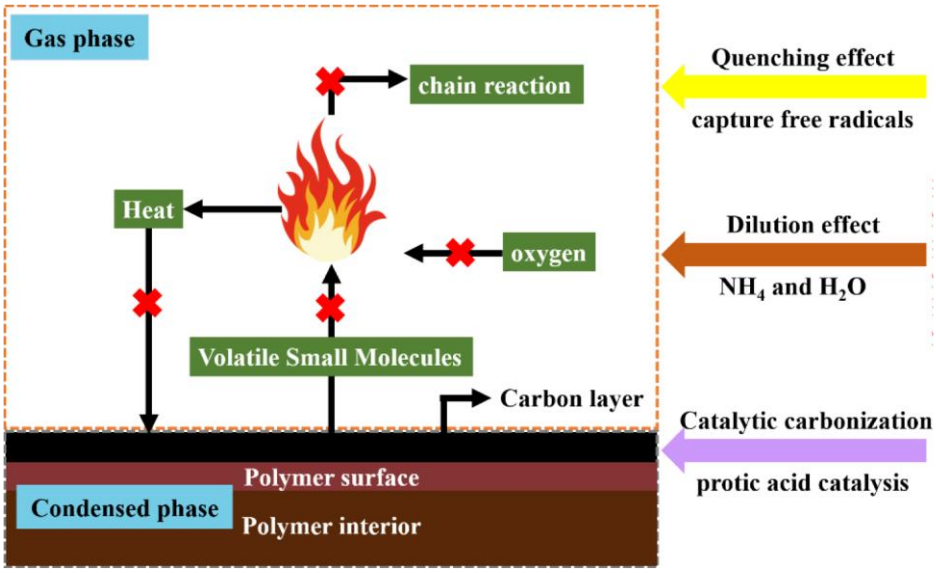
---

6 Dz. Cyt. Sihao Yin, *Surface coating of biomass...* s. 1-5,
[7] Dz. Cyt. Sihao Yin, *Surface coating of biomass...* s. 1-5,

Jednym z najczęściej używanych związków jest polifosforan amonu (APP) [8]. Zazwyczaj są połączone z uniepalniaczami fosforowymi, w celu usprawnienia niepalności PPUR w fazie gazowej i stałej. Rozwiązując problem kompatybilności, stosuje się mikrokapsułkowanie powierzchni APP. [9]

2.5.Grafit ekspandowalny

Innym uniepalniaczem dodawanym addytywnie jest grafit ekspandowalny (EG), jego efekt zależy od ilości i rozmiaru. Podczas spalania EG rozszerza się tworząc wermikularną strukturę zwęglania redukującą przenikanie ciepła. Jednak podczas dodawania go do PUR pogarsza właściwości mechaniczne pianki przez małą adhezję międzyfazową z matrycą, co może prowadzić do zapadania i zderzania się komórek. Podczas rozbicia cząsteczek EGznacznie zmniejsza się ognioodporność. W celu ominięcia tej wady poszukuje się innych uniepalniaczy częściowo zastępujących EG, co prowadzi do zmniejszenia używanych dawek. Podczas spalania tworzy w warstwie stałej warstwę zwęglania podobną do „robaczków", uzyskując w ten sposób efekt bariery fizycznej [10].



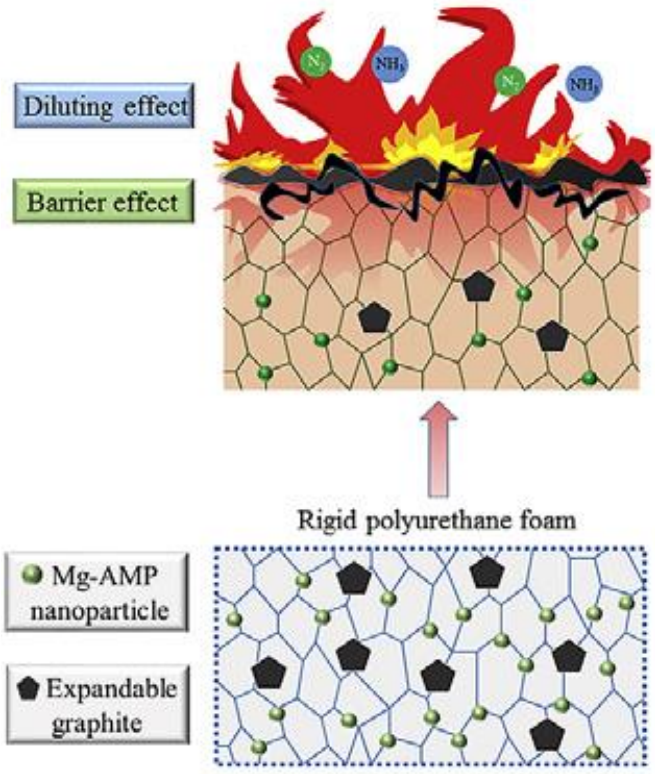
Rys.1 Schemat synergicznego działania Chitosanu (CS) i Czarnego Fosforu (BP) [11].

[8] Lei Liua, Zhengzhou Wanga, *Synergistic effect of nano magnesium amino-tris-(methylene-phosphonate)and expandable graphite on improving flame retardant, mechanical andthermal insulating properties of rigid polyurethane foam*, Materials Chemistry and Physics V:219 ELSEVIER 2018, s.319, 326.
[9] Yajun Chen, Linshan Li, Xiaoqing Qi , Lijun Qian, *The pyrolysis behaviors of phosphorus-containing organosilicon compound modified APP with different polyether segments and their flame retardant mechanism in polyurethane foam*,Composities Part B V:173, ELSEVIER 2019 s: 1-2.
[10] Dz. Cyt. Lei Liua *Synergistic effect of nano magnesium...* s.319, 326.
[11] Dz. Cyt. Sihao Yin, *Surface coating of biomass...* s. 1-5,

Na *Rys. 1* przedstawiono schemat synergicznego działania CS i BP, mieszanina tych uniepalniaczy wykazuje właściwości wygaszające *(z ang. Quenchingeffect)*, rozcieńczające *(z ang. Dilutioneffect)* i katalizujące proces zwęglania *(z ang. Catalyticcarbonization.)*

2.6.Nanozwiązki

Prowadzi się również prace badawcze mające na celu wytwarzanie PPUR o ograniczonej palności przy pomocy nanonapełniaczy. Dotychczasowe prace wskazują, że dodatek do PPUR montmorylonitu pozwala zintensyfikować uniepalniający efekt, przy dodaniu innych handlowych antypirenów. Dodatkowo wpływa na poprawę właściwości mechanicznych tak zmodyfikowanych PPUR.[12]. Poszukiwano również związki nanofosforu i azotu, posiadające właściwości zmniejszające palność, jednocześnie wzmacniające odporność PPUR na ogień i utrzymanie jej wytrzymałości na ściskanie. Amino-tris(metylenofosfonian) magnezu, (Mg-AMP) ma wysoką zawartość grup aminowych i hydroksylowych fosforu. Zastosowanie Mg-AMP poprawia zawartość zwęglania warstwy, uwalnia niepalne gazy rozcieńczająca palne gazy w fazie gazowej, również może katalizować powstanie zwęglania[13].



Rys.2 Synergiczne działanie Mg-AMP i GE podczas spalania.[14]

[12] Dz. Cyt Czech-Polak J., Oliwa R., Oleksy M., Budzik G., Sztywne *pianki poliuretanowe...* s. 115-124.
[13] Dz. Cyt. Lei Liua *Synergistic effect of nano magnesium...* s.319, 326.
[14] Dz. Cyt. Lei Liua *Synergistic effect of nano magnesium...* s.319, 326.

Schemat *(rys 2)* przedstawia synergiczne działanie Mg-AMP w fazie gazowej, powodując rozcieńczenie *(ang. Dilutioneffect)* gazów palnych i równoczesny efekt bariery fizycznej *(ang. Barriereffect)*w postaci warstwy zwęglonej („robaczków") grafitu ekspandowanego.

3. Podsumowanie

Wiele prowadzonych obecnie prac naukowych ma na celu uniepalnianie pianek poliuretanowych, sporo uwagi poświęca się brakowi toksyczności uniepalnionych materiałów. Szczególne zainteresowanie skierowane jest w stronę potrzeby biodegradowalności materiałów, co potwierdzają badania na temat Chitosanu przeprowadzone przez Sihao Yin i jego współpracownuków.[3] Przyjazny środowisku również okazał się być fosfor czarny badany przez tych samych autorów. Równie popularnymi antypirenami są nanozwiązki, a ich wykorzystanie w poprawie właściwości ognioodpornych materiałów pozwala na zachowanie niepogorszonych właściwości mechanicznych.

Literatura

1. Czech-Polak J., Oliwa R., Oleksy M., Budzik G., Sztywne *pianki poliuretanowe o zwiększonej odporności na płomień*, Wydawnictwo Sieć Badawcza Łukasiewicz – Instytut Chemii Przemysłowej, T: 63 nr 2,Warszawa 2018, s. 115-124.
2. Prociak A., Rokicki G., Ryszkowska J., Materiały poliuretanowe, PWN, Warszawa 2016, s 11-15; 162-184.
3. Sihao Yin, Yirou Du, Xiaodong Liang , YuhuiXie, Delong Xie, Yi Mei, *Surface coating of biomass-modified black phosphorus enhances flame retardancy of rigid polyurethane foam and its synergistic mechanism*, Applied Surface Science V: 637, s. 1-5, ELSEVIER 2023
4. Yao Yuan, Chao Ma, Yongqian Shi, Lei Song, Yuan Hu, Weizhao Hu, *Highly-efficient reinforcement and flame retardancy of rigid polyurethane foam with phosphorus-containing additive andnitrogen-containing compound*, Materials Chemistry and Physics V: 211, s.44, ELSEVIER 2018.
5. Lei Liua, Zhengzhou Wanga, *Synergistic effect of nano magnesium amino-tris-(methylene-phosphonate)and expandable graphite on improving flame retardant, mechanical andthermal insulating properties of rigid polyurethane foam*, Materials Chemistry and Physics V: 219 ELSEVIER 2018, s.319, 326.
6. Yajun Chen, Linshan Li, Xiaoqing Qi , Lijun Qian, *The pyrolysis behaviors of phosphorus-containing organosilicon compound modified APP with different polyether segments and their flame retardant mechanism in polyurethane foam*, Composities Part B V: 173, ELSEVIER 2019 s: 1-2.

Wiktoria Słąba
Koło Naukowe „Esprit”

dr inż. Dorota Głowacz-Czerwonka[1], prof. PRz

Właściwości fizyczne sztywnych pianek poliuretanowych z dodatkiem uniepalniaczy addytywnych

Streszczenie

Otrzymano sztywne pianki poliuretanowe z Rokopolu RF-151V z dodatkiem uniepalniaczy.Zbadano wpływ antypirenów na właściwości fizyczne uzyskanych kompozycji (z udziałemMasteretu 63560 i innych uniepalniaczy) poprzez przeprowadzenie badania gęstości pozornej, chłonności wody, stabilności wymiarów i przewodnictwa cieplnego. Uzyskano pianki, których właściwości były lepsze niż pianek komercyjnych (niższy współczynnik przewodzenia ciepła, mniejsza chłonność wody).

Słowa kluczowe: Sztywne pianki poliuretanowe, uniepalniacze, gęstość pozorna, chłonność wody, współczynnik przewodzenia ciepła.

1. Wprowadzenie

Przewodnictwo cieplne materiałów określa się za pomocą zastępczego współczynnika przewodzenia ciepła λ. Sztywne pianki poliuretanowe (SPPUR) zawdzięcza doskonałe właściwości termoizolacyjne niskiemu przewodnictwu cieplnemu używanych poroforów i korzystnej strukturze. Stwierdzono że, przewodnictwo cieplne gazów zawartych w komórkach jest najbardziej znaczącym mechanizmem transportu ciepła w PPUR i stanowi 60-80% wartości współczynnika przewodzenia ciepła. Zmniejszenie rozmiarów komórek w SPPUR zmniejsza transport ciepła przez promieniowanie, które jest istotną częścią przenoszenia ciepła [2].

Gęstość pozorna zależy głównie od rodzaju polimeru i udziału składników dodatkowych, a także temperatury i wilgoci. Dotyczy ona ciał stałych porowatych i cieczy zawierających pęcherze gazowe[3].

Eksploatacja pianek przy zmiennej temperaturze powoduje zmianę ciśnienia gazu w ich porach. Powoduje to wzrost (w wysokiej temp.) i obniżenie (w niskiej temp.) wymiarów.

[1]Politechnika Rzeszowska Im. Ignacego Łukasiewicza, katedra Chemii Organicznej
[2]Prociak A., Poliuretanowe materiały termoizolacyjne nowej generacji, PK Kraków 2008, s 34-42
[3]Żenkiewicz M., Steyńska M., Karasiewicz T., Moraczewski K., Rytlewski P., M*etody badań i oceny niektórych właściwości tworzyw polimerowych i metali.* Wydawnictwo Uniwersytetu Kazimierza Wielkiego Bydgoszcz 2012. s:11-18

SPPUR wykazują niewielkie zmiany wymiarów liniowych, które nie przekraczają1% liniowego. Plastyfikacja matrycy powodowana rozpuszczaniem się poroforów wewnątrz niej, zmniejsza właściwości mechaniczne i stabilność wymiarową PPUR [4].

Zawartość wody w tworzywie nazywana jest chłonnością wody, którą wyraża się w procentach wagowych lub objętościowych. Większość materiałów polimerowych charakteryzuje się niską chłonnością wody, wpływ ma jego budowa.Tworzywa niepolarne chłoną wodę znacznie gorzej od tych z licznymi grupami polarnymi. Zawartość wody znacznie pogarsza właściwości przetwórcze polimerów. W przypadku tworzyw higroskopijnych wilgotność prowadzi do pogorszenia właściwości mechanicznych.[5]

Otrzymane, sztywne pianki poliuretanowe(z udziałemMasteretu63560 połączonego odpowiednio z Roflamem P, Exolitem, Phoslitem B64AM, MPP, grafitem ekspandowany396 (EG 396), AddFor's APP201F i mieszaninieEG wraz Exolitem) poddano badaniu stabilności wymiarów, gęstości pozornej, chłonności wody i badaniu współczynnika ciepła.

2. Część doświadczalna

2.1. Materiały

Rokopol® RF-151V(PCC Rokita SA), Ongonat® 2100 (BorsodChem, Hungary), Trietyloamina (TEA),POCh S.A,NiaxSilikon L-6900 (Momentive), Roflam P(PCC Rokita SA), ExolitOP 935 (Clariant Plastics GmbH), Phoslite B64 AM(Italmatch Chemicals), MPP - Polifosforan melaminy(Hangzhou Mei Wang Chemical Co., Ltd.), EG 396 (Sinograf), MasteretTM63560 (Italmatch Chemicals), AddForce APP201F (WTH GmbH).

2.2. Aparatura

Aparat do badania przewodnictwa cieplnego IZOMET 2114 (Applied Precision, Słowacja), suszarka laboratoryjna SML (Zalmed Polska), waga analityczna (Radwag Radom),

2.3. Otrzymywanie pianek poliuretanowych

Do kubka (PP) wprowadzono Rokopol RF-151V, TEA, silikon i wodę. Całość wymieszano i dodano uniepalniacze(z wyjątkiem pianki referencyjnej)i MDI. Po otrzymaniu pianki sezonowano.

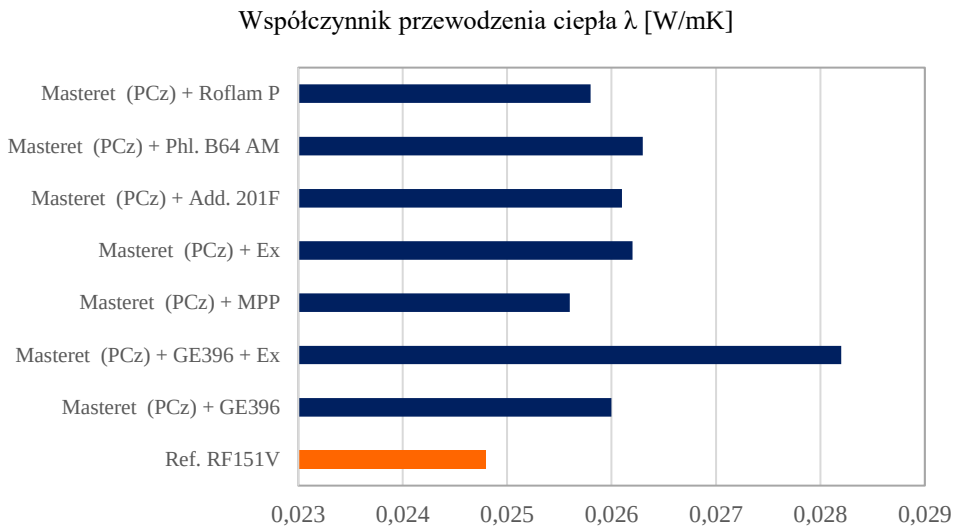
[4]Prociak A., Poliuretanowe... s 34-42
[5] Żenkiewicz M., M*etody badań i oceny...*. s:11-18

2.4. Przygotowanie kształtek do badań

Z otrzymanych kompozycji wycinano kształtki do dalszych badań.

3. Właściwości fizyczne

Niskie wartości współczynnika przewodzenia ciepła charakteryzują materiały izolacyjne.Zależy on głównie od obecności komórek zamkniętych. Pianki z dodatkiem uniepalniaczywykazują wyższe wartości współczynnika w porównaniu z pianką referencyjną *(Rys. 1),* jednak wszystkie są niższe od wartości 0,03 [W/mK], która charakteryzuje tradycyjne kompozycje dostępne na rynku i wykorzystywane do izolacji np. styropian czy wełnę mineralną[6]. Przykładem pianki komercyjnej dostępnej na rynku jest Purex NG-0810NF, której $\lambda = 0{,}0398$ [W/m·K][7].



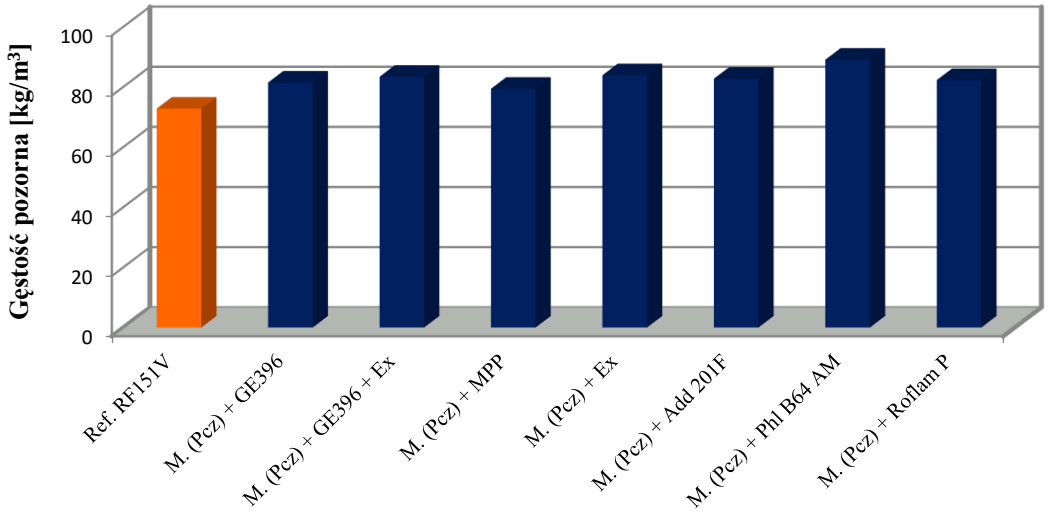
Rys. 1 Współczynnik przewodzenia ciepła otrzymanych kompozycji

3.1. Gęstość pozorna kompozycji

Gęstość pozorna jest wielkością dotyczącą ciał stałych porowatych i cieczy zawierających pęcherze gazowe [8]. Pianka komercyjna PUR-S80 firmy "JAG" PPH Sp. z o.o. której gęstość pozorna wynosi 80 [kg/m^3] o klasie palności F (palna) [9]. wykazuje niewiele niższą gęstość pozorną od otrzymanych kompozycji, a dodatkowo pianka komercyjna jest palna w przeciwieństwie do pianek z dodatkiem uniepalniaczy.

[6] inzynierbudownictwa.pl/materialy-i-wyroby-do-izolacji-cieplnej/ (dostęp 18.12.2023)
[7] www.piankisklep.pl/Systemy-natryskowe-pianki-poliuretanowej-Purex-NG-0440.html (dostęp 18.012024)
[8] Żenkiewicz M., M*etody badań i oceny*.... s:11-18
[9] https://jag.pl/oferta/sztywna-pianka-poliuretanowa-typ-pur-pir/ (dostęp 14.01.2024)

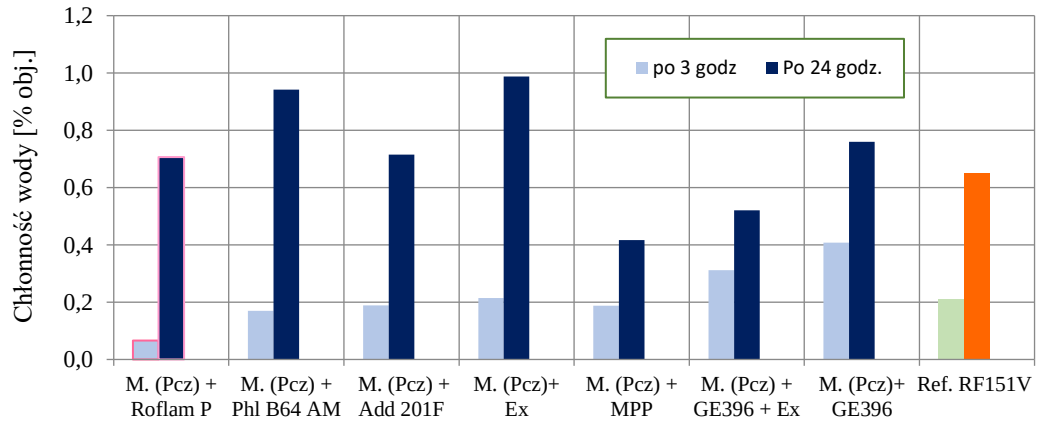
Gęstość pozorna pianki referencyjnej imieści się w granicach typowych dla sztywnych pianek poliuretanowych wytwarzanych w przemyśle *(Rys. 2)*, [10].



Rys. 2 Gęstość pozorna otrzymanych pianek

3.2. Chłonność wody pianek

Zawartość wody w tworzywie nazywamy chłonnością wody, wyrażaną w %obj. lub %mas.. Badanie przeprowadzono w celu sprawdzenia wpływu dodatków uniepalniających na zmianę tego parametru. Uzyskane kompozycje charakteryzują się mniejszą chłonnością wody (< 2.1% obj.)od pianki komercyjnej Purex-NG-0440 (chłonność wody < 3% obj.), *Rys. 3*) [11]



Rys.3 Chłonność wody kompozycji [%obj.]

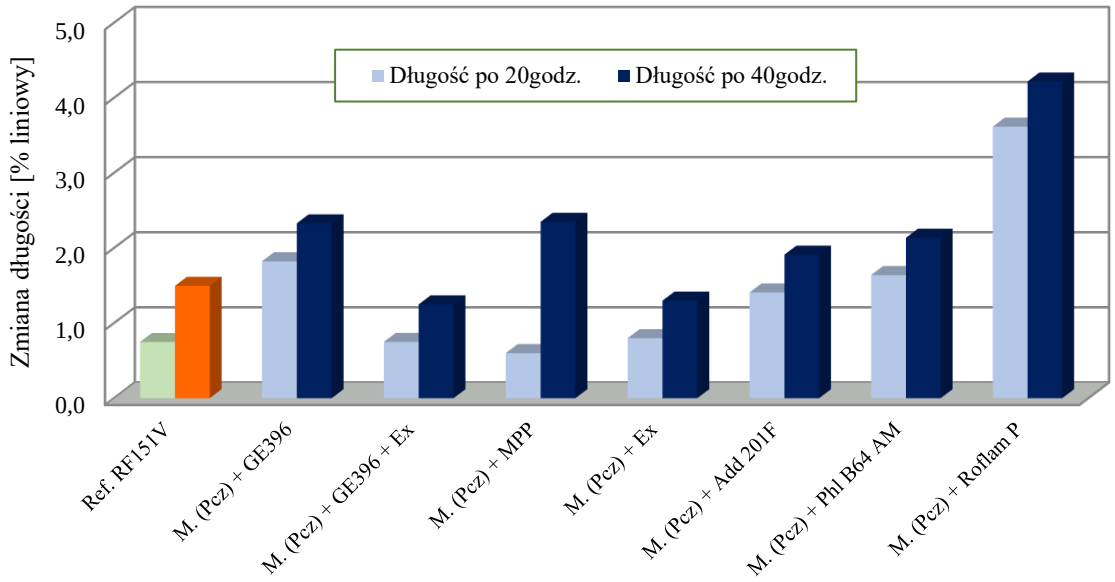
[10]Prociak A., Rokicki G., Ryszkowska J., Materiały poliuretanowe, PWN, Warszawa 2016, s 11-15; 162-184.
[11]www.piankisklep.pl/Systemy-natryskowe-pianki-poliuretanowej-Purex-NG-0440.html (dostęp 14.01.2024)

Najniższa chłonność wody wynika z obecności największej ilości komórek zamkniętych.

3.3.Stabilność wymiarów

Stabilność wymiarowa odnosi się do zdolności polimerów do zachowywania swoich rozmiarów w różnych warunkach otoczenia. Tworzywo sztuczne stabilne wymiarowo w niewielkim stopniu podlega rozszerzalności cieplnej. Współczynnik liniowej rozszerzalności cieplnej określa zakres zmiany długości materiału przy wzroście lub spadku temperatury[12].

Zbadano wpływ dodatku uniepalniaczy na stabilność wymiarową.

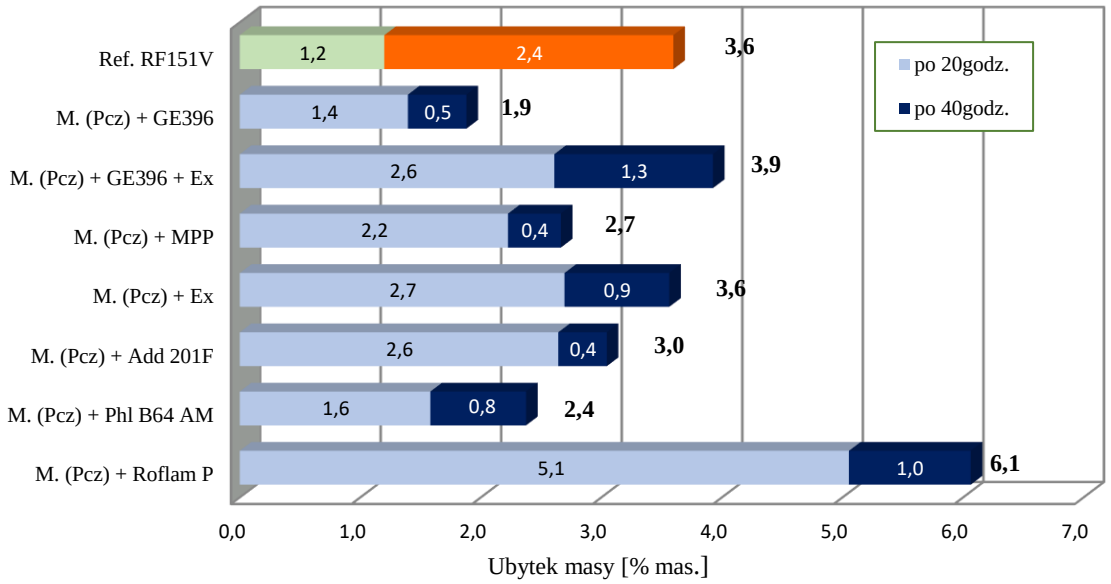


Rys. 4 Porównanie zmiany długości po 20 i 40godz wygrzewania w temp. 150°C

Wyniki badania stabilności wymiarów wykazały, że największą procentową zmianą wymiarów (długości) cechuje się kompozycja uzyskana przy udziale Roflamu P i Masteretu (Pcz) – jest to wartość wyższa niż uzyskana dla pianki referencyjnej *(Rys. 4).*

Przeprowadzono także badanie ubytku masy pianek poddanych działaniu podwyższonej temperatury (150°C) przez 40 godzin. Największym ubytkiem masy charakteryzuje się kompozycja Masteretu (Pcz.) z Roflamem, natomiast najmniejszym ubytkiem masy pianka Mastertet (Pcz) z EG 396 (1.9% mas.), por. *Rys. 5.*

[12]PN 92/C 89083, Sztywne tworzywa sztuczne Badanie stabilności wymiarów



Rys. 5 Wykres ubytku masy [% mas.] po 20 i 40 godz. wygrzewania w temp. 150°C.

Ubytek masy kompozycji zawierającej mieszaninę trzech uniepalniaczy jest porównywalny z wartością uzyskaną dla pianki referencyjnej.

4. Podsumowanie

Wprowadzenie uniepalniaczy znacząco wpływa na właściwości fizyczne otrzymanych pianek. W oparciu o przeprowadzone badania wykazano, że uzyskane kompozycje charakteryzują się niskim współczynnikiem przewodzenia ciepła, który mieści się w przyjętym zakresie wartości dla pianek komercyjnych.

Właściwości fizyczne otrzymanych (z udziałem uniepalniacza) kompozycji są dla większości pianek porównywalne z charakterystycznymi dla pianek komercyjnych, a niekiedy ich właściwości są korzystniejsze niż w przypadku pianki referencyjnej (chłonność wody jest mniejsza w przypadku pianek z dodatkiem Masteretu 63560 i Roflamu P, jak również Masteretu 63560 i MPP).

Literatura

1. Żenkiewicz M., Steyńska M., Karasiewicz T., Moraczewski K., Rytlewski P., *Metody badań i oceny niektórych właściwości tworzyw polimerowych i metali.* Wydawnictwo Uniwersytetu Kazimierza Wielkiego Bydgoszcz 2012. s:11-18.
2. Prociak A., Poliuretanowe materiały termoizolacyjne nowej generacji, PK Kraków 2008, s 34-42.

3. Prociak A., Rokicki G., Ryszkowska J., Materiały poliuretanowe, PWN, Warszawa 2016, s 11-15; 162-184.

Akty normatywne

1. PN 92/C 89083, Sztywne tworzywa sztuczne Badanie stabilności wymiarów
2. PN 93/C 89084, Tworzywa sztuczne - Oznaczanie chłonności wody przez sztywne tworzywa porowate

Źródła internetowe

1. www.piankisklep.pl/Systemy-natryskowe-pianki-poliuretanowej-Purex-NG-0440.html (dostęp 18.012024)
2. inzynierbudownictwa.pl/materialy-i-wyroby-do-izolacji-cieplnej/ (dostęp 18.12.2023)
3. https://jag.pl/oferta/sztywna-pianka-poliuretanowa-typ-pur-pir/ (dostęp 14.01.2024)
4. *www.piankisklep.pl/Systemy-natryskowe-pianki-poliuretanowej-Purex-NG-0440.html* (dostęp 14.01.2024)

Wiktoria Słąba
Koło Naukowe Esprit

dr inż. Dorota Głowacz-Czerwonka, prof. PRz [1]

Właściwości ogniowe sztywnych pianek poliuretanowych z udziałem addytywnych uniepalniaczy

Streszczenie

Otrzymano sztywne pianki poliuretanowe z Rokopolu RF-151V z udziałem wybranych uniepalniaczy addytywnych. Zbadano wpływ antypirenów na właściwości ogniowe kompozycji przeprowadzając badanie indeksu tlenowego i testu poziomego. Sprawdzono możliwość wystąpienia efektu synergicznego pomiędzy wprowadzonymi do kompozycji uniepalniaczami. Uzyskano kompozycje o zwiększonej odporności na płomień (LOI > 21% obj. tlenu), co potwierdziło właściwości samogasnące, a także kompozycje trudno zapalne (LOI > 28% obj. tlenu).

Słowa kluczowe: Sztywne pianki poliuretanowe, uniepalniacze, indeks tlenowy

1. Wprowadzenie

Stopień palności materiałów jest istotny ze względu na zagrożenie pożarowe jakie stanowią podczas produkcji i użytkowania, dlatego wszystkie materiały konstrukcyjne muszą być scharakteryzowane jako palne, samogasnące lub niepalne. Parametry takie jak: zapalność, szybkość wydzielania ciepła, rozprzestrzeniania płomienia po powierzchni materiału określają charakterystykę pożarową materiału[2].

Indeks tlenowy to metoda pomiaru wskaźnika tlenowego stosuje się do wszystkich tworzyw sztucznych w celu porównawczej oceny ich zapalności. Indeks tlenowy jest to najmniejsza procentowa zawartość objętościowa tlenu w mieszaninie tlenu i azotu, która podtrzymuje stałe palenie się badanej próbki.[3]

Test poziomego palenia dla klasy UL 94 HB , polega na zaznaczeniu linii 2,5 cm na badanej próbce o wymiarach 150x13x50mm. Na zamocowaną poziomo próbkę w statywie kieruje się płomień palnika pod kątem 45˚ o długości płomienia 2,5cm, obejmującego przednią część próbki. Przykłada się na 30 sekund, a następnie się go odsuwa. Mierzy się czas spalania, powyżej 38,1 mm/min uznaje się, że badany materiał posiada klasę palności UL94 HB[4].

---

[1]Politechnika Rzeszowska Im. Ignacego Łukasiewicza, Katedra Chemii Organicznej.

[2] Prociak A., Poliuretanowe materiały termoizolacyjne nowej generacji, PK Kraków 2008, s 34-42.

[3] PN-EN-ISO-4589 Tworzywa sztuczne | Oznaczanie zapalności metodą wskaźnika tlenowego Część 2: Badanie w temperaturze pokojowej

[4] https://www.energotech.pl/doc/File/download/KLASYFIKACJA_PALNOSCI.pdf (dostęp 14.12.2023)

W niniejszym artykule otrzymano sztywne pianki poliuretanowe oparte na Rokopolu RF-151V z udziałem wybranych uniepalniaczy. Kompozycje poddano testom ogniowym w celu zbadania odporności na płomień.

2. Część doświadczalna

2.1. Materiały

Zastosowano następujące surowce: Rokopol® RF-151V (PCC Rokita SA), Ongonat® 2100 (BorsodChem, Hungary), Trietyloamina (POCh SA), NiaxSilikon L-6900 (Momentive), Roflam P (PCC Rokita SA), Exolit OP 935 (Clariant Plastics GmbH), Phoslite B64 AM (Italmatch Chemicals, Italy), Polifosforan melaminy (MPP), EG 396 (Sinograf,) Masteret TM 63560 (Italmatch Chemicals, Italy), AddForce APP 201F (WTH GmbH, Niemcy).

2.2. Aparatura

Aparat do wyznaczania indeksu tlenowego (Concept Equipment, Wielka Brytania), waga analityczna (Radwag, Polska), zestaw do wykonania testu poziomego (wg UL-94 HB).

2.3. Otrzymywanie pianek poliuretanowych

Do polipropylenowego kubka wprowadzono Rokopol RF-151V, TEA, silikon i wodę. Całość mieszano i dodawano uniepalniacze (z wyjątkiem pianki referencyjnej) i MDI. Jako antypireny zastosowano: Masteret 63560, Roflam P, Exolit OP 935, Phoslite B64AM, MPP, EG 396 oraz AddForce APP201F. Po otrzymaniu pianki sezonowano przez 72 godz.

2.4. Przygotowanie kształtek do badań

Z otrzymanych kompozycji wycinano kształtki do dalszych badań.

3. Omówienie wyników

Otrzymano sztywne pianki poliuretanowe z udziałem uniepalniaczy (60% mas. względem Rokopolu RF-151V) oraz bez ich dodatku (kompozycja referencyjna). W celu sprawdzenia wpływu wprowadzonych antypirenów na właściwości ogniowe uzyskanych pianek poliuretanowych (PPUR) zbadano ich właściwości ogniowe przy pomocy testu poziomego palenia[5] i badania indeksu tlenowego zgodnie z normą PN-EN-ISO-4589.

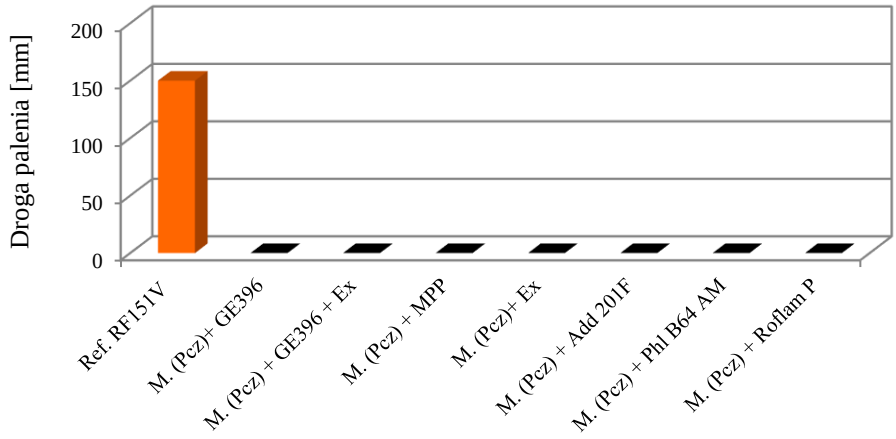
[5] https://www.energotech.pl/doc/File/download/KLASYFIKACJA_PALNOSCI.pdf (dostęp 14.12.2023)

3.1. Test poziomy

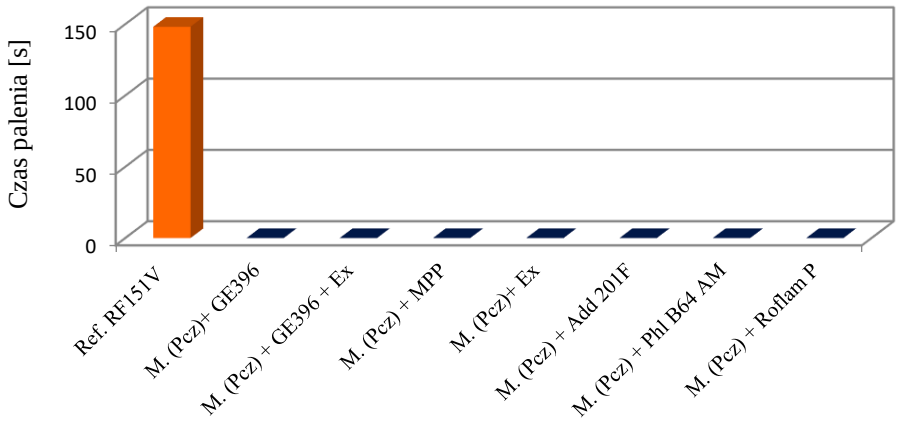
Test poziomy palenia dla klasy UL 94 HB wykonuje się aby sprawdzić czy dany materiał spełnia wymogi materiału samogasnącego [6]. Podczas badania wyznaczono ubytek masy, czas palenia, drogę palenia i szybkość palenia uzyskanych kompozycji.

Pianka referencyjna po odsunięciu płomienia uległa całkowitemu spaleniu. Droga palenia objęła całą kształtkę (długość normatywna, *Rys. 1*), a jej czas palenia był długi (wynosił 148 s), co spowodowane strukturą porowatą materiału (*Rys. 2*). Oznacza to, że pianka bez dodatku uniepalniacza nie spełnia wymogów testu UL 94 HB. Kompozycje z dodatkiem uniepalniaczy przestawały się palić natychmiast po usunięciu płomienia.

Przeprowadzone badanie pozwala zatem na zaklasyfikowanie pianek otrzymanych przy udziale uniepalniacza (według normy UL 94 HB) jako kompozycji samogasnących.



Rys. 1 *Porównanie drogi palenia dla kompozycji PUR*

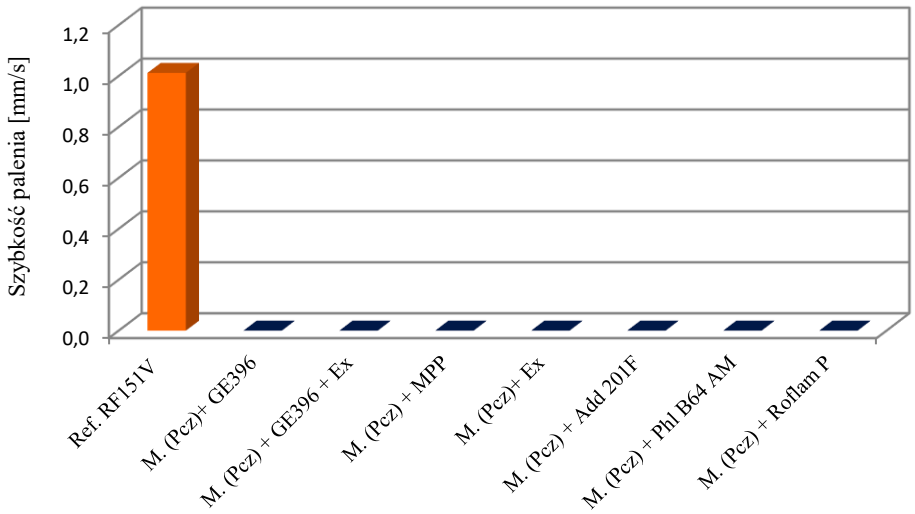


Rys.2 Czas palenia otrzymanych kompozycji podczas testu poziomego.

[6] *Ibidem* KLASYFIKACJA_PALNOSCI.pdf (dostęp 14.12.2023)

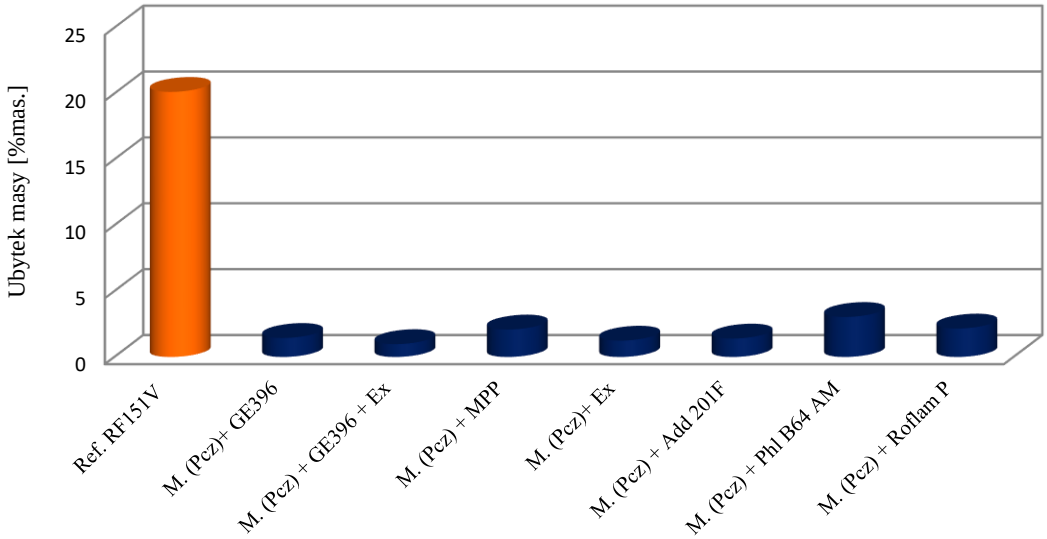
Antypireny skutecznie ograniczyły rozprzestrzenianie się ognia w piankach. Podczas badania zaobserwowano wytworzenie się warstwy zwęglonej na powierzchni pianek z dodatkiem uniepalniaczy. Powstała warstwa ma na celu blokowanie dostępu powietrza do pianki, a także zmniejszenie szybkości jej palenia się.

Pianka referencyjna spłonęła w całości, natomiast żadna z pozostałych kompozycji (uzyskanych z udziałem uniepalniacza) nie uległa zapaleniu (szybkość palenia przyjęto zatem jako 0 mm/s), *Rys. 3*.



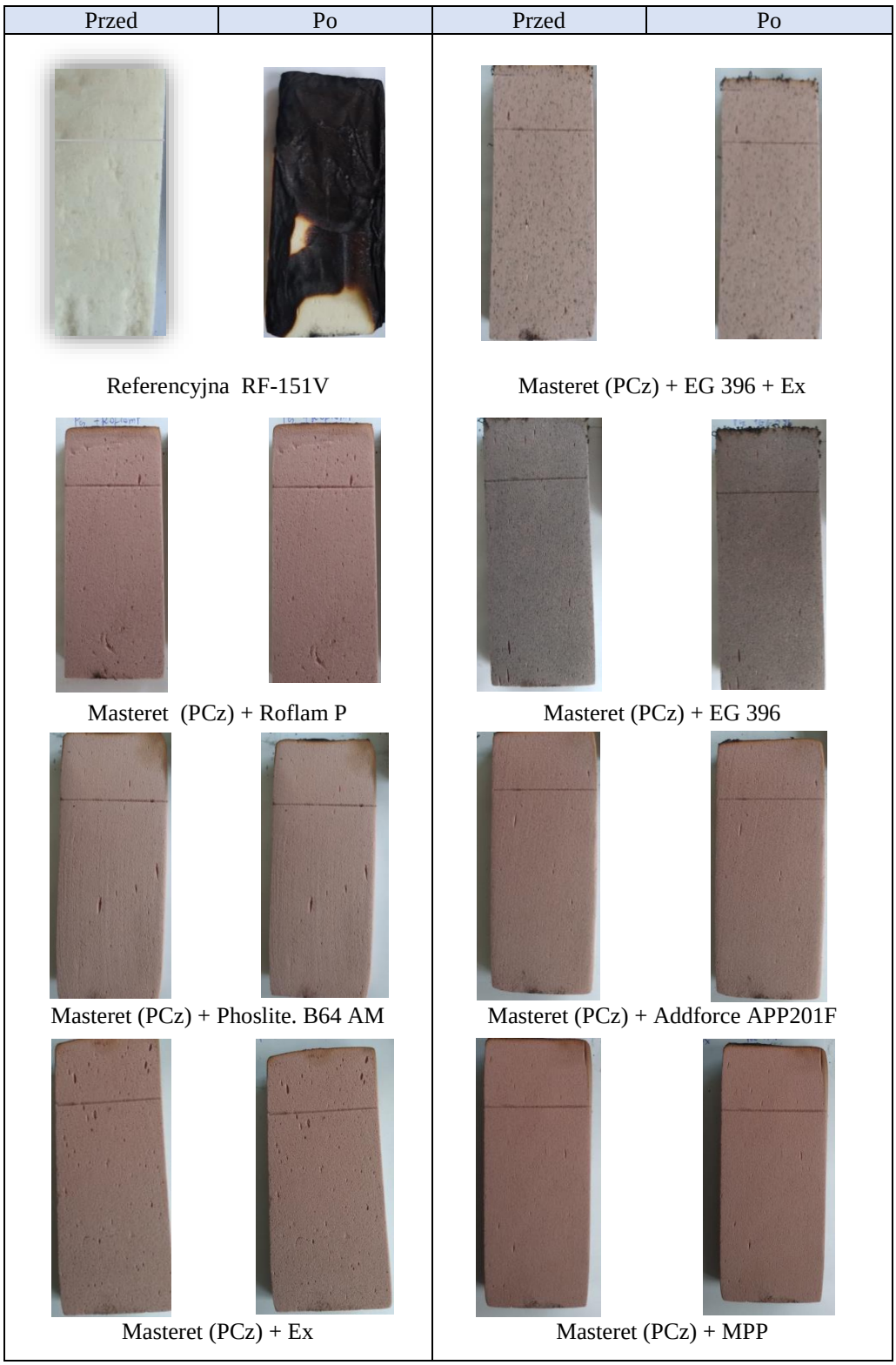
Rys. 3 Szybkość palenia uzyskanych kompozycji podczas testu poziomego

Zbadano ubytek masy kompozycji podczas przeprowadzania testu poziomego według normy UL-94 HB.



Rys. 4 Ubytek masy kompozycji po przeprowadzeniu testu poziomego (wg Ul-94 HB)

Największym ubytkiem masy charakteryzuje się pianka referencyjna (20,1% mas.), najmniejszym kompozycja z dodatkiem Masteretu 63560 i EG 396 (0,98% mas.). Wszystkie otrzymane i uniepalnione kompozycje wykazują niewielki ubytek masy (poniżej 3% mas.), co świadczy o ich odporności na działanie ognia.



Rys. 5 Pianki przed i po przeprowadzonym teście poziomym

Wygląd pianek przed i po przeprowadzeniu testu poziomego przedstawiono na *Rys. 5*. Jak wykazał badanie wprowadzone do badanych kompozycji antypireny działają skutecznie - powodują gaśnięcie płomienia i nie pozwalają na jego rozprzestrzenianie się.

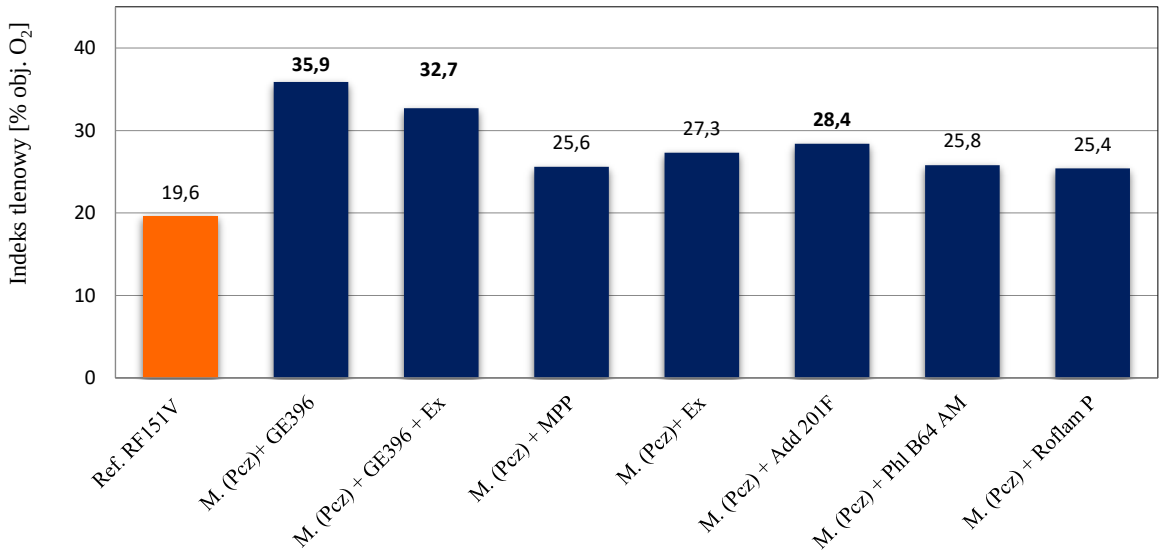
3.2. Indeks tlenowy

Amerykański przemysł wprowadził podział tworzyw polimerowych (w oparciu o wartości indeksu tlenowego) na następujące klasy palności *(Tab. 1)*:[7]

Tabela 1 Klasy palności eg wartości indeksu tlenowego *(Low Oxygen Indeks (LOI))* [8]

Klasa palności tworzyw	LOI (% obj. tlenu)
palne	< 21
samogasnące	21 < LOI < 28
niepalne	28 < LOI

W oparciu o inną klasyfikację tworzywa dzieli się na dwie grupy: palne (LOI < 26% obj. tlenu)l ub trudno zapalne (LOI ≥ 26% obj. tlenu) konieczna literatura.



Rys. 6 Wyniki badania indeksu tlenowego

W wyniku badania indeksu tlenowego stwierdzono, iż wyłącznie pianka referencyjna (uzyskana bez dodatku antypirenów) jest palna (LOI =19,6% obj.). W wypadku wszystkich pozostałych, tj. uniepalnionych kompozycji jest on wyższy niż w przypadku pianki

[7] Norwiński S., Postawa P., *Ocena palności metodą indeksu tlenowego (OI) wybranych kompozytów na osnowie polipropylenu*, Przetwórstwo Tworzyw, T.22 ,s. 285-293, Częstochowa 2016
[8] *Ibidem* s 285-293

referencyjnej, co wskazuje na pożądany efekt synergii uniepalniaczy zawierających atomy azotu i fosforu.

W przypadku trzech kompozycji osiągnięto klasę palności – niepalną lub trudno zapalną. Niepalne okazały się pianki:

- Masteret 63560 + EG 396 (najwyższa wartość LOI (35,9% obj. tlenu),
- Masteret 63560 + EG 396 + Exolit OP 935 (LOI: 32,7% obj. tlenu),
- Masteret 63560 + Adforce FR APP 201F (LOI: 28,4% obj. tlenu).

Pozostałe, otrzymane pianki okazały się samogasnące. Potwierdzono synergiczne działanie zastosowanych uniepaniaczy np. grafitu ekspandowanego (działającego w fazie stałej tworzącego barierę fizyczną przed płomieniem) i Masteretu działającego w fazie gazowej.

4. Podsumowanie

Otrzymano sztywne pianki poliuretanowe z Rokopolu RF-151V z udziałem wybranych uniepalniaczy addytywnych. Zbadano wpływ antypirenów na właściwości ogniowe kompozycji przeprowadzając badanie indeksu tlenowego i testu poziomego. Sprawdzono możliwość wystąpienia efektu synergicznego pomiędzy wprowadzonymi do kompozycji uniepalniaczami. Uzyskano kompozycje o zwiększonej odporności na płomień charakteryzujące się OI większym niż 21%, co potwierdziło właściwości samogasnące i otrzymano trzy kompozycje powyżej 28% oznaczające pianki trudno zapalne.

Literatura

1. *Prociak A., Poliuretanowe materiały termoizolacyjne nowej generacji, PK Kraków 2008, s 34-42.*
2. Norwiński S., Postawa P., *Ocena palności metodą indeksu tlenowego (OI) wybranych kompozytów na osnowie polipropylenu*, Przetwórstwo Tworzyw, T.22 ,s. 285-293, Częstochowa 2016

Akty normatywne

1. PN-EN-ISO-4589 *Tworzywa sztuczne | Oznaczanie zapalności metodą wskaźnika tlenowego Część 2: Badanie w temperaturze pokojowej*

Źródła internetowe

1. https://www.energotech.pl/doc/File/download/KLASYFIKACJA_PALNOSCI.pdf (dostęp 14.12.2023)

Magdalena Cebula
Koło Naukowe Studentów Chemii ESPRIT

Prof. dr hab. inż. Wiktor Bukowski
Opiekun Koła Naukowego

Chromatografia flash jako nowoczesna technika rozdziału i oczyszczania substancji chemicznych

Streszczenie

W niniejszym artykule zaprezentowano możliwości wykorzystania techniki chromatografii cieczowej typu flash do rozdziału i oczyszczania substancji chemicznych. Ten typ preparatywnej chromatografii jest stosowany m.in. do oczyszczania półproduktów i substancji aktywnych na potrzeby przemysłu farmaceutycznego. Przedstawiono ogólną zasadę działania chromatografii. Omówiono główne elementy aparatury wyróżniające chromatografię typu flash w porównaniu do wysokosprawnej chromatografii cieczowej. Przedstawiono przykłady wykorzystania chromatografii typu flash do wstępnego rozdziału i oczyszczenia substancji chemicznych. Wyszczególniono zalety, jakie niesie za sobą sprzężenie wysokosprawnej chromatografii cieczowej z chromatografią typu flash.

Słowa kluczowe: chromatografia, preparatywna HPLC, chromatografia typu flash, przemysł farmaceutyczny

1. Wprowadzenie do chromatografii

Różne techniki chromatograficzne należą do narzędzi wykorzystywanych powszechnie przez chemików i specjalistów dziedzin pokrewnych do analizy próbek substancji organicznych, wyodrębniania produktów organicznych z mieszanin poreakcyjnych oraz izolacji składników z ekstraktów zawierających skomplikowany materiał biologiczny. Obecny stopień rozwoju metod chromatograficznych umożliwia wykrywanie analitu i oznaczanie jego zawartości w próbce przy stężeniach na bardzo niskim poziomie, często wobec wielu innych substancji. Chromatografia preparatywna jest z kolei wykorzystywana jako metoda otrzymywania czystych substancji, także coraz częściej w ilościach

przemysłowych, szczególnie w przemyśle: farmaceutycznym, kosmetycznym oraz spożywczym.[1]

Chromatografia sama w sobie stanowi proces separacji polegający na rozdzieleniu składników pomiędzy dwie fazy: stacjonarną oraz ruchomą, która przemieszcza się względem fazy stacjonarnej. Chromatografię dzieli się na dwa podstawowe typy: cieczową i gazową. Różnica związana jest ze stanem fizycznym fazy ruchomej. W chromatografii gazowej fazą ruchomą jest gaz, który transportuje lotne składniki próbki przez stałą fazę stacjonarną, natomiast w chromatografii cieczowej fazą ruchomą jest rozpuszczalnik lub mieszanina rozpuszczalników. Do rozdziału substancji wykorzystuje się różnice w oddziaływaniu składników próbki z fazą stacjonarną, związane z różnicą polarności, rozmiaru cząsteczek lub wzajemnym powinowactwie grup funkcyjnych składników fazy ruchomej z powierzchnią fazy nieruchomej.[2] W wypadku metod chromatografii cieczowej faza stacjonarna charakteryzuje się porowatą naturą i dużą powierzchnią właściwą. W chromatografii gazowej współcześnie stosuje się najczęściej kapilarne kolumny chromatograficzne z fazą stacjonarną w postaci cienkiego filmu naniesionego na wewnętrzne ścianki kapilar o różnych długościach i średnicach.

2. Chromatografia preparatywna w przemyśle farmaceutycznym

W przemyśle farmaceutycznym wszystkie wytwarzane produkty muszą być najwyższej jakości, co stanowi podstawę bezpieczeństwa, jeżeli chodzi o zdrowie i życia pacjenta. Aby potwierdzić, że produkty farmaceutyczne spełniają wymagane standardy, badacze, producenci i projektanci zobowiązani są do oceny jakościowej, ilościowej oraz wyjaśnienia struktury składników w złożonych mieszaninach. W ciągu ostatnich kilku dekad nastąpił ogromny postęp technologiczny w zakresie oprzyrządowania i technik chromatograficznych. Wśród nich stale rozwijają się techniki chromatografii preparatywnej.[3]

[1] Z. Witkiewicz, J. Kałużna – Czaplińska, *Podstawy chromatografii i technik elektromigracyjnych,* Wydawnictwo Naukowe PWN, Warszawa 2017, 21-26.

[2] Z. Witkiewicz, W. Wardencki, *Chromatografia gazowa. Teoria i praktyka*, Wydawnictwo Naukowe PWN, Warszawa 2018, 11-12.

[3] N. Surve, A. Thomas R. Bhole, C. Patil, *Flash Chromatography and Semi-Preparative HPLC: Review on the Applications and Recent Advancements over the Last Decade*, Eurasian Journal of Chemistry, 2023, 28.1 (109)

Podstawowym celem chromatografii preparatywnej jest oddzielenie oraz oczyszczenie składników mieszanin przed ich dalszym wykorzystywaniem. Wśród metod chromatografii preparatywnej można wyróżnić następujące rodzaje: chromatografię grawitacyjną (niskie ciśnienie), typu flash (średnie ciśnienie) oraz preparatywną HPLC (wysokie ciśnienie). Wymienione typy różnią się ciśnieniem wytwarzanym przez przepływ fazy ruchomej przez fazę stacjonarną. Chromatografia z zasilaniem grawitacyjnym, znana również jako chromatografia na otwartej kolumnie, wykorzystuje wyłącznie grawitację do transportu rozdzielanych substancji przez kolumnę, co stanowi proces raczej czasochłonny. Nowoczesne systemy chromatograficzne zaopatrzone są w pompę, która pozwala na osiąganie większej prędkości przepływu i sprawia, że proces separacji jest bardziej wydajny. Zalety oraz wady tradycyjnych i nowoczesnych sposobów podsumowano w *Tabeli 1*.

Tabela 1. Porównanie tradycyjnej chromatografii na otwartej kolumnie z metodami flash i preparatywną HPLC.

Chromatografia na otwartej kolumnie	Flash/preparatywna HPLC
• Niskie wydatki na oprzyrządowanie i materiały eksploatacyjne • Niskie ciśnienie • Czasochłonna • Niska rozdzielczość • Małe natężenie przepływu • Wysokie zużycie rozpuszczalników • Brak regulowanego przepływu i ciśnienia • Niekompatybilna z wysoką wydajnością detektorów	• Wysokie koszty przyrządów i konserwacji • Wysoka automatyzacja • Szybki proces • Duża rozdzielczość • Duże natężenie przepływu • Niskie zużycie rozpuszczalnika • Wysoka automatyzacja • Kompatybilność z szeroką gammą detektorów i materiałów eksploatacyjnych

Źródło: Opracowanie własne

Obecnie na etapie przygotowawczym, docelowy związek oczyszcza się w dużych ilościach za pomocą chromatografii typu flash, preparatywnej HPLC lub kombinacji obu metod. Jeżeli te techniki stosuje się łącznie, to na etapie wstępnego oczyszczania i przygotowania, wykorzystuje się chromatografię typu flash, w celu uzyskania końcowej, wysokiej czystości. W dalszej części artykułu główna uwaga zostanie skupiona na rozwinięciu zagadnień związanych bezpośrednio z tym typem chromatografii.

3. Elementy aparatury wyróżniające chromatografię typu flash

Współczesna aparatura do chromatografii typu flash jest dostosowywana do indywidualnych potrzeb każdego użytkownika, co umożliwia planowanie procesów rozdziału w różnej skali. Istnieje większa różnorodność materiałów eksploatacyjnych przeznaczonych do chromatografii typu flash w porównaniu z materiałami do preparatywnej HPLC. Przed ich wyborem należy wziąć pod uwagę takie czynniki jak:

- trudność separacji
- wymagana czystość
- ilość próbki
- wydajność oczyszczania

Jednym z głównych czynników decydujących o możliwości rozdziału składników metodami chromatograficznymi jest rodzaj fazy stacjonarnej. Kolumny stosowane w chromatografii typu flash różnią się od tradycyjnych kolumn używanych w HPLC. Producenci przewidują możliwość ręcznego pakowania kolumn. Z tego też względu ważne jest uzyskanie równomiernego rozkładu cząstek i zapewnienie możliwości jednolitej gęstości upakowania. Ręcznie pakowane materiały eksploatacyjne zwykle nie gwarantują skuteczności separacji i powtarzalności rozdziału na poziomie uzyskiwanym z wykorzystaniem gotowych kolumn (kardridży) pakowanych w warunkach przemysłowych (*Rysunek 1*). Takie rozwiązanie stosuje się, gdy użytkownik chce w najbardziej ekonomiczny sposób oczyścić proste mieszaniny.

Istnieją dwa sposoby nanoszenie rozdzielanej mieszaniny na złoże adsorbentu w kolumnie chromatograficznej – tzw. suche i mokre. W metodzie mokrej próbka do oczyszczenia lub rozdzielana na składniki, rozpuszcza się w małej ilości lotnego rozpuszczalnika, takiego jak np. heksan czy aceton. Uzyskany roztwór nanosi się następnie ostrożnie na złoże adsorbentu wypełniającego kolumnę. Czasami rozpuszczalnik wybrany do załadowania próbki jest bardziej polarny niż rozpuszczalniki eluujące. W tym przypadku bardzo ważne jest, aby do załadowania próbki użyć jak najmniej takiego rozpuszczalnika. Zbyt duża ilość rozpuszczalnika polarnego może mieć wpływ na zakłócenie elucji, a tym samym oczyszczanie lub rozdzielanie mieszaniny. W takich przypadkach zaleca się suchą metodę obciążania kolumny. Polega ona na zwieszaniu rozdzielanej mieszaniny z odpowiednią porcją adsorbentu w celu zaadsorbowania rozdzielanych substancji. Po odparowaniu rozpuszczalnika, suchy sorbent przenosi się na szczyt przygotowanej kolumny.[4]

Głównym adsorbentem stosowanym w chromatografii typu flash jest krzemionka, zapewniająca lekko kwaśne środowisko. Jest to porowata forma dwutlenku krzemu, składająca się z nieregularnej trójwymiarowej struktury naprzemiennych atomów krzemu i tlenu. Grupy silanolowe rozmieszczone są na całej powierzchni materiału wypełniającego, łącznie z porami.

[4] A. B. Roge, S. N. Firke, R. M. Kawade, S. K. Sarje, S. M. Vadvalkar, *Brief review on: flash chromatography.* International Journal of Pharmaceutical Sciences and Research, 2011, 2(8), 1930-1937.

Materiał pozwala osiągać dobrą separację szerokiej gamy związków organicznych, szczególnie o wysokiej i średniej polarności. Ponadto krzemionka może występować w układzie faz odwróconych, w których składa się z niepolarnych grup organicznych, takich jak łańcuchy oktadecylu (C18), oktylu (C8) lub butylu (C4) związane z grupami silanolowymi żelu krzemionkowego. Zwykle C18 z długim łańcuchem alkilowym stosuje się dla związków mniejszych natomiast C8 i C4 dają lepsze wyniki dla cząsteczek o większej masie molowej.

Chętnie wykorzystywaną fazą stacjonarną jest również tlenek glinu, w którym aktywne punkty stanowią centra Al^{3+} i łączące je atomy O^{2-}. Żele tlenku glinu są dostępne w rozmiarach cząstek podobnych do krzemionkowych. Materiał cechuje lepsza stabilność pH, a jego neutralna postać jest często stosowana w przypadku związków wrażliwych na kwasy, których nie można oczyścić za pomocą krzemionki. Mimo że tlenek glinu ma niższą nośność niż krzemionka, w niektórych zastosowaniach zapewnia wyjątkową skuteczność separacji.

Chromatografia typu flash różni się od konwencjonalnej techniki bezciśnieniowej pod dwoma głównymi względami: po pierwsze, wykorzystuje się nieco mniejsze cząstki żelu krzemionkowego (250-400 Mesh), a po drugie, ze względu na ograniczony przepływ rozpuszczalnika spowodowany gęstym upakowaniem małych cząstek żelu, wykorzystuje się sprężony gaz (ok. 10-15 Psi), który zapewnia zwiększenie przepływu eluentów przez złoże sorbentu.[4]

Rysunek 1. Kolumny stosowane w chromatografii typu flash.
Źródło: https://haas.com.pl/produkt/sepaflash-ilok-sl/ (dostęp: 29.04.2024)

Innym aspektem, na który warto zwrócić uwagę to metody detekcji związków rozdzielanych metodą flash chromatografii. W chromatografii preparatywnej najczęściej stosuje się detektory UV. Ich podstawowa cecha to selektywność, co oznacza, że mierzą jedynie substancje, które absorbują światło o wybranej długości fali w zakresie ultrafioletu (200-400 nm) lub zakresu widzialnego (400-800 nm). Do substancji odpowiednich do detekcji metodą UV zaliczają się związki z grupą chromoforową, np. zawierające aromatyczny pierścień, dwa sprzężone wiązania podwójne, wiązanie podwójne sąsiadujące z atomem z jedną parą elektronów.

Detektory UV stosowane we flash chromatografii sprzężone są często z detektorami ELSD, które są preferowanym wyborem do oczyszczania związków bez grupy chromoforowej. Takie związki obejmują m.in. węglowodany, lipidy, tłuszcze i polimery. Czułość detektorów ELSD jest niezależna od właściwości fizycznych i chemicznych rozdzielanych analitów. Wpływ natomiast ma ich ilość, co oznacza, że wysoki sygnał wskazuje, że eluowana jest duża ilość związku. Z racji faktu, że detektory ELSD zaliczane są do detektorów niszczących, ilość próbki kierowana do takich detektorów powinna być jak najmniejsza.[5]

4. Przykłady praktycznego zastosowania chromatografii typu flash

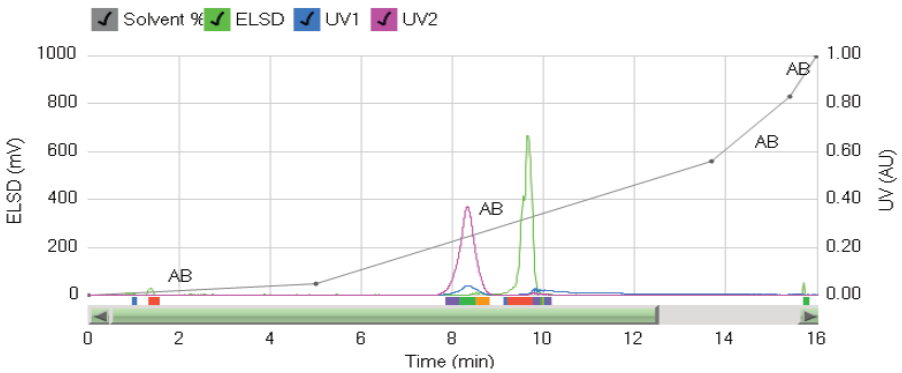
2. Rozdzielenie ekstraktu *Caryophylli flos* za pomocą chromatografii typu flash.

Przy pomocy techniki chromatografii typu flash możliwe jest rozdzielanie ekstraktu kwiatów Goździkowca zawierającego eugenol. Związek ten ze względu na swoje właściwości znalazł zastosowanie m.in. stomatologii, przemyśle farmaceutycznym oraz aromaterapii. Stosowany jest jako środek przeciwbólowy, antyseptyczny oraz rozgrzewający.[6] Posiada również właściwości antyoksydacyjne. Lipidy błony komórkowej mogą ulegać peroksydacji pod wpływem wolnych rodników, co prowadzi do śmierci komórki, a także wywołuje szereg chorób, takich jak miażdżyca, cukrzyca, czy też nowotwory. Rośliny występujące naturalnie, mają tendencję do hamowania peroksydacji lipidów ze względu na obecność związków fenolowych takich jak flawonoidy oraz ich estry.

Wykorzystując sprzężoną pracę detektorów UV oraz ELSD do detekcji składników w odbieranych frakcjach chromatograficznych możliwe było skrócenie czasu analizy (*Rysunek*

[5] Guide: *Chromapedia: Detection Methods*, BUCHI Laboratory Equipment, 2020, 39-42.
[6] K. Nowak, J. Ogonowski, M. Jaworska, K. Grzesik. *Olejek goździkowy - właściwości i zastosowanie*. Chemik, 2012, 66(2), 145-152.

2). W ten sposób zoptymalizowano rozdział ekstraktu, czyniąc go bardziej wydajnym i produktywnym, nawet w obecności małych ilości składników próbki.[7]



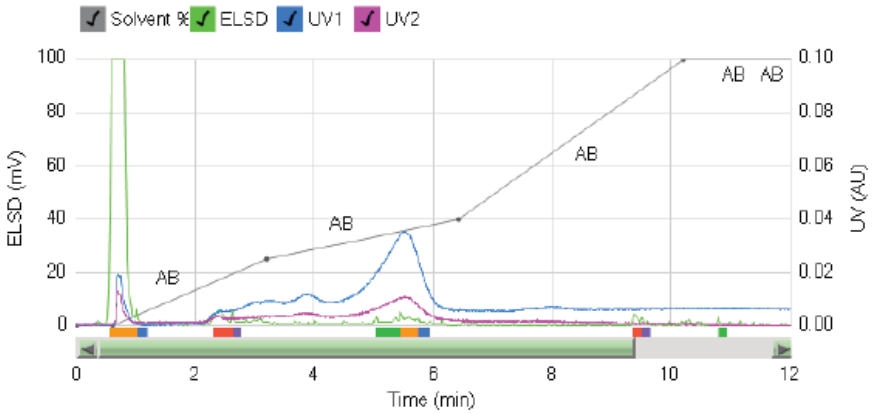
Rysunek 2. Oddzielenie ekstraktu Caryophylli flos za pomocą chromatografii typu flash.
Źródło: Application Note Book Pure Technologies. A Full Spectrum of Purification Solutions.

3. Oczyszczanie ekstraktu z korzenia *Pseudolarix kaemferi*

Wraz ze wzrostem zainteresowania branży leczniczej naturalnymi środkami przeciwgrzybicznymi pozyskiwanymi z kory/korzeni niektórych roślin konieczne stało się ustalenie analitycznej metody kontroli jakości uzyskiwanych ekstraktów. Do tego celu wykorzystano np. zintegrowany system chromatografii typu flash, Pure C-815, za pomocą którego analizowano obecność diterpenoidów, takich jak kwas pseudolarowy B zawarty w korzeniu *Pseudolarix kaemferi*. Kwas ten okazał się najbardziej aktywnym środkiem przeciwgrzybiczym, wyizolowanym z roślin wyższych. Wykazuje silne działanie przeciwgrzybicze przeciwko gatunkom *Candida* oraz *Torulopsis*. Sygnał ELSD pokazał elucję związków niepolarnych, podczas gdy detektor UV umożliwił wykrycie związków na niskich poziomach stężeń (*Rysunek 3*). Dalszą analizę jakości otrzymanego produktu przeprowadzono z wykorzystaniem techniki HPLC. [8]

[7] M. Ogata, M. Hoshi, S. Urano, T. Endo. *Antioxidant activity of Eugenol and related monomeric and dimeric compounds*, Chem. Phar. Bull, 2000, 1467-1469

[8] C. Qiao, Q. Han. J. Song, S. Mo, C. Tai, H. Xu. *HPLC analysis of Bioactive diterpenoids from the root bark of pseudolarix kaemferi*, J. of Food and Drug Analysis, 2006 ,14(4), 353-356.

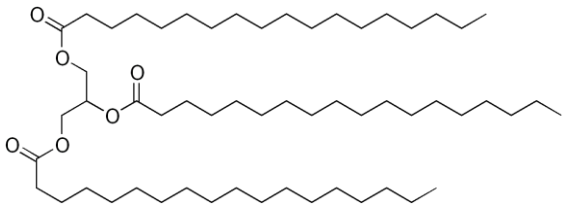


Rysunek 2. Oczyszczanie ekstraktu przy użyciu chromatografii Pure Flash.

Źródło: Application Note Book Pure Technologies. A Full Spectrum of Purification Solutions.

4. Separacja mono- di-, triglicerydów przy użyciu fazy C18 i aminowej

Trójglicerydy to kwasy tłuszczowe występujące w postaci estrów w połączeniu z gliceryną. Jednym z ich przedstawicieli jest tristearynian glicerolu, który występuje zarówno u roślin, jak i zwierząt i może być stosowany jako nośnik leku dla docelowych związków leczniczych.[9]



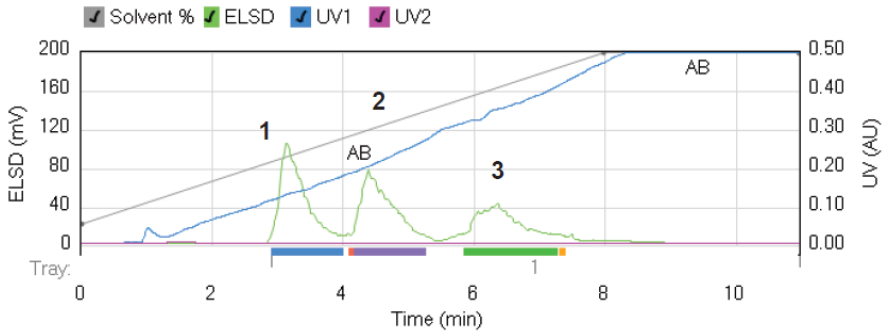
Rysunek 3. Struktura tristearynianu glicerolu.

Źródło: Opracowanie własne

Używając systemu chromatografii flash Pure C-815 wraz z kolumnami wypełnionymi selektywnymi fazami stacjonarnymi zoptymalizowano sposób separacji mono- di- i triglicerydów. Korzystne okazało się zastosowanie do tego celu fazy C18 i aminowej.

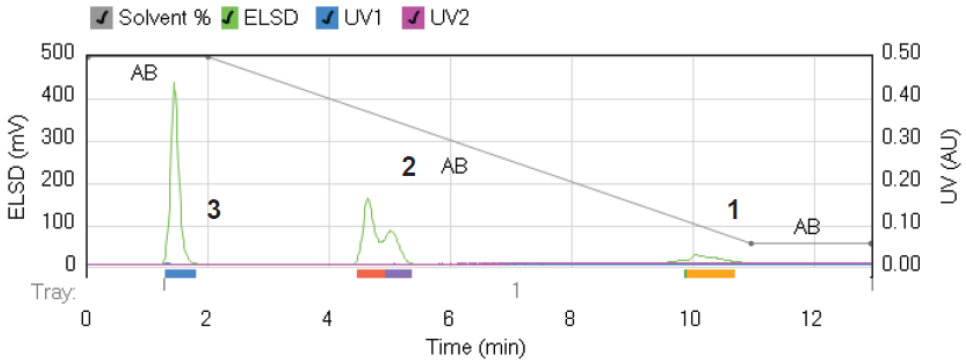
Tak zwana faza C18 wykorzystuje w procesie separacji oddziaływania hydrofobowe, podczas gdy faza aminowa oddziaływania lipofilowe pomiędzy fazą stacjonarną a łańcuchem tłuszczowym. Istotną rolę w procesie rozdziału (elucji i retencji) składników mieszaniny glicerydów odegrał również odpowiedni dobór gradientu rozpuszczalników. Technologia detekcji Pure w systemie chromatografii flash Pure C-815 pozwoliła na szybkie wyizolowanie oraz oczyszczenie związków na bazie lipidów, które nie są chromoforowe (*Rysunek 4 i 5*).

[9] P. Dewick, J. Wiley & Sons, Inc. *Medicinal natural products*; *A biosynthetic approach*, 3rd edition; Hoboken, UK, 2009, 40-44.



Rysunek 4. Separacja mono- di-, triglicerydów przy użyciu fazy C18.

Źródło: Application Note Book Pure Technologies A Full Spectrum of Purification Solutions.



Rysunek 5. Separacja mono- di-, triglicerydów przy użyciu fazy aminowej.

Źródło: Application Note Book Pure Technologies. A Full Spectrum of Purification Solutions.

5. Podsumowanie

Poprzez zastosowanie techniki chromatografii typu flash, która w dzisiejszych czasach została całkowicie zautomatyzowana, możliwe jest przyspieszenie procesu separacji związków z ich mieszanin. Ponadto przyczynia się do zwiększenia wydajności samej izolacji i oczyszczania produktów. Warto zwrócić uwagę, że oprócz oszczędności czasu związanego z samą analizą, zastosowanie wstępnego oczyszczania przy użyciu chromatografii typu flash przyczynia się do osiągnięcia znacznie lepszej rozdzielczości pomiędzy analizowanymi pikami. W przypadku związków ulegających w łatwy sposób degradacji lub zmianie podczas rozdziału chromatograficznego, zastosowanie flash chromatografii pozwala na ich odzysk z większą czystością, ze względu na krótszy czas kontaktu z układem chromatograficznym. Zastosowanie chromatografii typu flash do oczyszczania, frakcjonowania oraz izolacji związków docelowych, pozwala w znaczący sposób uprościć oczyszczanie z zastosowaniem preparatywnej chromatografii wysokociśnieniowej (HPLC).

Literatura

1. Z. Witkiewicz, J. Kałużna – Czaplińska, Podstawy chromatografii i technik elektromigracyjnych, Wydawnictwo Naukowe PWN, Warszawa 2017, 21-26.
2. Z. Witkiewicz, W. Wardencki, Chromatografia gazowa. Teoria i praktyka, Wydawnictwo Naukowe PWN, Warszawa 2018, 11-12.
3. N. Surve, A. Thomas R. Bhole, C. Patil, Flash Chromatography and Semi-Preparative HPLC: Review on the Applications and Recent Advancements over the Last Decade, Eurasian Journal of Chemistry, 2023, 28.1 (109)
4. A. B. Roge, S. N. Firke, R. M. Kawade, S. K. Sarje, S. M. Vadvalkar, Brief review on: flash chromatog-raphy. International Journal of Pharmaceutical Sciences and Research, 2011, 2(8), 1930-1937.
5. Guide: Chromapedia: Detection Methods, BUCHI Laboratory Equipment, 2020, s. 39-42.
6. K. Nowak, J. Ogonowski, M. Jaworska, K. Grzesik. Olejek goździkowy - właściwości i zastosowa-nie. Chemik, 2012, 66(2), 145-152.
7. M. Ogata, M. Hoshi, S. Urano, T. Endo. Antioxidant activity of Eugenol and related monomeric and di-meric compounds, Chem. Phar. Bull, 2000, 467-1469
8. C. Qiao, Q. Han. J. Song, S. Mo, C. Tai, H. Xu. HPLC analysis of Bioactive diterpenoids from the root bark of pseudolarix kaemferi, J. of Food and Drug Analysis, 2006 ,14(4), 353-356.
9. P. Dewick, J. Wiley & Sons, Inc. Medicinal natural products; A biosynthetic approach, 3rd edition; Hobo-ken, UK, 2009, 40-44.

Źródła internetowe

1. https://haas.com.pl/produkt/sepaflash-ilok-sl/ (dostęp: 29.04.2024)

KOŁO

NAUKOWE

INŻYNIERII

MEDYCZNEJ

X-MED

Abigail Machaj
Studenckie Koło Naukowe X-Med

mgr inż. Wiktoria Wojnarowska
Opiekun naukowy

Personalizowana orteza dłoni wykonana z zastosowaniem druku 3D

Tradycyjna metoda tworzenia ortez jest czasochłonna i kosztowna. Szybki rozwój druku 3D otworzył nowe możliwości w tej dziedzinie. Stabilizatory wykonane technologią druku 3D są lżejsze, bardziej higieniczne i estetycznie dopasowane. Celem artykułu było przedstawienie zastosowania technik takich jak skanowanie, modelowanie i druk 3D do tworzenia indywidualnych, spersonalizowanych ortez. Proces obejmował uzyskanie cyfrowego modelu 3D górnej kończyny pacjenta poprzez jej zeskanowanie oraz utworzenie modelu ortezy w programie Autodesk Inventor. Model został wytworzony przy pomocy jednej z technologii druku 3D - osadzania stopionego materiału (ang. fused deposition modeling, FDM). Efektem była orteza dostosowana do indywidualnych potrzeb pacjenta. Nowoczesne techniki inżynierskie umożliwiają uzyskanie ortezy idealnie dopasowanej do anatomicznych cech pacjenta, co zwiększa komfort noszenia i poprawia jakość życia.

Słowa kluczowe: druk 3D, szybkie prototypowanie, skanery 3D, ortezy, FDM.

1. Wprowadzenie

Współczesna medycyna wymaga nowoczesnych rozwiązań, które mogą sprostać rosnącym oczekiwaniom pacjentów w krajach o wysokim i średnim poziomie rozwoju. Dotychczasowe metody leczenia są nieustannie doskonalone dzięki postępom nie tylko w medycynie, ale także w technologii i inżynierii. W ostatnich latach popularność zyskało szybkie prototypowanie, zwłaszcza wytwarzanie przyrostowe, jako alternatywa w procesie tworzenia prototypów. Ze względu na szybkość i niskie koszty, metody te zyskały zainteresowanie nie tylko we wzornictwie przemysłowym, projektowaniu, motoryzacji i przemyśle, ale również w medycynie, inżynierii biomedycznej oraz inżynierii tkankowej. Druk 3D jest wykorzystywany w medycynie do wytwarzania spersonalizowanych wyrobów medycznych, takich jak protezy, implanty, narzędzia chirurgiczne oraz ortezy, które są dobrze dopasowane do indywidualnej anatomii pacjenta.

Orteza to zewnętrzne urządzenie medyczne stosowane w celu stabilizacji, ochrony, wsparcia lub korekcji dysfunkcji układu mięśniowo-szkieletowego. Ortezy są najczęściej kojarzone z ograniczeniem ruchomości stawów i wykorzystywane w tym celu. Jednakże służą one nie tylko do usztywniania złamanych kończyn; popularne są także warianty pooperacyjne,

rekonwalescencyjne oraz rehabilitacyjne. Między innymi są one rutynowo przepisywane w celu poprawy zdolności ruchowej dzieci i dorosłych z zaburzeniami neurologicznymi, takimi jak porażenie mózgowe, choroba Charcota-Marie-Tootha, udar mózgu i stwardnienie rozsiane[1].

Ortezy mogą być stosowane zarówno w terapii pourazowej, jak i w prewencji kontuzji, szczególnie u osób aktywnych fizycznie. Istnieją różne rodzaje ortez, w zależności od kończyny, którą mają usztywniać, takie jak ortezy nadgarstka, kolana czy kostki. Tradycyjne stabilizatory nadgarstka są zazwyczaj ręcznie wykonywane z gipsu, formowanego bezpośrednio na nadgarstku pacjenta. To podejście, choć powszechne, jest pracochłonne, oferuje ograniczone opcje projektowe, może być kosztowne i często wiąże się z długim czasem oczekiwania. Dodatkowo, gipsowe stabilizatory są ciężkie, a ich długotrwałe noszenie osłabia siłę mięśniową kończyny. Nie przepuszczają powietrza i nie można ich moczyć, co utrudnia utrzymanie higieny. Jednym z problemów jest również alergia na gips, która występuje u niektórych pacjentów. Ponadto gips szybko się on kruszy, co obniża trwałość takich ortez[2].

Wytwarzanie przyrostowe ma ogromny potencjał do wyeliminowania kilku kroków związanych z tradycyjnymi metodami produkcji ortez nadgarstka. Drukowanie 3D umożliwia między innymi swobodę projektowania, pozwalając na odejście od tradycyjnych paradygmatów projektowych i tym samym umożliwia opracowanie lepszych ortez nadgarstka. Dzięki metodom szybkiego prototypowania (ang. Rapid prototyping, RP) ortezy nadgarstka mogą być optymalizowane do indywidualnych wymagań biomechanicznych, aby zapewnić nie tylko lepszą funkcjonalność i dopasowanie, ale również uwzględnić aspekty estetyczne.

Całość procedury można opisać w 5 krokach[3]:

1. Wykorzystanie skanera 3D do uzyskania cyfrowego obrazu wybranej kończyny pacjenta.
2. Opracowanie cyfrowego modelu ortezy.
3. Wydrukowanie fizycznej ortezy przy użyciu drukarki 3D.
4. Dokończenie procesu obejmuje dopasowanie kształtu, wygładzenie krawędzi i dodanie elementów stabilizujących, takich jak rzepy.
5. Weryfikacja i sprawdzenie dopasowania ortezy, najlepiej przez certyfikowanego terapeutę zajęciowego lub ortopedę z przeszkoleniem w zakresie wytwarzania ortez.

---

[1] Oud T. A. M., Lazzari E., Gijsbers H. J. H., Gobbo M., Nollet F., Brehm M. A. *Effectiveness of 3D-printed orthoses for traumatic and chronic hand conditions: A scoping review*, PLOS One, 2021, 16(11), e0260271.

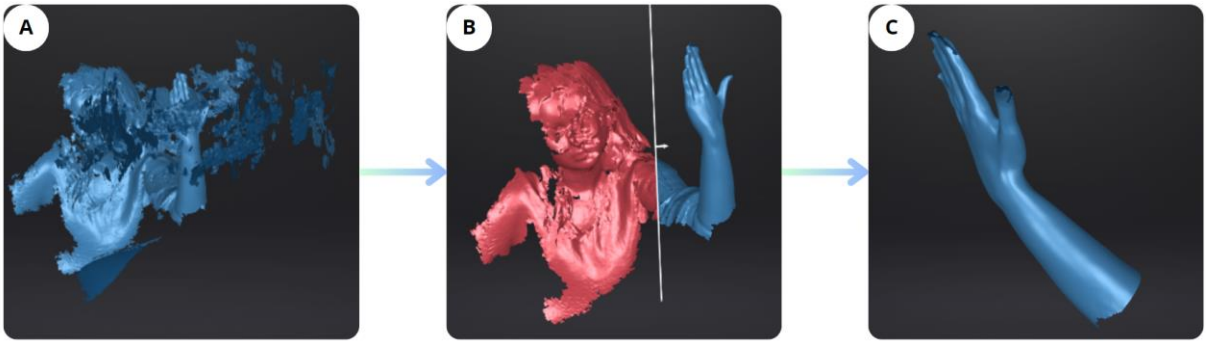
[2] Haque A., Parsons H., Parsons N., Costa M., Redmond A., Mason J., Nwankwo H., Kearney R., *Use of cast immobilization versus removable brace in adults with an ankle fracture: two-year follow-up of a multicentre randomized controlled trial*, The Bone & Joint Journal, 2023, 105-B(4), s. 382-388.

[3] Schwartz D. A., & Schofield K. A., *Utilization of 3D printed orthoses for musculoskeletal conditions of the upper extremity: A systematic review*, Journal of Hand Therapy, 2023, 36(1), s. 166-178.

Dlatego celem artykułu było przedstawienie możliwości zastosowania nowoczesnych technologii, takich jak skanowanie 3D, zaawansowane modelowanie komputerowe i druk 3D, do tworzenia indywidualnych, spersonalizowanych ortez nadgarstka. Na początku realizacji projektu założono, że orteza będzie wyrobem indywidualnym, przeznaczona dla osoby z urazem w obrębie nadgarstka. Orteza miała za zadanie usztywniać ten region. Ponadto orteza miała być łatwa do zakładania i zdejmowania, mieć prostą konstrukcję i małą wagę. Opracowanie ortezy obejmowało wykonanie cyfrowego modelu anatomii górnej kończyny pacjenta przy pomocy skanowania 3D, wykonanie cyfrowego modelu ortezy przy pomocy oprogramowania do projektowania wspomaganego komputerowo (ang. Computer aided design, CAD), wykonanie fizycznego modelu ortezy przy pomocy metod szybkiego prototypowania oraz wykończenie ortezy m.in. poprzez dodanie elementów stabilizujących.

2. Wykonanie skanu

W ramach projektu wykonano skan lewej ręki przy użyciu skanera Shining 3D Einstar. Jest to skaner wyposażony w trzy podczerwone projektory emitujące bezpieczne dla oczu światło strukturalne VCSEL. Proces skanowania przeprowadzono przy użyciu dedykowanego oprogramowania EXStar Software, zaprojektowanego specjalnie do obsługi tego skanera. Proces skanowania rozpoczął się od przygotowania ręki badanego do skanowania, co obejmowało ustawienie kończyny w odpowiedniej pozycji oraz zapewnienie optymalnych warunków oświetleniowych. Cała procedura skanowania trwała mniej niż trzydzieści minut, co świadczy o efektywności i zaawansowaniu technologii zastosowanej w urządzeniu. W wyniku skanowania uzyskano wirtualną chmurę punktów odpowiadającą geometrii ciała pacjenta.

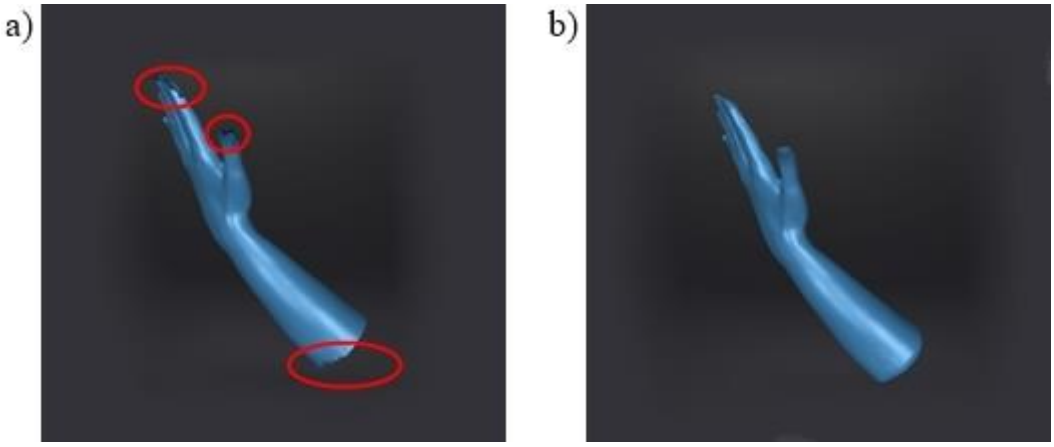


Rysunek 6. Trzy etapy przetwarzania modelu 3D lewej kończyny górnej: (A) surowy skan z artefaktami, (B) wyodrębnienie wybranego obszaru modelu poprzez podział płaszczyzną tnącą, (C) wyodrębniony cyfrowy model kończyny
Źródło: opracowanie własne.

Początkowy model 3D, widoczny na rysunku 2A, zawierał liczne artefakty. Mogły one wynikać z różnych czynników, takich jak ruch ręki w trakcie skanowania czy refleksy świetlne. Do usunięcia artefaktów zastosowano narzędzia dostępne w oprogramowaniu EXStar Software. Proces ten polegał na wykorzystaniu algorytmów filtrujących, które automatycznie identyfikowały i eliminowały niepożądane elementy.

Kolejnym etapem była segmentacja modelu, czyli wyodrębnienie wybranego obszaru od reszty zeskanowanego obiektu. Zostało to osiągnięte poprzez przecięcie modelu za pomocą odpowiednio ustawionej płaszczyzny tnącej (rys. 2B). Było to kluczowe, aby uzyskać model obejmujący wyłącznie część ramienia i dłoni, na które nakładana miała być orteza (rys. 2C). Dzięki wyodrębnieniu tylko tej części zeskanowanego modelu zmniejszono rozmiar pliku. To z kolei sprawiło, że dalsze przetwarzanie modelu wymagało mniejszej mocy obliczeniowej.

Jednakże cyfrowy model kończyny w dalszym ciągu charakteryzował się pewnymi nieprawidłowościami, w tym ubytkami w obrębie palców (rys. 3a). Wobec tego poddano go dalszym korektom. Naprawa modelu została wykonana w tym samym oprogramowaniu poprzez zastosowanie dostępnych narzędzi edycyjnych. Jednym z nich z nich było narzędzie umożliwiające rekonstrukcję brakujących fragmentów modelu.

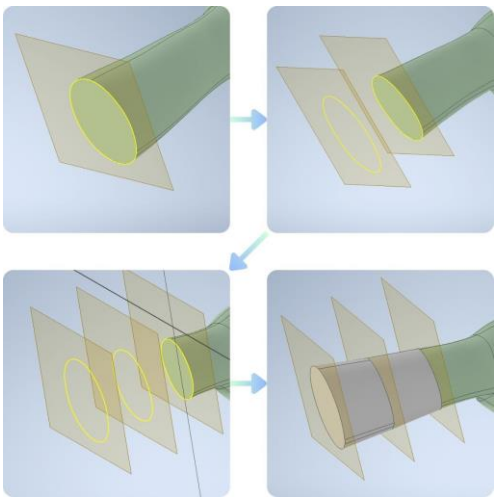


Rysunek 7. Cyfrowy model kończyny górnej: a) z zaznaczonymi ubytkami, b) po uzupełnieniu ubytków powierzchni w oprogramowaniu EXStar – model końcowy
Źródło: opracowanie własne.

W wyniku wprowadzonych poprawek otrzymano cyfrowy model kończyny górnej przedstawiony na rysunku 3b. Model 3D ręki był nie tylko wolny od zakłóceń, ale także odzwierciedlał rzeczywiste proporcje i struktury anatomiczne. Otrzymany model posłużył jako odniesienie podczas modelowania docelowej ortezy, umożliwiając spełnienie założenia o indywidualnym charakterze ortezy.

3. Modelowanie ortezy

Ortezę zamodelowano w środowisku Autodesk Inventor. Pierwszym etapem było zaimportowanie modelu kończyny do tego programu. Import wymagał konwersji formatu pliku, ponieważ początkowy model był chmurą punktów. Konwersję te wykonano w oprogramowaniu SpaceClaim poprzez aproksymację krzywych splajnowych. Polegała ona na pokryciu chmury punktów elementarnymi płatami powierzchni typu NURBS (ang. non-uniform rational b-spline). Otrzymany model zapisano w formacie STEP.



Rysunek 8. Etapy tworzenia modelu w inventorze
Źródło: opracowanie własne.

Po zaimportowaniu modelu do Inventora, pierwszym krokiem było wygenerowaniu kilku przekrojów kończyny pacjenta. Dokonano tego poprzez podzielenie modelu wzdłuż osi ręki za pomocą płaszczyzn. Na utworzonych płaszczyznach stworzono obrysy kończyny. Następnie użyto funkcji wyciągnięcia złożonego, umożliwiającego zaznaczenie tych obrysów i utworzenie figury na ich podstawie. Dzięki temu uzyskano model ręki oparty na wcześniejszym skanie, na którym można było dalej pracować (Rysunek 4).

Następnym krokiem było stworzenie offsetu, co polegało na utworzeniu nowej bryły odsuniętej od ręki o 5 mm. Zastosowanie odsunięcia miało na celu zapewnienie odpowiedniego luzu między skórą pacjenta a ortezą. Nawet niewielki luz zapewnienia komfort użytkowania ortezy poprzez wyeliminowanie ucisku w miejscach wrażliwych. Następnie dla wykonanej bryły stworzono kolejny offset, który tworzył bryłę również odsuniętą o 5 mm, co pozwoliło uzyskać szkielet ortezy. Po wykonaniu tych operacji można było wyłączyć model ręki, jak i pierwszego offsetu i pracować bezpośrednio na szkielecie ortezy. Przedstawione postępowanie umożliwiło uzyskanie modelu dopasowanego do kształtu kończyny pacjenta. Takie

dopasowanie ogranicza ryzyko ucisku oraz powstawania otarć, czego nie zapewniają ortezy dostępne w sklepach medycznych.

W dalszej kolejności na szkielecie ortezy dodano wycięcia na palce oraz wycięcie umożliwiające wkładanie ręki do ortezy. Dodano również otwory wentylacyjne. Kolejnym etapem było zaokrąglenie wszystkich części, aby wyeliminować ewentualne źródła dyskomfortu. Dodano również miejsca do mocowania rzepów, co pozwalało na stabilne zamknięcie ortezy. Miało to na celu zapobiec jej spadaniu podczas użytkowania, co okazało się bardziej praktycznym rozwiązaniem niż podzielenie ortezy na dwie części. Dzięki rzepom użytkownik może łatwiej dostosować ortezę do codziennych potrzeb rehabilitacyjnych.

Ostateczny projekt ortezy uwzględniał nie tylko funkcjonalność, ale także komfort użytkowania. Wzór otworów na ortezie został zaprojektowany w taki sposób, aby zapewnić odpowiednią wentylację skóry, co zwiększało komfort noszenia i zapobiegało podrażnieniom. Orteza, oprócz zapewnienia wsparcia i stabilizacji, umożliwia także właściwe oddychanie skóry, co jest szczególnie ważne podczas długotrwałego noszenia.

4. Wydruk i postprocessing

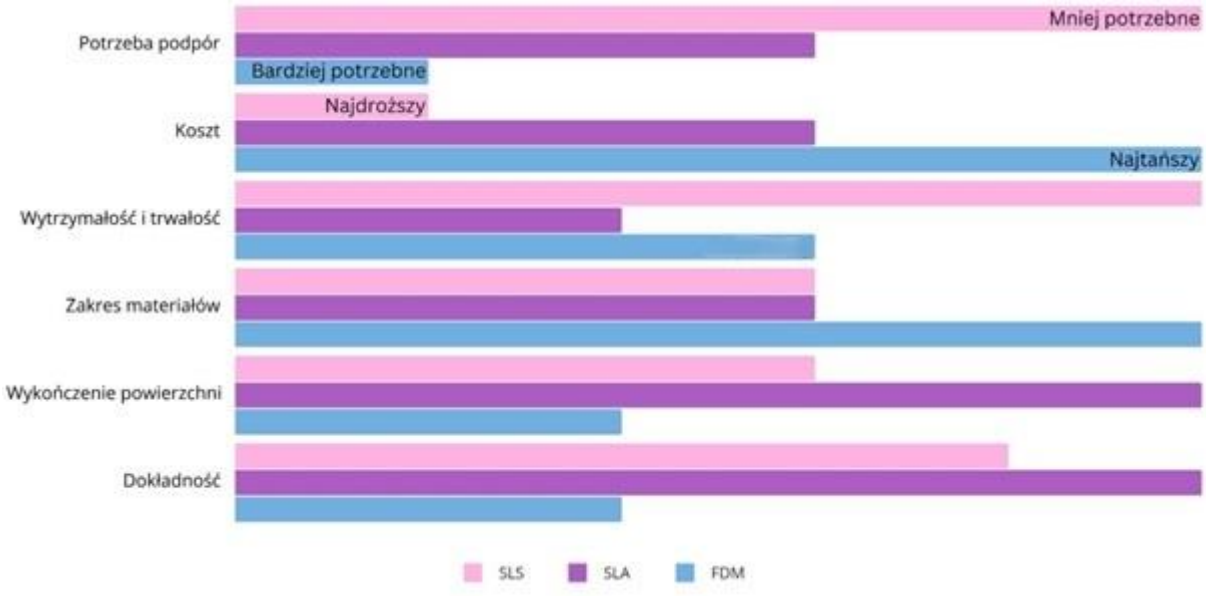
4.1. Wybór technologii wytwarzania

By wytworzyć opracowaną ortezę należy dobrać odpowiednią technikę wytwarzania. Niedoskonałości tradycyjnych technik wytwarzania ortez zainspirowały wiele jednostek do eksperymentowania z technologiami szybkiego prototypowania, aby tworzyć coraz bardziej wygodne i funkcjonalne ortezy. Metody RP to to techniki pozwalające na szybkie i efektywne tworzenie fizycznych modeli lub prototypów komponentów za pomocą zaawansowanych technologii, takich jak metody przyrostowe (ang. additive manufacturing, AM) oraz obróbka CNC. Metody przyrostowe opierają się na tworzeniu fizycznego obiektu trójwymiarowego metodą warstwa po warstwie[4].

Do technik szybkiego prototypowania należą głównie stereolitografia (ang. stereolithography, SLA), selektywne spiekanie laserowe (ang. selective laser sintering, SLS), modelowanie osadzania stopionego materiału (ang. fused deposition modeling, FDM) oraz druk 3D[5]. Na rysunku 1 przedstawiono porównanie trzech metod RP: FDM, SLA i SLS pod

[4] Laska-Leśniewicz A., *Wykorzystanie metod szybkiego prototypowania (rapid prototyping) w nowoczesnej medycynie*, Zeszyty Naukowe Towarzystwa Doktorantów Uniwersytetu Jagiellońskiego. Nauki Ścisłe, 2017, 15(2), s. 39–48.
[5] Ibidem.

względem konieczności stosowania podpór, kosztów, wytrzymałości wytworzonych modeli, zakresu stosowanych materiałów, jakości wykończenia oraz dokładności wytwarzania.



Rysunek 9. Porównanie pod względem ilości podpór, kosztów, wytrzymałości, zakresu materiałów, wykończenia i precyzji metod FDM, SLA i SLS
Źródło: opracowanie własne na podstawie źródła[6].

Przedstawione dane wskazują na istotne różnice między technologiami druku 3D, takimi jak FDM, SLA i SLS, które mają kluczowe znaczenie dla ich zastosowania w produkcji ortez i innych zaawansowanych aplikacjach. Poniżej szczegółowo omówiono te różnice pod kątem poszczególnych parametrów:

- Podpory:

W przypadku technologii SLA, wymagane jest więcej podpór w porównaniu do FDM i SLS. FDM charakteryzuje się umiarkowaną ilością podpór, natomiast SLS zazwyczaj nie wymaga ich wcale, co może być dużą zaletą przy wytwarzaniu bardziej złożonych modeli.

- Koszty:

FDM jest najbardziej ekonomiczną metodą druku 3D. Koszty druku za pomocą SLA są znacznie wyższe z powodu droższych materiałów oraz większej ilości wymaganego sprzętu. SLS, choć oferuje wiele zalet, jest również kosztowny, co może ograniczać jego zastosowanie w projektach wymagających optymalizacji budżetowej..

- Wytrzymałość i trwałość:

[6] FDM vs. SLA vs. SLS: Which is the best 3D Printing Technology for Your Project?: https://www.linkedin.com/pulse/fdm-vs-sla-sls-which-best-3d-printing-technology-your-jer%C3%B3nimo?trk=public_profile_article_view (dostęp: 12.06.2024).

SLS oferuje najwyższą wytrzymałość i trwałość drukowanych obiektów, co czyni go idealnym do zastosowań wymagających dużej odporności mechanicznej. Z kolei FDM i SLA charakteryzują się umiarkowaną wytrzymałością. Ta jednak jest wystarczająca dla wielu aplikacji, zwłaszcza w medycynie. Tam są używane do tworzenia spersonalizowanych ortez i innych urządzeń medycznych.

- Zakres materiałów:

FDM pozwala na stosowanie szerokiego zakresu materiałów, co zapewnia większą elastyczność w doborze surowców do specyficznych potrzeb projektu. SLA i SLS także oferują różnorodność materiałów, jednak są one często bardziej specjalistyczne i droższe.

- Wykończenie powierzchni:

SLA przewyższa inne technologie pod względem jakości wykończenia powierzchni. Obiekty drukowane metodą SLA charakteryzują się bardzo gładką powierzchnią, co jest istotne w przypadku przedmiotów wymagających wysokiej precyzji i estetyki. Z kolei FDM i SLS oferują wykończenie na zadowalającym poziomie, choć mogą wymagać dodatkowej obróbki post-processingowej.

- Dokładność wymiarowa:

SLA i SLS oferują większą dokładność wykonania modeli w porównaniu do FDM, co jest bardzo ważne w aplikacjach, gdzie dokładność wymiarowa jest kluczowa. FDM, mimo że mniej dokładne, często wystarcza do wielu zastosowań, zwłaszcza w prototypowaniu i wytwarzaniu modeli medycznych.

Podsumowując, analizowane metody RP charakteryzują się różnymi wadami i zaletami. Wybór konkretnej metody powinien wynikać z konkretnych wymagań projektowych. W przypadku opracowywanej ortezy najważniejsze były czynniki ekonomiczne oraz łatwość wytworzenia. Dlatego też do wykonania ortezy wybrano technologię FDM, która ze wszystkich analizowanych metod charakteryzuje się najniższymi kosztami druku. Dodatkowo, technologia FDM nie wymaga użycia rozpuszczalników, zapewnia dużą łatwość i elastyczność w obsłudze oraz przetwarzaniu materiałów. Należy jednak zauważyć, że technologia ta ma pewne ograniczenia w wyborze materiałów do wytwarzania modeli, skupiając się głównie na polimerach termoplastycznych, co w konkretnym przypadku nie było istotnym czynnikiem.

4.2. Wybór materiału do wykonania ortezy

Do wytworzenia i wykończenia ortez bardzo istotne jest odpowiednie dobranie materiałów. Najbardziej popularnymi wariantami są termoplastyczne polimery, takie jak polipropylen (PP) i polietylen (PE), które mogą być formowane na ciepło, aby idealnie dopasować się do anatomii

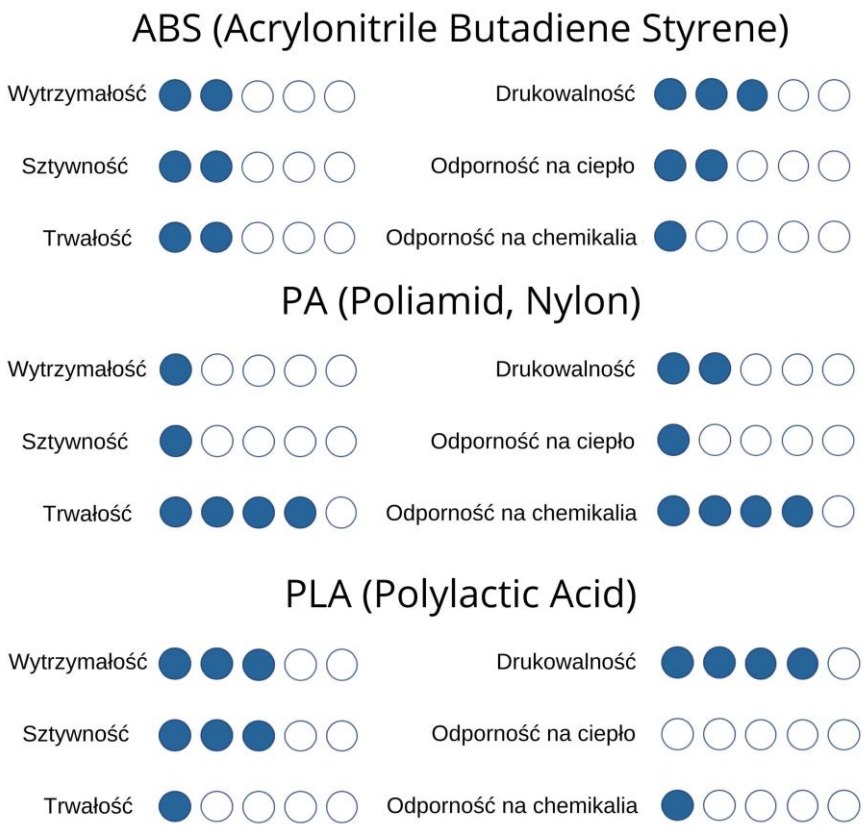
pacjenta. Ponadto stosuje się również materiały kompozytowe, takie jak włókna węglowe lub szklane połączone z żywicami, które zapewniają lekkość i wytrzymałość. Metale, takie jak stopy aluminium, tytanu i stal nierdzewna, są używane w konstrukcjach wymagających dużej wytrzymałości. Pianki i żele amortyzujące, takie jak pianki EVA (etylen-octan winylu) i silikonowe wkładki, zapewniają komfort i ochronę przed uciskiem. Wykorzystuje się także neopren, spandex i inne elastyczne materiały, które zapewniają wsparcie i kompresję. Do druku 3D metodą FDM można używać materiałów takich jak polilaktyd (PLA), akrylonitryl-butadien-styren (ABS), nylon, poli(tereftalan etylenu z domieszką glikolu) (PETG), termoplastyczny poliuretan (TPU), polistyren wysokoudarowy (HIPS) i poli(alkohol winylowy) (PVA), które oferują różnorodne właściwości fizyczne i chemiczne. Podczas projektowania ortezy do jej wydruku dostępne były trzy materiały: ABS, PA i PLA. Na rysunku 5 przedstawiono porównanie wybranych właściwości tych materiałów.

ABS charakteryzuje się wysoką wytrzymałością, sztywnością i trwałością. Jest odporny na ciepło i chemikalia, co czyni go odpowiednim do zastosowań wymagających dużej odporności mechanicznej. Jego drukowalność jest również wysoka, co pozwala na uzyskanie dokładnych modeli. Jednak ABS jest mniej ekologiczny w porównaniu do PLA i wymaga wyższej temperatury druku, co może wpływać na koszty energetyczne.

PA wyróżnia się wyjątkową wytrzymałością i trwałością, a także odpornością na chemikalia i ciepło. Jest elastyczny i odporny na ścieranie, co czyni go idealnym do zastosowań wymagających wysokiej odporności na obciążenia mechaniczne. Drukowanie nylonem może być jednak bardziej wymagające ze względu na jego higroskopijność, co oznacza, że materiał ten pochłania wilgoć z otoczenia.

PLA jest materiałem biodegradowalnym i łatwym do drukowania, co sprawia, że jest przyjazny dla środowiska i odpowiedni dla mniej doświadczonych użytkowników. Oferuje umiarkowaną wytrzymałość i sztywność, które są wystarczające dla wielu zastosowań medycznych, takich jak ortezy. Jego odporność na ciepło i chemikalia jest mniejsza niż w przypadku ABS i nylonu, jednak jego potencjał ekologiczny i łatwość przetwarzania sprawiają, że jest odpowiedni do wykonania ortezy. Ponadto PLA charakteryzuje się biokompatybilnością, co jest ważne przy wyrobach mających styczność z ludzkim ciałem[7].

[7] Da Silva D., Kaduri M., Poley M., Adir O., Krinsky N., Shainsky-Roitman J., Schroeder A., *Biocompatibility, biodegradation and excretion of polylactic acid (PLA) in medical implants and theranostic systems*, Chemical Engineering Journal, 340, 2018, s. 9-14.



Rysunek 10. Porównanie wytrzymałości, sztywności, trwałości, łatwości wydruku, odporności na ciepło i chemikalia dla ABS, PA (nylon) i PLA
Źródło: opracowanie własne na podstawie źródła[8]

Rozwijając kwestię ekologiczności PLA (i pochodnych) to badania wykazały spory potencjał tego materiału w dążeniu do zrównoważonego rozwoju. Przykładowo w 2020 roku produkcja i zużycie PLA na całym świecie szacowano na około 800 000 ton rocznie. Wielkość ta zwiększa się z roku na rok, w 2010 zużycie wynosiło tylko 120 000 ton w porównaniu z niedawnymi statystykami. Dane te sugerują, że popularność PLA tylko rośnie i rośnie. Dlatego powstają nowe różne mieszanki takie jak Wound Up, który jest kompozytem PLA i ziaren kawy pochodzących z recyklingu Z drugiej strony należy zaznaczyć, że nawet ten najtańszy PLA jest biodegradowalny. Na szybkość biodegradacji PLA wpływa stopień jego krystaliczności (poprawę właściwości użytkowych uzyskano w Tajwańskim Przemysłowym Instytucie Technologii, w którym po 8 latach pracy opracowano nietoksyczny środek zarodkujący, przyśpieszający krystalizację tego materiału). Do PLA można dodawać różnego rodzaju napełniacze lub włókna. Poprzez zastosowanie mieszanki z polisacharydami, jak np. ze skrobią,

[8] PLA vs ABS vs Nylon: https://markforged.com/resources/blog/pla-abs-nylon (dostęp: 12.06.2024).

obniża się cenę, a przede wszystkim skraca czas rozkładu biologicznego. Z kolei celuloza w postaci włókien zwiększa sztywność i odporność na temperaturę[9].

Podsumowując, do wykonania ortezy wybrano materiał PLA, który przede wszystkim w przeciwieństwie do pozostałych materiałów charakteryzuje się biokompatybilnością. Ponadto ważne były korzyści ekologiczne. PLA jest biodegradowalny, łatwy do drukowania i przyjazny dla środowiska, co czyni go odpowiednim wyborem dla aplikacji medycznych. Dodatkowo, PLA jest stosunkowo ekonomiczny w porównaniu do innych materiałów, co również przemawiało za jego wyborem.

4.3. Przygotowanie modelu do druku 3D oraz proces wydruku

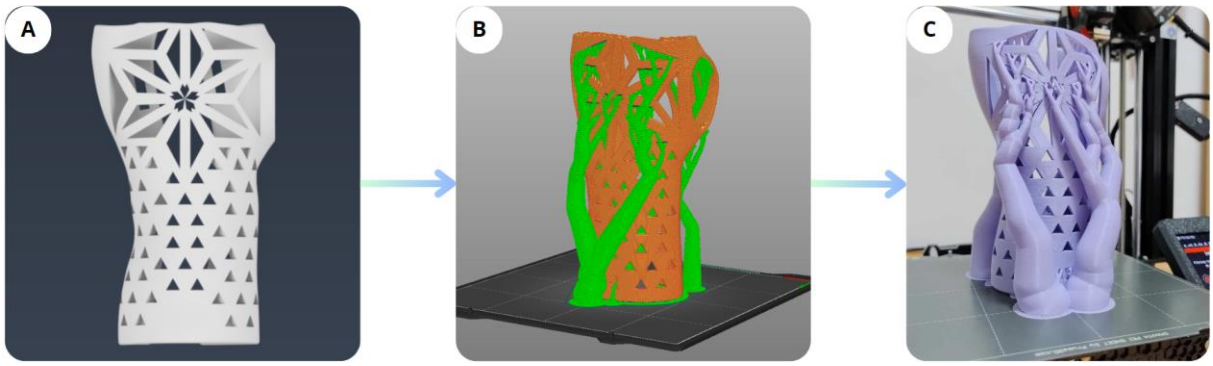
Kolejnym etapem było przygotowanie modelu do druku 3D. Proces ten rozpoczął się już w w programie Autodesk Inventor (Rysunek 6A), gdzie już na etapie projektowania, należało model zoptymalizować pod kątem druku. Po zakończeniu projektowania, model został wyeksportowany do formatu STL. Następnie, w odpowiednim oprogramowaniu, tzw. slicerze, przygotowano sekwencję ruchów głowicy, czyli G-codów. W tym przypadku był to PrusaSlicer. G-code to instrukcja dla drukarki 3D, która zawiera trajektorie, po których maszyna ma ekstrudować materiał. Oprócz trajektorii, G-code zawiera również informacje dotyczące m.in. wysokości drukowanej warstwy oraz procenta wypełnienia modelu[10]. Podczas przygotowania g-code'a, czyli tzw. plasterkowania, oprogramowanie dzieli model na poszczególne warstwy, które będą drukowane jedna po drugiej. Na tym etapie szczególną uwagę zwraca się na konstrukcję podpór, które wspierają model podczas druku. W tym przypadku zastosowano organiczne podpory, które są łatwe do usunięcia po zakończeniu druku, co minimalizuje zużycie filamentu.

Gotowy model został przesłany do drukarki 3D, w tym przypadku Prusa Mini, która została wybrana ze względu na krótki czas druku. Materiałem użytym do drukowania jest PLA, wybrany ze względu na swoje właściwości mechaniczne i biodegradowalność. Drukowanie trwało około 6 godzin. Efekt wykonania wydruku przedstawiono na rysunku 6C.

---

[9] Jaworska N., Podsiadło H., *Technologia druku 3D jako szansa dla środowiska naturalnego*, Acta Poligraphica, 2019, 14, s. 55-70.

[10] Wróbel, E. *Wielofunkcyjna orteza rehabilitacyjna dłoni z możliwością blokowania chwytu przeznaczona do wytwarzania w technologii przyrostowej* [w:] *Inżynieria biomedyczna. Metody przyrostowe w technice medycznej,* Politechnika Lubelska, Lublin 2016.



Rysunek 11. Etapy procesu drukowania. A - model w Inventorze. B - model zaimportowany do dedykowanego programu drukarki typu slicer, aby przygotować go do druku 3D. C - wydrukowany model
Źródło: opracowanie własne.

4.4. Postprocessing

Po zakończeniu druku model został zdjęty ze stołu roboczego drukarki 3D, a następnie usunięto podpory. Organiczne podpory użyte w tym procesie umożliwiają łatwe i szybkie ich usunięcie. Ich specjalna struktura umożliwia precyzyjne usunięcie ręczne, minimalizując ryzyko uszkodzenia drukowanego obiektu. Dzięki temu proces usuwania podpór jest efektywny i nie wymaga zastosowania skomplikowanych narzędzi ani dodatkowych obróbek, co przyspiesza produkcję i zapewnia wysoką jakość finalnego produktu.

Poza obróbką wydruku należało jeszcze w odpowiedni sposób wykończyć ortezę. Między innymi w celu zwiększenia komfortu pacjenta do wnętrza ortezy przyklejono specjalną gąbkę o dużych oczkach. Ten szczególny rodzaj gąbki został wybrany ze względu na swoje właściwości umożliwiające lepsze oddychanie skóry, co przyczynia się do redukcji pocenia się ręki oraz minimalizuje ryzyko podrażnień skórnych.

Ponadto we wcześniej zaplanowanych miejscach należało umieścić elementy zapinające, mające na celu zapewnienie odpowiedniego dopasowania i stabilności ortezy na kończynie pacjenta. W tym przypadku były to rzepy, które zostały przyszyte. Ten sposób zapięcia umożliwia użytkownikowi łatwe i szybkie dostosowanie ortezy do swoich indywidualnych potrzeb oraz gwarantuje, że orteza będzie się trzymała na miejscu. Efektem przeprowadzonych prac był fizyczny model ortezy przedstawiony na rysunku 7.

Poza zastosowanymi rozwiązaniami wykończenia ortezy, istnieją również pewne alternatywy, które warto rozważyć w zależności od specyficznych potrzeb użytkownika. Jedną z takich alternatyw dla gąbki o dużych oczkach jest żel silikonowy, powszechnie stosowany ze względu na zdolność do równomiernego rozkładania nacisku, co jest korzystne dla osób o wrażliwej skórze. Żel silikonowy zapewnia również doskonałą amortyzację i jest niezwykle trwały, co czyni go idealnym wyborem dla ortez przeznaczonych do długotrwałego

użytkowania. Inną opcją jest pianka z pamięcią kształtu, która dopasowuje się do konturów ręki użytkownika, zapewniając indywidualne wsparcie i zwiększając komfort. Ta technologia jest szczególnie przydatna w przypadkach rehabilitacji po urazach, ponieważ redukuje punktowe naciski na uszkodzone obszary.

Rysunek 12. Przednia i tylna strona gotowej ortezy nadgarstka
Źródło: opracowanie własne.

5. Podsumowanie

W artykule przedstawiono proces tworzenia spersonalizowanej ortezy dłoni przy użyciu technologii druku 3D, skanowania 3D oraz zaawansowanego modelowania komputerowego. Tradycyjne metody wytwarzania ortez, oparte na gipsie, są czasochłonne, kosztowne i często niewygodne dla pacjentów. W odpowiedzi na te wyzwania, projekt zakładał zastosowanie nowoczesnych technik inżynierskich, aby stworzyć ortezę idealnie dopasowaną do anatomicznych cech pacjenta, co znacząco poprawia komfort noszenia i jakość życia użytkowników. Proces rozpoczął się od uzyskania cyfrowego modelu 3D górnej kończyny pacjenta poprzez jej zeskanowanie za pomocą skanera Shining 3D Einstar. Następnie model ortezy został stworzony w programie Autodesk Inventor. Technologia druku 3D, wykorzystana w projekcie, to osadzanie stopionego materiału (FDM), która jest ekonomiczna i odpowiednia do produkcji spersonalizowanych urządzeń medycznych. Proces składał się z pięciu kluczowych kroków: skanowania 3D kończyny, opracowania modelu w oprogramowaniu CAD, drukowania fizycznej ortezy, dopasowania i wygładzenia kształtu (postprocessing).

Wybór technologii FDM podyktowany był jej ekonomicznością, łatwością obsługi oraz możliwością stosowania różnych materiałów termoplastycznych. Choć FDM oferuje umiarkowaną wytrzymałość i precyzję w porównaniu do innych technik takich jak SLA czy SLS, była ona wystarczająca do realizacji celów projektu. Proces modelowania ortezy w programie Inventor obejmował importowanie modelu ręki, tworzenie offsetów, dodawanie wycięć na palce oraz miejsc do mocowania rzepów, co zapewniało stabilne zamknięcie ortezy i komfort użytkowania. Do produkcji ortezy wybrano termoplastyczny polimer - PLA, który można formować na ciepło, aby idealnie dopasować się do anatomii pacjenta. Materiał ten jest lekki, wytrzymały - co jest istotne podczas długotrwałego noszenia. Należy także zaznaczyć, że jest biodegradowalny i stale jest ulepszana ta właściwość, tak by szybciej się rozkładał. Efektem projektu była spersonalizowana orteza, która nie tylko spełniała funkcje stabilizacyjne i ochronne, ale również była komfortowa i estetycznie dopasowana do pacjenta. W celu zwiększenia komfortu pacjenta do wnętrza ortezy przyklejono specjalną gąbkę o dużych oczkach. Ten szczególny rodzaj gąbki został wybrany ze względu na swoje właściwości umożliwiające lepsze oddychanie skóry, co przyczynia się do redukcji pocenia się ręki oraz minimalizuje ryzyko podrażnień skórnych. Istnieje jednak kilka alternatyw, które również można rozważyć w zależności od specyficznych potrzeb użytkownika

Dzięki wykorzystaniu skanowania 3D można dobrze odwzorować anatomiczne szczegóły kończyny pacjenta, co pozwala na dopasowanie ortezy do indywidualnej anatomii pacjenta. Zaawansowane modelowanie komputerowe umożliwia projektowanie ortez o optymalnej strukturze i funkcjonalności, a 3DS pozwala na szybkie i efektywne wytwarzanie końcowych produktów, dostosowanych do specyficznych potrzeb każdego pacjenta. W celu zwiększenia komfortu pacjenta do wnętrza ortezy przyklejono specjalną gąbkę o dużych oczkach. Ten szczególny rodzaj gąbki został wybrany ze względu na swoje właściwości umożliwiające lepsze oddychanie skóry, co przyczynia się do redukcji pocenia się ręki oraz minimalizuje ryzyko podrażnień skórnych. Istnieje jednak kilka alternatyw, które również można rozważyć w zależności od specyficznych potrzeb użytkownika

Literatura

1. Da Silva D., Kaduri M., Poley M., Adir O., Krinsky N., Shainsky-Roitman J., Schroeder A., *Biocompatibility, biodegradation and excretion of polylactic acid (PLA) in medical implants and theranostic systems*, Chemical Engineering Journal, 2018, 340, s. 9-14.

2. Haque A., Parsons H., Parsons N., Costa M., Redmond A., Mason J., Nwankwo H., Kearney R., *Use of cast immobilization versus removable brace in adults with an ankle fracture: two-year follow-up of a multicentre randomized controlled trial*, The Bone & Joint Journal, 2023, 105-B(4), s. 382-388.
3. Jaworska N., Podsiadło H., *Technologia druku 3D jako szansa dla środowiska naturalnego*, Acta Poligraphica, 2019, 14, s. 55-70.
4. Laska-Leśniewicz A., *Wykorzystanie metod szybkiego prototypowania (rapid prototyping) w nowoczesnej medycynie*, Zeszyty Naukowe Towarzystwa Doktorantów Uniwersytetu Jagiellońskiego. Nauki Ścisłe, 2017, 15(2), s. 39–48.
5. Oud T. A. M., Lazzari E., Gijsbers H. J. H., Gobbo M., Nollet F., Brehm M. A., *Effectiveness of 3D-printed orthoses for traumatic and chronic hand conditions: A scoping review*, PLOS One, 2021, 16(11), e0260271.
6. Schwartz D. A., & Schofield K. A., *Utilization of 3D printed orthoses for musculoskeletal conditions of the upper extremity: A systematic review*, Journal of Hand Therapy, 2023, 36(1), s. 166-178.
7. Wróbel, E., *Wielofunkcyjna orteza rehabilitacyjna dłoni z możliwością blokowania chwytu przeznaczona do wytwarzania w technologii przyrostowej* [w:] *Inżynieria biomedyczna. Metody przyrostowe w technice medycznej*, Politechnika Lubelska, Lublin 2016.

Źródła internetowe

1. FDM vs. SLA vs. SLS: Which is the best 3D Printing Technology for Your Project?: https://www.linkedin.com/pulse/fdm-vs-sla-sls-which-best-3d-printing-technology-your-jer%C3%B3nimo?trk=public_profile_article_view (dostęp: 12.06.2024).
2. PLA vs ABS vs Nylon: https://markforged.com/resources/blog/pla-abs-nylon (dostęp: 12.06.2024).

Magdalena Dul
Koło Naukowe X-Med. Politechniki Rzeszowskiej

mgr inż. Michał Wanic
Opiekun Koła Naukowego

Metody dekontaminacji i sterylizacji instrumentarium medycznego: przegląd technik

Streszczenie

W artykule omówiono kluczowe procesy związane z dekontaminacją, myciem, dezynfekcją, sterylizacją i pakowaniem narzędzi medycznych, które są niezbędne do zapewnienia bezpieczeństwa pacjentów i personelu medycznego poprzez eliminację drobnoustrojów chorobotwórczych. Każdy z tych procesów pełni istotną rolę w procesie przygotowywania narzędzi medycznych do ponownego użycia. Szczególną uwagę zwrócona na procesy dezynfekcji i sterylizacji, ze względu na ich kluczową rolę. Instrumentarium medyczne podzielono na kilka głównych kategorii, istotnych w procesach sterylizacji i dezynfekcji. Omówiono także opakowania sterylizacyjne i ich rolę zapewnieniu skuteczności procesu sterylizacji oraz bezpieczeństwa użytkowania.

Słowa kluczowe: sterylność, wyrób medyczny, dezynfekcja, dekontaminacja.

1. Wprowadzenie

Zdecydowana większość procedur medycznych wykonywanych w placówkach ochrony zdrowia związana jest z wykorzystaniem specjalistycznych narzędzi i urządzeń zwanych ogólnie instrumentarium medycznym. Zapewnienie odpowiedniego poziomu sterylności jest kluczowym aspektem pozwalającym na ograniczenie ryzyka zakażeń związanych z wykorzystywanym sprzętem medycznym. Całość instrumentarium medycznego można podzielić na elementy jednorazowe (np. igły, strzykawki, nici chirurgiczne, elektrody, itp.) oraz wielorazowe. Za sterylność elementów jednorazowych odpowiada producent dostarczając produkt w odpowiednim opakowaniu. Natomiast w przypadku elementów wielokrotnego użytku za dbanie o ich sterylność odpowiada użytkownik końcowy (szpital, przychodnia, itp.). Niniejszy artykuł opisuje sposoby postępowania niezbędne do zapewnienia bezpieczeństwa pacjentów i personelu medycznego takie jak dekontaminacja, dezynfekcja i sterylizacja.

Instrumentarium medyczne to zbiór narzędzi i urządzeń stosowanych przez personel medyczny w celach między innymi diagnostycznych, chirurgicznych, terapeutycznych. Instrumentarium to podzielić można na kilka głównych kategorii: narzędzia chirurgiczne, diagnostyczne, laboratoryjne, endoskopowe, do podtrzymywania funkcji życiowych.

W celu uskutecznienia sterylizacji i przechowywania narzędzi medycznych, w sposób zachowujący ich sterylność, konieczne jest stosowanie odpowiednich opakowań. Opakowania sterylizacyjne można podzielić na jednorazowe i wielorazowe. Każde z nich musi spełniać określone kryteria, aby zapewnić skuteczność procesu sterylizacji oraz bezpieczeństwo użytkowania.

Celem niniejszego artykułu jest kompleksowe przedstawienie wszystkich aspektów związanych z procesem sterylizacji narzędzi medycznych. Artykuł ten ma na celu edukację i zwiększenie świadomości pracowników służby zdrowia, specjalistów zajmujących się sterylizacją oraz innych zainteresowanych osób na temat kluczowych etapów i technik związanych z dekontaminacją, myciem, dezynfekcją, sterylizacją oraz pakowaniem narzędzi medycznych. Poprzez szczegółowy opis każdego z tych procesów, artykuł dąży do podkreślenia znaczenia utrzymania wysokich standardów higieny i bezpieczeństwa w placówkach medycznych, co ma bezpośredni wpływ na zdrowie pacjentów oraz efektywność zabiegów medycznych

2. Reprocesowanie, mycie, dezynfekcja i dekontaminacja

Reprocesowanie, czyli ponowna obróbka instrumentarium medycznego obejmująca zakres czynności gwarantujących bezpieczeństwo ponownego użycia narzędzi. Jest to wieloetapowy proces, obejmujący przede wszystkim: czyszczenie, dezynfekcję, sterylizację, a także odpowiednie opakowanie narzędzi. Warto pamiętać, że zgodnie z Opinią Konsultanta Krajowego w dz. Pielęgniarstwa chirurgicznego i operacyjnego na temat reprocesowania wyrobów medycznych do jednorazowego używania[1], nie jest dozwolone poddanie temu procesowi jednorazowego wyrobu. Nie ma możliwości przeprowadzenia takiej procedury z zachowaniem zasad bezpieczeństwa. Produkty jednokrotnego użycia nie gwarantują utrzymania swoich właściwości po przeprowadzeniu procesu dekontaminacji, konserwacji oraz sterylizacji.

Mycie narzędzi medycznych to kluczowy etap w procesie ich przygotowania do ponownego wykorzystania. Obejmuje on usuwanie zanieczyszczeń, resztek tkanek i innych substancji organicznych, które mogą pozostać na narzędziach po użyciu. Proces mycia ma na celu eliminację mikroorganizmów oraz redukcję ryzyka zakażenia pacjenta. Narzędzia chirurgiczne czyści się z wykorzystaniem między innymi myjek ultradźwiękowych, a także specjalnych

[1]M. Szewczyk. , prof. UMK, Opinia Konsultanta Krajowego w dz. Pielęgniarstwa chirurgicznego i operacyjnego na temat reprocesowania wyrobów medycznych do jednorazowego użycia, Okręgowa Izba Pielęgniarek i Położnych w Elblągu, 2013.

roztworów myjących. Myjka ultradźwiękowa wykorzystuje generowane fale ultradźwiękowe, które przemieszczając się w ośrodku przewodzącym (roztworze myjącym). Powoduje to szybkie zmiany ciśnienia w cieczy, które powodują powstawanie mikropęcherzyków powietrza, które implodując oddzielają cząsteczki zanieczyszczeń od przedmiotu. Zjawisko powstawania pęcherzyków powietrza w cieczy pod wpływem zmian cieśnienia nosi nazwę kawitacji.

Dezynfekcja i dekontaminacja są kluczowymi procesami w zapewnianiu bezpieczeństwa pacjentów i personelu medycznego poprzez eliminację drobnoustrojów chorobotwórczych.

Dezynfekcja redukuje liczbę patogennych mikroorganizmów do bezpiecznych poziomów, prawdopodobieństwo wystąpienia drobnoustrojów wynosi 1 : 1 000 000 (SAL 10^{-6}). Występowanie mikroorganizmów poniżej tego poziomu pozwala uznać wyrób medyczny za bezpieczny dla zdrowia pacjentów i personelu medycznego, a ryzyko zakażenia jest minimalne lub praktycznie nie istnieje. W przypadku tej metody nie są niszczone wszystkie przetrwalniki, czyli formy spoczynkowe pozwalające organizmom na przetrwanie w niekorzystnych warunkach. Jednak ich ilość jest zredukowana to bezpiecznych poziomów.[2]

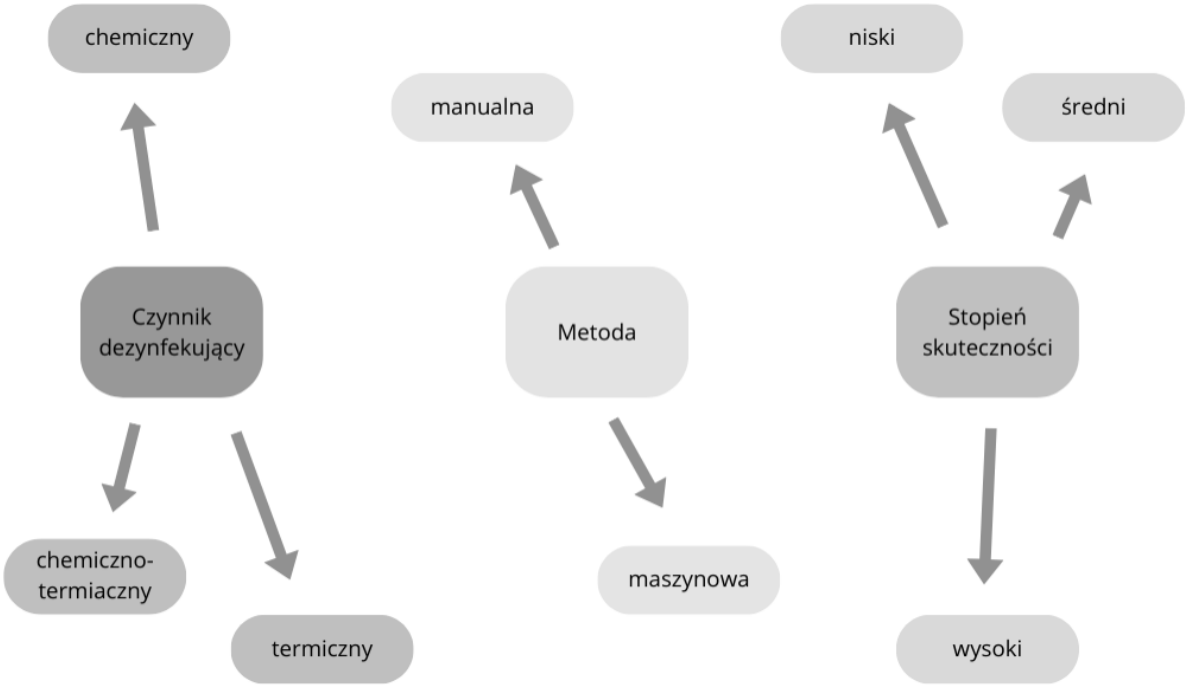
Dezynfekcja pozwala na zapobieganie zakażeniom krzyżowym i utrzymanie odpowiednich standardów higieny w placówkach medycznych. Proces dezynfekcji obejmuje stosowanie środków chemicznych lub fizycznych, które są skuteczne w eliminacji lub redukcji populacji drobnoustrojów do poziomów bezpiecznych dla pacjenta. Przed przystąpieniem do dezynfekcji narzędzi są one najpierw starannie myte oraz przechodzą kontrolę pod kątem ich uszkodzeń. Dezynfekcję można podzielić na:

- Dezynfekcja manualna wykonywana jest ręcznie przy użyciu środków chemicznych, np. przecieranie powierzchni płynami dezynfekującymi.
- Dezynfekcja maszynowa wykorzystuje urządzenia mechaniczne, np. myjnie-dezynfektory, myjki ultradźwiękowe.
- Dezynfekcja termiczna to dezynfekcja za pomocą wysokiej temperatury (ok. 100°C), np. para wodna.
- Dezynfekcja chemiczna to dezynfekcja wykorzystująca chemiczne środki dezynfekujące, np. alkohol.

[2] E.Malatyńska-Jasak, B.Wanot, A.Biskupek-Wanot, *Procedury dezynfekcji i sterylizacji w ochronie zdrowia*, Wydawnictwo Naukowe Uniwersytetu Humanistyczno-Przyrodniczego, Częstochowa, 2022.

- Dezynfekcja chemiczno-termiczna łączy obydwie przedstawione powyżej metody. Jest połączeniem ciepła ok. 60°C z zastosowaniem środków chemicznych o niższych stężeniach.
- Dezynfekcja niskiego stopnia eliminuje większość bakterii, wybrane wirusy i grzyby. Natomiast nie eliminuje wszystkich przetrwalników. Wykorzystuje się ją głównie do powierzchni i sprzętu niemedycznego.
- Dezynfekcja średniego stopnia zabija większość bakterii, wybrane prątki, wirusy i grzyby, ale również nie eliminuje wszystkich przetrwalników. Stosuje się ją do sprzętu medycznego mającego kontakt z błonami śluzowymi.
- Dezynfekcja wysokiego stopnia pozwala na eliminacje wszystkich mikroorganizmów, poza grupą dużej liczby przetrwalników. Używana jest do dezynfekcji narzędzi, wchodzących w bezpośredni kontakt z krwią lub sterylnymi częściami ciała.

Powyższy podział rodzajów dezynfekcji został przedstawiany na Rys. 1.



Rysunek1. Podział rodzajów dezynfekcji ze względu na czynnik dezynfekujący, metodę i stopień skuteczności. Źródło: opracowanie własne

Dekontaminacja obejmuje usunięcie zanieczyszczeń oraz redukcję i eliminację drobnoustrojów. Zawiera zarówno fizyczne usuwanie zanieczyszczeń, dezynfekcję oraz sterylizację (szczegółowo opisana w kolejnym rozdziale). Proces dekontaminacji polega na

usuwaniu i neutralizacji materiałów toksycznych (odpady radioaktywne, chemikalia) lub na zabijaniu i usuwaniu drobnoustrojów (sterylizacja narzędzi chirurgicznych)[3]. Etapy dekontaminacji powierzchni różnią się w zależności od stopnia ich zanieczyszczenia. Dla powierzchni o niewielkim zanieczyszczeniu materią organiczną, wystarczy mycie oraz suszenie. Dla powierzchni bardziej zanieczyszczonych, po etapie suszenia następuje jeszcze dezynfekcja. Proces dekontaminacji narzędzi i sprzętu medycznego obejmuje kilka etapów, które muszą być dokładnie przestrzegane, w celu uniknięcia powtórnej kontaminacji sprzętu. Prawidłowa dekontaminacja narzędzi medycznych wymaga odpowiedniego doboru metod, ścisłego przestrzegania procedur oraz ciągłego szkolenia personelu. Wśród kluczowych etapów można wymieć:

- mycie,
- dezynfekcje,
- suszenie,
- kontrolowanie,
- pakowanie,
- znakowania,
- sterylizacje,
- przechowywanie.

3. Sterylizacja medyczna – proces i metody

Sterylizacja medyczna to proces przemyślanego działania mający na celu eliminację wszelkich drobnoustrojów i przetrwalników bakteryjnych oraz stanowi kolejny etap reprocesowania instrumentów medycznych. Sterylność medyczna może być również definiowana jako obniżenie ilości drobnoustrojów w produkcie do granicy zapewniającej całkowite zabezpieczenie przed rozwojem. Metody sterylizacji medycznej muszą spełniać następujące kryteria:

- bezpieczeństwo i nieszkodliwe działanie na lekarza i pacjenta,
- zapewnienie niezmienności własności użytkowych materiału (brak zmian cech funkcjonalnych, wpływających na użytkowanie wyrobu medycznego).

[3] https://www.medonet.pl/zdrowie,dekontaminacja--usuniecie-drobnoustrojow--oczyszczenie-ciala-lub-przedmiotow,artykul,1731226.html

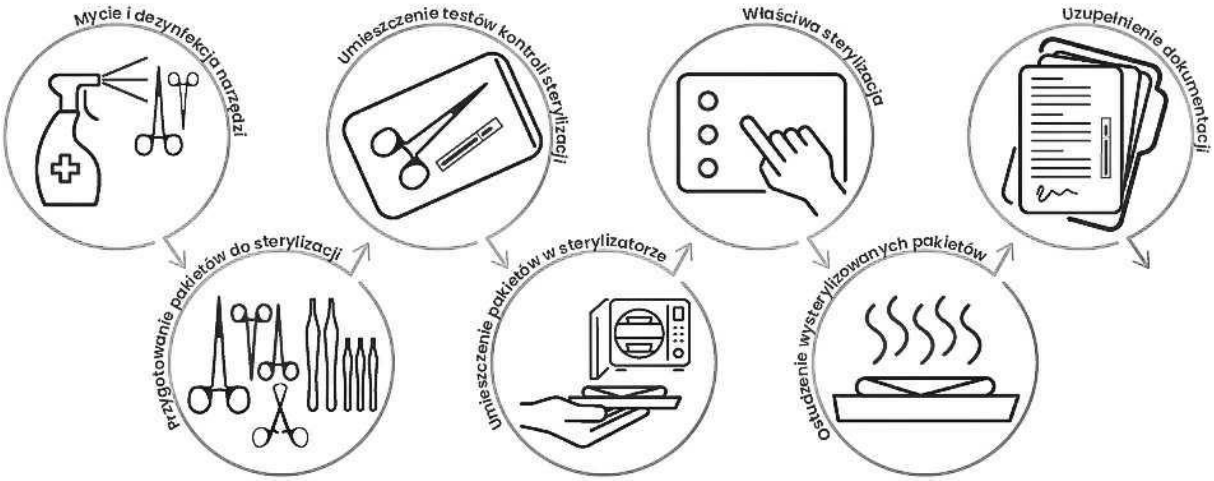
Coraz częściej wymaga się także ekologicznej czystości procesu przemysłowej sterylizacji medycznej, a także przykłada się coraz większą wagę do zapewnienia bezpieczeństwa osobom pracującym przy tym procesie.

W celu zapewnienia sterylności wyrobów medycznych należy postępować zgodnie z zasadami przeprowadzenia tego procesu określonymi przez Państwowy Zakład Higieny. Niezbędne jest właściwe oddziaływanie z wyrobem przed poddaniem go sterylizacji, jak i po tym procesie. Obowiązuje kilka podstawowych zasad podczas przygotowywania wyrobu medycznego[4]:

- Należy stosować się do instrukcji danego wyrobu medycznego.
- Przy procesie dekontaminacji zaleca się wykorzystanie wody oczyszczonej w wymaganym stopniu, właściwym dla danej technologii. Ma to na celu zapobieganie ewentualnym uszkodzeniom narzędzi.
- Wszystkie etapy dekontaminacji oraz proces sterylizacji powinny być wykonywane w jednym, tym samym miejscu.
- Gdy niezbędny jest transport narzędzi, czy innych instrumentariów medycznych, nie jest konieczne mycie z dezynfekcją wyrobów w miejscu ich użycia, pod warunkiem, że czas ich transportu nie trwał dłużej niż 3 godziny (od momentu użycia do rozpoczęcia dekontaminacji). Skażone wyroby medyczne transportowane są w szczelnych opakowaniach odpornych na uszkodzenia.
- Należy stosować określone preparaty chemiczne o działaniu bakteriostatycznym w celu zapobiegnięcia zasychaniu krwi.
- Zalecana jest kontrola wewnętrzna przeprowadzanych procesów oraz dokumentacja ich wyników.

Cały proces można podzielić na szereg etapów: mycie i dezynfekcja, przygotowanie pakietów, umieszczenie testów kontrolnych, proces sterylizacji, kontrola testów, przygotowanie dokumentacji (Rys. 2). Główne procesy zostały omówione poniżej.

[4]Praca zbiorowa pod redakcją P. Grzesiowski, *Ogólne wytyczne dla wszystkich podmiotów wykonujących procesy dekontaminacji, w tym wyrobów medycznych i innych przedmiotów wielokrotnego użytku wykorzystujących przy udzielaniu świadczeń zdrowotnych oraz innych czynności, podczas których może dojść do przeniesienia choroby zakaźnej lub zakażenia*, , Warszawa, 2017.



Rysunek2. Etapy sterylizacji narzędzi chirurgicznych,.
Źródło: URL: https://sterim.eu/blog/proces-sterylizacji-krok-po-kroku

Pierwszym etapem w procesie sterylizacji jest prawidłowe przygotowanie wyrobu medycznego do pracy nad nim. Polega ono na uprzednim umyciu i dezynfekcji sprzętu medycznego w celu ułatwienia dotarcia czynnika sterylizującego do każdego jego zakamarka.

Przygotowanie pakietów polega na umieszczeniu umytych i zdezynfekowanych narzędzi chirurgicznych w odpowiednim opakowaniu do sterylizacji. Wyroby medyczne poddawane sterylizacji obowiązkowo przechodzą kontrolę pod kątem czystości, jak i zarówno tego, czy nie zostały uszkodzone. Dokładnie sprawdzane są miejsca krytyczne, np. szczeliny. Instrumenty medyczne, które są poddawane sterylizacji i wymagają czasowego przechowywania lub transportu, muszą być odpowiednio opakowane. Zadaniem takiego opakowania jest zachowanie sterylności zawartości od momentu zakończenia procesu sterylizacji do chwili użycia. Systemy opakowań sterylizacyjnych są znormalizowane, aby spełniać wszystkie kryteria utrzymania sterylności wyrobu medycznego (PN EN 868). Producent opakowań musi wykazać przydatność opakowania do konkretnego zastosowania, dokumentując możliwość wykorzystania danego opakowania do określonej metody sterylizacji.

Kolejnym etapem jest umieszczenie pakietów w sterylizatorze. Specjalnym aparatem służącym do wyjaławiania, czyli eliminacji drobnoustrojów między innymi z narzędzi chirurgicznych. Sposób ułożenia instrumentarium medycznego jest zależny jest w głównej mierze od: kształtu, wielkości i parametrów geometrycznych narzędzia. To właśnie w sterylizatorach odbywa się właściwy proces wyjaławiania, który przebiega całkowicie automatycznie, za pomocą wybranego programu. Sterylizację przeprowadza się w warunkach nisko- lub wysokotemperaturowych, za pomocą różnych czynników aktywnych (Rys. 3).

Sterylizacja parą wodną w nadciśnieniu wykorzystuje nasyconą parę wodną, która skutecznie niszczy drobnoustroje przez koagulację białek. Metoda ta jest bezpieczna dla środowiska, jest jednak nieodpowiednia dla materiałów wrażliwych na temperaturę i wilgoć. Może być realizowana w autoklawach grawitacyjnych dla prostych przedmiotów lub próżniowych w przypadku złożonych struktur. Skuteczność jej zależy od usunięcia powietrza z komory i jakości pary wodnej.

Sterylizacja suchym gorącym powietrzem stosowana dla materiałów nieprzepuszczalnych dla wilgoci. Przeprowadzana w temperaturze 160°C przez 120 minut lub 180°C przez 30 minut. Ze względu na słabą penetrację suchego powietrza i długi czas trwania procesu nie jest zalecana do sterylizacji narzędzi medycznych.

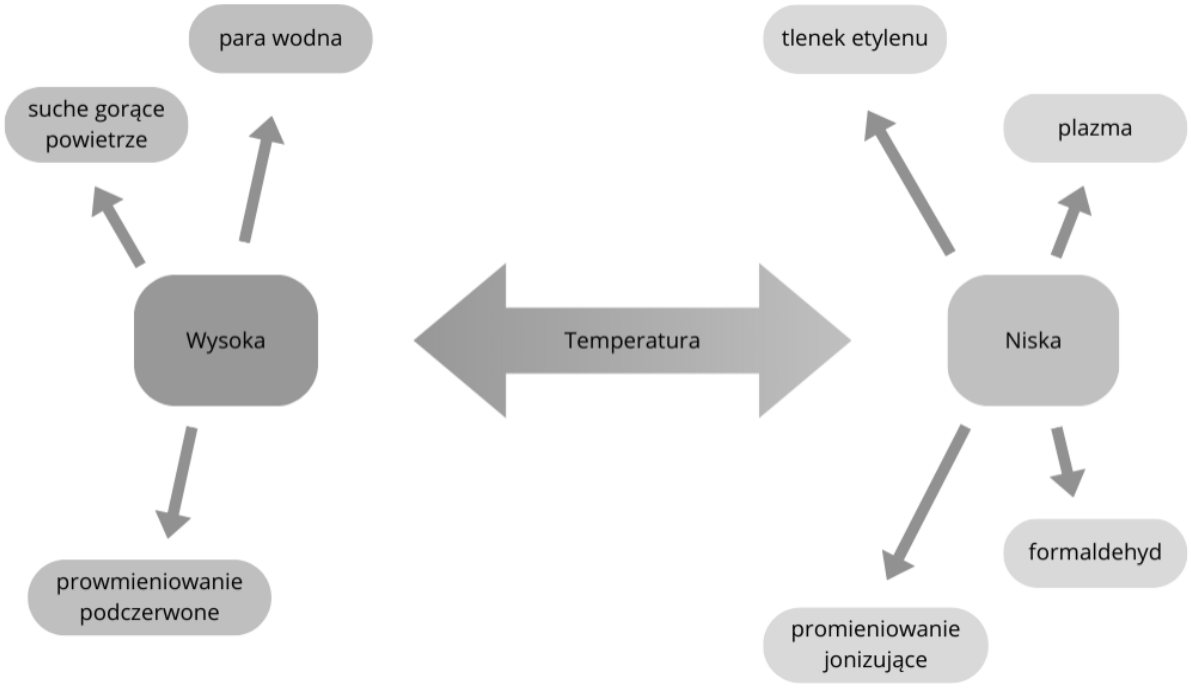
Sterylizacja promieniowaniem podczerwonym stosowana jest obecnie w przemyśle spożywczym i farmaceutycznym. Na chwilę obecną nie jest dopuszczona do wykorzystania w placówka medycznych. Metoda ta charakteryzuje się skutecznością w niszczeniu przetrwalników bakteryjnych, krótkim czasem cyklu i brakiem toksyczności, potwierdzonym laboratoryjnie.

Sterylizacja tlenkiem etylenu niszczy drobnoustroje przez alkilację białek, DNA i RNA. Proces odbywa się w temperaturze 30-65°C i wilgotności 40-80%, trwa 2-5 godzin. Jest toksyczny i karcinogenny, a materiał wymaga długiej degazacji.

Sterylizacja plazmą wykorzystuje nadtlenek wodoru w warunkach próżni w temperaturze pomiędzy 38-50°C przez 30-75 minut. Produktem końcowym są tlen i woda. Nie można jej stosować do sterylizacji bielizny, celulozy, proszków i płynów.

Sterylizacja formaldehydem stosuje parę wodną i formaldehyd w temperaturze 48-70°C przez 2-4 godziny. Jej zastosowanie ograniczone jest poprzez toksyczność oraz ograniczoną skuteczność penetracji.

Sterylizacja promieniowaniem jonizującym stosowana jest do przemysłowej sterylizacji sprzętu medycznego. Wykorzystuje się tu głównie promieniotwórczy izotop kobaltu: kobalt-60. Zaletami tej metody są: krótki czas sterylizacji, temperatura zbliżona do pokojowej i brak toksycznych pozostałości.



Rysunek2. Rodzaje sterylizacji ze względu na temperaturę przeprowadzania procesu i użytych czynników.

Źródło: opracowanie własne na podstawie: Fleischer M., Dezynfekcja, Sterylizacja, Antyseptyka, Katedra i Zakład Mikrobiologii, Uniwersytet Medyczny im. Piastów Śląskich we Wrocławiu, 2017.

4. Opakowanie i kompletowanie narzędzi medycznych.

Norma PN-EN ISO 11607:2006 dotycząca opakowań przeznaczonych do finalnie sterylizowanych produktów wprowadza pojęcie systemu bariery sterylnej, określając wymagania dotyczące systemów bariery sterylnej, materiałów i systemów opakowaniowych oraz także wymagania dotyczące procesów zestawiania, kształtowania i uszczelniania. Norma PN-EN 868 2-10 uwzględnia materiały i systemy opakowaniowe dla wyrobów medycznych przeznaczonych do sterylizacji, zapewniając wymagania dotyczące przenikania czynnika sterylizującego, odporności na uszkodzenia, szczelności zamknięcia oraz barierowości dla drobnoustrojów i niepożądanych substancji. Opakowanie musi umożliwiać skuteczne przenikanie czynnika sterylizującego (np. pary wodnej, gazów) do jego wnętrza. Konieczna jest również odporność na uszkodzenia podczas przeprowadzenia procesu sterylizacji. Opakowanie powinno zapewniać szczelne i trwałe zamknięcie zawartości, jak i umożliwiać bezpieczne wyjęcie zawartości do ponownego użycia. Opakowanie sterylizacyjne stanowi barierę dla drobnoustrojów oraz niepożądanych substancji, takich jak kleje, tusze z nadruku czy testy chemiczne.

Proces kompletowania i pakietowania narzędzi medycznych do sterylizacji jest kluczowym etapem w zapewnieniu skuteczności sterylizacji i utrzymaniu jałowości

instrumentów medycznych. Obejmuje on kilka istotnych kroków, które muszą być dokładnie przestrzegane. W zestawy kompletuje się umyte, sprawdzone i zdezynfekowane narzędzia oraz załącza test chemiczny, umożliwiający monitorowanie skuteczności procesu sterylizacji. Wybór odpowiedniego materiału do pakowania jest ważnym czynnikiem dla skuteczności procesu sterylizacji. Opakowania muszą być kompatybilne z wybraną metodą sterylizacji oraz muszą zapewniać długotrwałą ochronę mikrobiologiczną zapakowanych narzędzi. Dobór materiału opakowaniowego zależy od specyfiki sterylizowanych instrumentów i wybranej metody sterylizacji, a także wymagań dotyczących przechowywania i transportu wysterylizowanych narzędzi.

5. Mechanizmy inaktywacji drobnoustrojów

Zasada działania sterylizacji opiera się na inaktywacji drobnoustrojów z wykorzystaniem jednego z mechanizmów: oksydacji, alkilacji lub denaturację.

Oksydacja to proces chemiczny, który polega na przeniesieniu elektronów od jednego związku chemicznego do drugiego, co prowadzi do blokowania działania enzymów i koenzymów. W kontekście inaktywacji drobnoustrojów, oksydacja powoduje rozległe uszkodzenia i ostateczną inaktywację wirusów oraz innych patogenów. Obecnie stosowane substancje utleniające o działaniu sterylizacyjnym to kwas nadoctowy oraz nadtlenek wodoru. Kwas nadoctowy to silny utleniacz, który uszkadza błony komórkowe, białka oraz materiał genetyczny drobnoustrojów poprzez wytwarzanie reaktywnych form tlenu. Skutkuje to ich inaktywacją, czyli obniżeniem lub całkowitym zanikiem aktywności substancji biologicznie czynnej. Działanie nadtlenku wodoru polega na wytwarzaniu wolnych rodników tlenowych, które uszkadzają błony komórkowe, białka oraz DNA drobnoustrojów. Jest on skuteczny przeciwko szerokiemu spektrum mikroorganizmów, w tym bakterii, wirusów, grzybów i przetrwalników. Jest używany zarówno w postaci płynnej, jak i gazowej do dezynfekcji narzędzi medycznych, powierzchni oraz w procesach sterylizacji plazmowej.

Alkilacja to proces chemiczny polegający na przyłączaniu grup alkilowych do różnych cząsteczek. Prowadzi do modyfikacji strukturalnych i funkcjonalnych kluczowych komponentów komórkowych. Przyłączenie tych grup do zasad azotowych w DNA prowadzi do błędów w replikacji i transkrypcji. Skutkuje to mutacjami oraz śmiercią komórek. Proces alkilacji zamienia także strukturę białek, co może doprowadzić do utraty ich funkcji biologicznych, przez co zaburzone zostają procesy metaboliczne komórki. W sterylizacji medycznej wykorzystuje się związki alkilujące takie jak: formaldehyd i tlenek etylenu.

Denaturacja to zmiany w strukturze białka natywnego. Prowadzą one do utraty aktywności biologicznej lub innej charakterystycznej cechy przy zachowaniu sekwencji aminokwasów. Podczas tych zmian niszczone są wiązania wodorowe, a także w określonych warunkach zerwaniu ulegają także mostki dwusiarczkowe.

6. Podsumowanie

Artykuł przedstawia procesy zapewnianiające sterylność instrumentów medycznych, które są kluczowe dla ograniczenia ryzyka zakażeń w placówkach ochrony zdrowia. Instrumentarium medyczne, podzielone na jednorazowe i wielorazowe, wymaga odpowiedniego postępowania, aby zapewnić jego sterylność. Jednorazowe elementy są sterylizowane przez producenta, natomiast za sterylność wielorazowych odpowiada użytkownik końcowy. Procesy te obejmują mycie, dezynfekcję, dekontaminację i sterylizację. Dezynfekcja redukuje liczbę patogennych mikroorganizmów do bezpiecznych poziomów. Procesy dezynfekcji charakteryzują się różnym stopniem skuteczności, który zależy od wykorzystanej metody i środka dezynfekującego. Dobór odpowiedniego postępowania zależy od rodzaju urządzania oraz sposobów kontaktów z pacjentem. Sterylizacja całkowicie niszczy wszystkie formy życia biologicznego. Procesy sterylizacji można podzielić na dwie główne grupy nisko- i wysokotemperaturowe, dla każdej z tej grup występuje szereg czynników aktywnych. Dobór odpowiedniego postepowania zależy przede wszystkim od materiałów z jakich wykonany jest sterylizowany przedmiot oraz od jego geometrii. Aby proces sterylizacji przebiegł prawidłowo, ważne jest stosowanie odpowiednich opakowań sterylizacyjnych, aby utrzymać sterylność narzędzi. Artykuł zwiększa świadomość na temat technik i etapów tych procesów, podkreślając znaczenie wysokich standardów higieny i bezpieczeństwa, co ma bezpośredni wpływ na zdrowie pacjentów oraz efektywność zabiegów medycznych.

Literatura

2. Bielacka A., *Sterylizacja narzędzi i materiałów medycznych. Metody i kontrola*, PZWL, Warszawa, 1994.
3. Break M.R., *Metody i kontrola sterylizacji*, PZWL, Warszawa, 1974.
4. Buchrieser V., Miorini T., *Podstawy Mycia, Dezynfekcji i Sterylizacji, Skrypt Podstawowy*, WFHSS, 2009.
5. Jaguś A., *Sterylizacja niskotemperaturowa tlenkiem etylenu w pytaniach i odpowiedziach*, Pielęgniarka EPIDEMIOLOGICZNA, Kwartalnik Polskiego Stowarzyszenia Pielęgniarek Epidemiologicznych, Marzec, 2013.

6. Malatyńska-Jasak E., Wanot B., Biskupek-Wanot A., *Procedury dezynfekcji i sterylizacji w ochronie zdrowia*, Wydawnictwo Naukowe Uniwersytetu Humanistyczno-Przyrodniczego, Częstochowa, 2022.
7. Praca zbiorowa pod redakcją P. Grzesiowski, *Ogólne wytyczne dla wszystkich podmiotów wykonujących procesy dekontaminacji, w tym wyrobów medycznych i innych przedmiotów wielokrotnego użytku wykorzystujących przy udzielaniu świadczeń zdrowotnych oraz innych czynności, podczas których może dojść do przeniesienia choroby zakaźnej lub zakażenia*, Warszawa, 2017.
8. Szewczyk. M., *Opinia Konsultanta Krajowego w dz. Pielęgniarstwa chirurgicznego i operacyjnego na temat reprocesowania wyrobów medycznych do jednorazowego użycia*, Okręgowa Izba Pielęgniarek i Położnych w Elblągu, 2013.
9. Zespół Roboczy ds. Przygotowania Instrumentarium Medycznego: Prawidłowy sposób przygotowania-Przygotowanie instrumentarium medycznego zachowujące jego wartość"; Wydanie 11, 2017.

Akty normatywne

1. PN-EN 868-5:2019-01, *Opakowania dla finalnie sterylizowanych wyrobów medycznych - Część 5: Torebki z zamknięciem samoprzylepnym oraz rękawy z materiałów porowatych i folii z tworzywa sztucznego - Wymagania i metody badań*, Polski Komitet Normalizacyjny, Warszawa, 2020.
2. PN-EN ISO 11607:2006, *Opakowania przeznaczone do finalnie sterylizowanych wyrobów medycznych -- Część 1: Wymagania dotyczące materiałów, systemów barier sterylnych i systemów opakowaniowych*, Polski Komitet Normalizacyjny, Warszawa, 2006.

Źródła internetowe

1. https://www.medonet.pl/zdrowie,dekontaminacja--usuniecie-drobnoustrojow--oczyszczenie-ciala-lub-przedmiotow,artykul,1731226.html(dostęp:13.06.2024).
2. https://sani.pl/blog/dezynfekcja-i-sterylizacja-jakie-sa-najwazniejsze-roznice(dostęp:13.06.2024).
3. https://sterim.eu/blog/proces-sterylizacji-krok-po-kroku(dostęp:13.06.2024).
4. https://dezpro.pl/sanityzacja-sterylizacja-dekontaminacja-i-dezynfekcja-czym-sie-roznia/?fbclid=IwZXh0bgNhZW0CMTAAAR35n_xYL09xAFr8aDagJ_2HJPmnsnaIi8yK7JRCEbaci6K3oOwIHru641Y_aem_ZmFrZWR1bW15MTZieXRlcw(dostęp:13.06.2024).

KOŁO

NAUKOWE

ELEKTRONIKI

I TECHNOLOGII

INFORMACYJNYCH

Piotr Dubaj, Jakub Bocek, Patryk Krupa, Sławomir Pareniak, Katarzyna Maternia
Koło naukowe Elektroniki i Technologii Informacyjnych

dr inż. Bartosz Pawłowicz
Opiekun Koła Naukowego

Autonomiczne środki transportu- szanse i zagrożenia dla społeczeństwa

Streszczenie

Autonomiczne środki transportu otwierają drzwi do rewolucji w mobilności, obiecując poprawę bezpieczeństwa, efektywności czasowej i dostępności transportu. Jednakże, wraz z ich rozwojem pojawiają się również obawy dotyczące prywatności danych, bezrobocia, nierówności społecznych oraz konieczności dostosowania prawa i regulacji. Wdrożenie tych innowacji wymaga zrównoważonego podejścia do tematyki zagadnień związanych z autonomizacją pojazdów, uwzględniającego zarówno korzyści, jak i potencjalne zagrożenia dla społeczeństwa.

Słowa kluczowe: samochód autonomiczny, poziomy autonomiczności, bezpieczeństwo drogowe, infrastruktura drogowa.

1. Wprowadzenie do tematyki.

Samochód autonomiczny, definiowany jako pojazd zdolny do poruszania się bez konieczności ingerencji ze strony kierowcy, stanowi rewolucję w dziedzinie transportu. W tradycyjnym rozumieniu, samochód to narzędzie, które wymaga aktywnego sterowania przez człowieka. Jednak, dzięki zaawansowanej technologii komputerowej, pojazdy autonomiczne mogą samodzielnie kontrolować swoją jazdę, regulować prędkość i omijać przeszkody.

Aktualnie samochody autonomiczne są już produkowane i testowane na drogach. Ich funkcje autonomicznej jazdy stopniowo stają się coraz bardziej zaawansowane, z wykorzystaniem zaawansowanych systemów radarowych, kamer, laserów oraz technologii GPS. Te pojazdy, wyposażone w różne stopnie autonomiczności, od prostych systemów asystujących po pełną automatyzację, mogą być już spotykane w wybranych miejscach komercyjnych na całym świecie.

Jednak, choć obiecująca, technologia samochodów autonomicznych stawia także przed nami wiele wyzwań. W niniejszym artykule przyjrzymy się bliżej temu, jak jest aktualny rozwój pojazdów autonomicznych i jak działają pojazdy autonomiczne, jakie są ich różne poziomy autonomiczności oraz jakie kwestie, takie jak bezpieczeństwo i ubezpieczenia, wiążą się z ich wprowadzeniem na drogi, a także przeanalizujemy szanse i zagrożenia dla społeczeństwa.

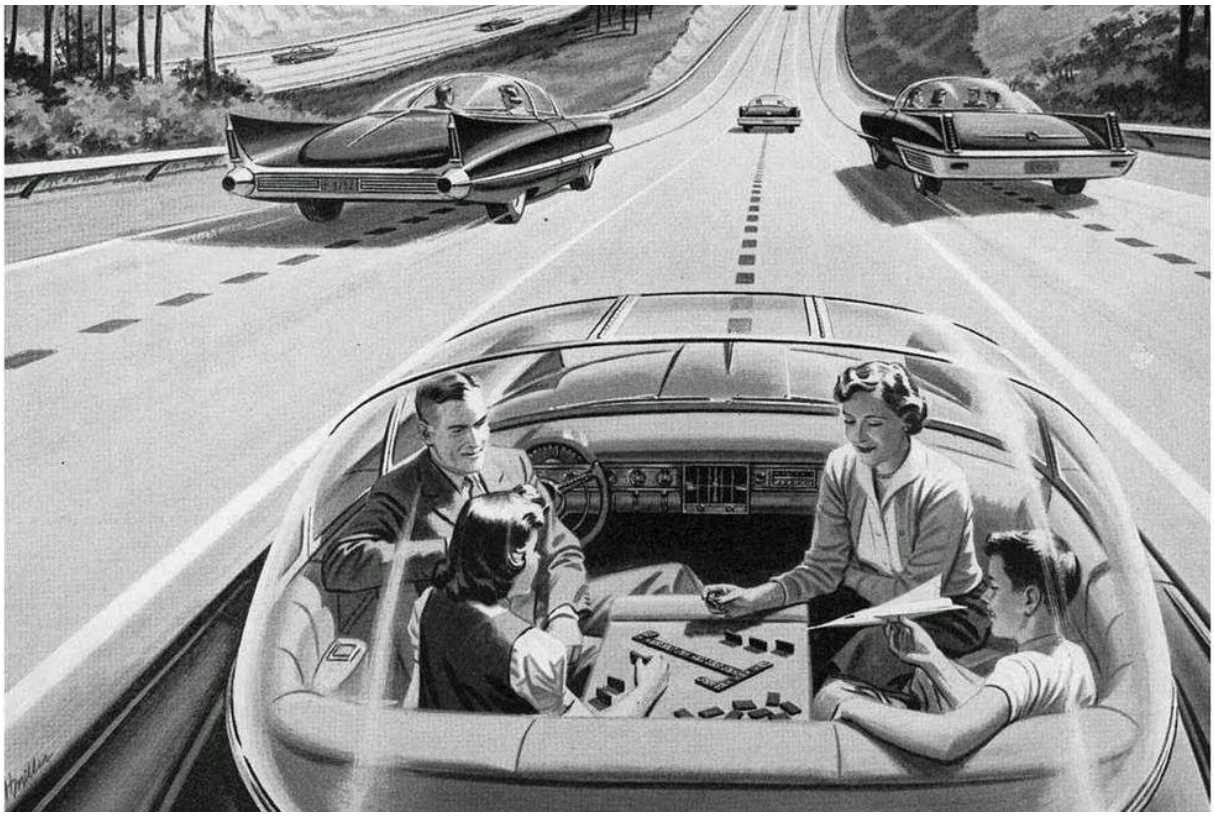
2. Historia samochodów autonomicznych.

O idei pojazdów autonomicznych pisał już Leonardo da Vinci. Pierwszy prawdziwy wóz bez kierowcy pojawił się w latach 20. w Stanach Zjednoczonych. Był to samochód, w którym kierowcę zastępował układ zdalnego sterowania. Kontrolowało go auto jadące z tyłu. W tamtych czasach było to spore wydarzenie. Dziennikarze gazety The Milwaukee Sentinel pisali o krążącym po ulicach samochodzie widmo.

W 1939 roku w Nowym Jorku firma General Motors zorganizowała ekspozycję Futurama. Pokazywała ona, jak zmieni się świat do 1960 roku. Koncepcja zakładała separację ruchu samochodowego od otoczenia. Przewidywała również bezkolizyjny ruch i samochody pozbawione kierowców, poruszające się po specjalnie przystosowanych autostradach.

W 1953 roku firma RCA Labs przeprowadzała próby z miniaturowym modelem, który poruszał się samodzielnie, dzięki umieszczonym w drodze, magnetycznym paskom. W 1958 roku ta sama grupa ludzi przeprowadziła próbę z samochodem w tradycyjnych wymiarach, na specjalnie przygotowanym odcinku publicznej drogi w Nebrasce. Współpraca z General Motors pozwoliła na bardziej zaawansowane rozwiązanie – drogę wykrywającą znajdujące się na niej pojazdy. Dzięki tym informacjom, autonomiczny samochód mógł przyspieszać, hamować oraz skręcać. W latach 60. i 70. podobne eksperymenty prowadzono także w Wielkiej Brytanii z wykorzystaniem Citroena DS. Sterowanie umożliwiały umieszczone w drodze kable sygnałowe. Te nowatorskie pomysły ingerowały w infrastrukturę – samochody mogły samodzielnie poruszać się jedynie w specjalnie przygotowanych miejscach. To wymagałoby zbudowania wszystkich dróg od nowa.

Na początku lat 80. Ernst Dickmanns, ekspert od komputerowego przetwarzania obrazów z Monachium, rozpoczął próby z całkowicie autonomicznym pojazdem, bazującym na 5-tonowym vanie Mercedesa. Konstruktor wyposażył auto w kilka kamer i komputer oraz opracował nowatorski algorytm, symulujący tzw. ruchy sakkadowe ludzkiego oka. Pozwalał on komputerowi na “widzenie” trójwymiarowe. Pojazd o nazwie VaMoRs jeździł jedynie po wyłączonych z użytku fragmentach dróg, żeby nie naruszać obowiązujących przepisów.



Rys1. Retrofuturystyczna wizja samochodu bez kierowcy Źródło: https://v.wpimg.pl/N2VkM2YuYVM3CTsBdg5sRnRRb1swV2IQI0l3EHYWYQNhWX1KdhhgVTcPN1ctGiNTOAIsHCkZYUAwWmBXOBB2CjUKeQJrR3kKNQ8uVDwTLVZvDXpXPUYoHDwbKhAk

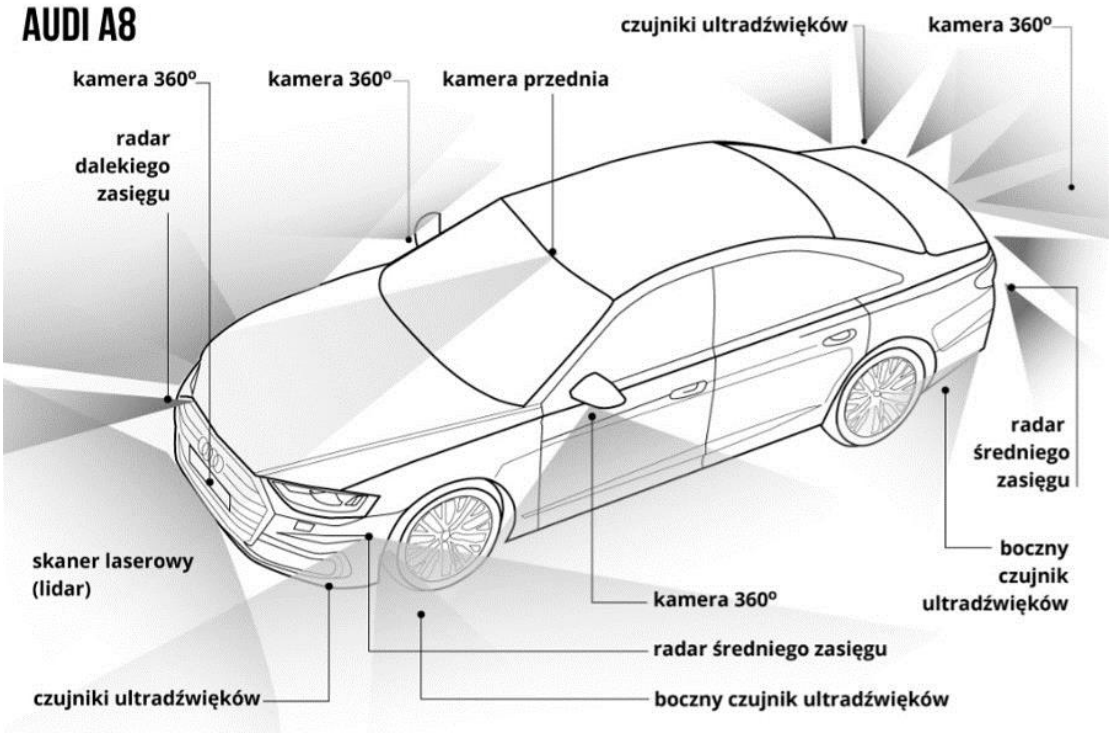
Niedługo później rozpoczęto program PROMETHEUS (PROgraMme for a European Traffic of Highest Efficiency and Unprecedented Safety). Jego efektem był między innymi VaMP, czyli przerobiony Mercedes 500 SEL. Mógł on samodzielnie śledzić innych uczestników ruchu, zmieniać pasy i wyprzedzać wolniejsze samochody. Komputer sterujący potrafił nawet kontrolować kilka pojazdów jednocześnie. W 1992 roku przeprowadzono próby autonomicznych aut w normalnym ruchu ulicznym. 3 lata później autonomiczny samochód przejechał około 1700 kilometrów – z Monachium do duńskiego Odense. Co pewien czas musiał jednak interweniować człowiek. Najdłuższy odcinek bez reakcji kierowcy wynosił 158 kilometrów.

Prace nad autonomicznymi samochodami prowadziła także DARPA. Pojazd Autonomous Land Vehicle (ALV) mógł przemieszczać się w trudnym, nieznanym terenie z licznymi przeszkodami. W tych samych latach we Włoszech prowadzono program ARGO. Przebudowana Lancia Thema bazowała na śledzeniu poziomych znaków, które były namalowane na drogach. To było całkiem skuteczne rozwiązanie. Pojazd na dystansie 1900 kilometrów mógł samodzielnie jechać przez 94% trasy.

Niedługo później cały świat otrzymał niezakłócony sygnał GPS, a powszechny dostęp do nawigacji satelitarnej otworzył nowy rozdział w historii autonomicznych samochodów. Jeszcze stosunkowo niedawno temat interesował przede wszystkim firmy technologiczne i uniwersytety. W tym momencie przewiduje się, że około 2030 roku rynek aut autonomicznych będzie wart 10 bilionów dolarów.

3. Technologie stosowane w samochodach autonomicznych

Jak to możliwe, że samochód bez kierowcy jedzie w dobrą stronę? Co sprawia, że nie zderza się z innymi obiektami? Pojazdy bezzałogowe poruszają się we właściwym kierunku i omijają przeszkody dzięki takim technologiom jak radar, lidar, GPS czy widzenie komputerowe. Co to takiego?



Rys. 2 Przedstawienie funkcji i elementów w samochodzie autonomicznym; Źródło: https://www.motofaktor.pl/jak-dziala-samochod-autonomiczny/

- Radar to urządzenie, które wyszukuje obiekty za pomocą fal radiowych.
- Lidar działa podobnie do radaru, lecz zamiast fal wykorzystuje światło lasera.
- GPS to system nawigacji satelitarnej, który dostarcza informacji o położeniu na podstawie wysyłanych na orbitę okołoziemską sygnałów radiowych.
- Widzenie komputerowe (inaczej zwane rozpoznawaniem obrazu) polega na przetwarzaniu obrazu przez maszynę w opis cyfrowy w celu dalszego wykorzystania.

4. Samochody autonomiczne – Poziomy autonomiczności samochodów

Należy zauważyć, że nie wszystkie systemy samojezdne są jednakowe. Norma SAE J3016 określa poziomy zdolności do samodzielnej jazdy danego pojazdu, które omówimy szczegółowo poniżej.

Poziom 0: Tu nie można mówić o samojezdności. Na tym poziomie kierowca sprawuje całkowitą kontrolę przez cały czas, nawet jeśli wspierają go aktywne układy bezpieczeństwa (które nie zmieniają autonomicznie toru ruchu pojazdu w sposób trwały). Określenie „trwały" jest o tyle istotne, że SAE oprócz różnych funkcji ostrzegawczych (np. monitorowanie martwego pola, ostrzeganie przed zjechaniem z pasa ruchu) wymienia jako przykład automatyczny hamulec awaryjny, który może w danym momencie załączyć się samoczynnie.

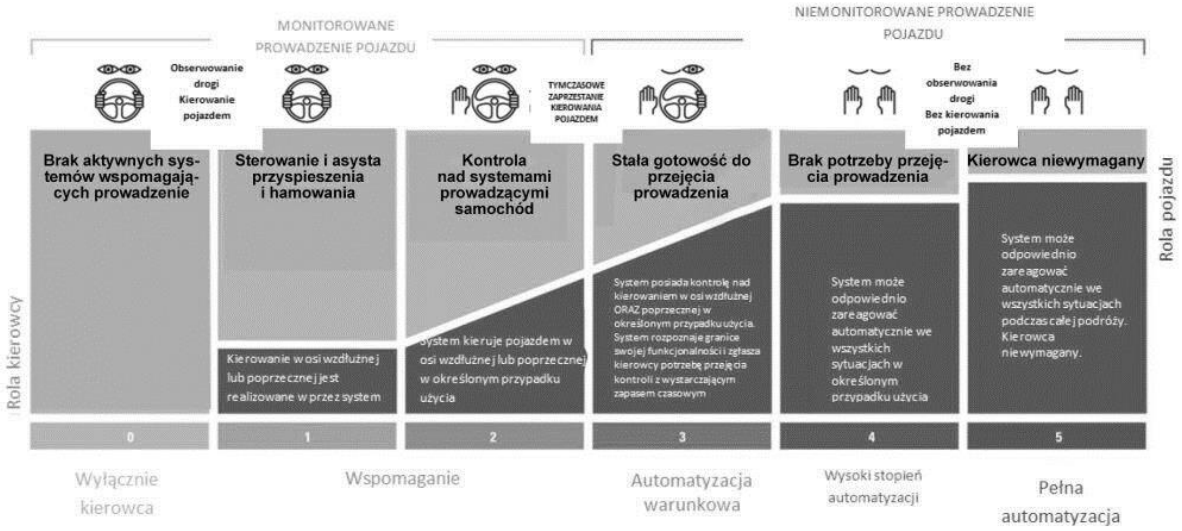
Poziom 1: Tutaj w ruch pojazdu mogą ingerować niektóre funkcje wspomagające kierowcę (advanced driver assistance systems – ADAS) (samochód może albo zmienić prędkość albo kierować samochodem w sposób ograniczony, ale nie obie czynności jednocześnie). Istotne jest to, że człowiek nadal jest odpowiedzialny za kierowanie pojazdem. Przykładem może być system utrzymywania pasa ruchu i tempomat, z których tylko jeden jest aktywny w danym momencie.

Poziom 2: Tutaj mamy do czynienia z częściową automatyzacją jazdy. Na tym poziomie pojazd może wykonywać oba rodzaje manewrów autonomicznie, ale nadal jest ściśle nadzorowany przez człowieka. Do tego poziomu zalicza się większość samochodów dostępnych obecnie na rynku – w tym modele Tesli.

Poziom 3: Na tym poziomie system jest w stanie wykonywać złożone zadania, tzn. jednostki zainstalowane w pojeździe mogą autonomicznie wykonywać cały proces prowadzenia pojazdu, jeśli spełnione są pewne wymagania – ale ważnym zastrzeżeniem jest to, że osoba za kierownicą na poziomie 3 musi nadal być gotowa przejąć kontrolę, jeśli samochód tego zażąda lub jeśli wystąpi awaria. Wśród systemów wspomagających kierowcę, SAE wymienia tzw. asystenta zatorów drogowych.

Poziom 4: Dopiero na tym poziomie można mówić o samojezdności. W tym przypadku pojazd nie oczekuje już od kierowcy interwencji w razie problemu, pojazd samodzielnie radzi sobie w trudnych sytuacjach.

Poziom 5: Obecnie najwyższy znany poziom samojezdności, na którym pojazd jedzie całkowicie sam, w odróżnieniu od poziomu 4, w którym ograniczeniem dla uruchomienia funkcji są prędkość, pora dnia i warunki drogowe.



Rys.3 Przedstawienie ingerencji kierowcy w prowadzenie samochodu według poziomów autonomiczności; Źródło: https://r-scale-20.dcs.redcdn.pl/scale/o2/tvn/web-content/m/p1/i/8c3039bd5842dca3d944faab91447818/1c51d5ea-7593-466b-80ec-eb5b3d79dee5.jpg?type=1&srcmode=4&srcx=0/1&srcy=0/1&srcw=1018&srch=2000&dstw=1018&dsth=2000&quality=80

5. Jak samochody autonomiczne zmienią nasze życie?

Autonomiczne samochody mają wiele potencjalnych zalet. Jednym z głównych korzyści jest poprawa bezpieczeństwa drogowego. Według badań, większość wypadków drogowych wynika z błędów ludzkich. Techniki autonomiczne eliminują ten czynnik, co może znacznie zmniejszyć liczbę wypadków na drogach. Ponadto, autonomiczne samochody są w stanie reagować szybciej niż człowiek, co może pomóc uniknąć kolizji.

Kolejną korzyścią jest większa mobilność osób niepełnosprawnych i starszych. Autonomiczne samochody mogą dostarczać niezależność i swobodę podróżowania tym, którzy nie są w stanie prowadzić samochodu z powodu ograniczeń fizycznych lub wieku. Dzięki temu, osoby niepełnosprawne i starsze będą miały łatwiejszy dostęp do miejsc, do których wcześniej miały trudności w dotarciu.

Jednak autonomiczne samochody stoją również przed wieloma wyzwaniami. Jednym z największych wyzwań jest zapewnienie odpowiedniej infrastruktury drogowej. Systemy autonomiczne muszą być zintegrowane z inteligentną infrastrukturą, taką jak sygnalizacja świetlna, znaki drogowe i sieć komunikacyjna. Ponadto, konieczne jest stworzenie odpowiednich ram prawnych, które będą regulować używanie autonomicznych samochodów na drogach.

6. Autonomiczne samochody - Wpływ na bezpieczeństwo drogowe

Systemy autonomiczne są wyposażone w różne czujniki i kamery, które monitorują otoczenie samochodu w czasie rzeczywistym. Dzięki temu, samochód może rozpoznawać inne pojazdy, pieszych, znaki drogowe i inne elementy infrastruktury drogowej. Systemy te analizują dane, przetwarzają je i podejmują odpowiednie decyzje w celu uniknięcia kolizji lub innych niebezpiecznych sytuacji.

Rys.4 Bezpieczeństwo na drodze Źródło: https://ocdn.eu/pulscms-transforms/1/O4YktkpTURBXy85MTEzZThhZGNmNTA5N2Q1OGY4OTYxZjg2Y2MyYjcyYi5qcGeSlQMAOc0HgM0EOJMFzQSwzQKj

Ponadto, autonomiczne samochody stosują zaawansowane algorytmy i sztuczną inteligencję, które pozwalają im przewidywać zachowanie innych uczestników ruchu drogowego. Dzięki temu, mogą dostosowywać swoją jazdę do zmieniających się warunków na drodze i reagować na sytuacje awaryjne w sposób skuteczny i bezpieczny.

7. Infrastruktura drogowa i ramy prawne autonomicznych pojazdów

Wprowadzenie autonomicznych samochodów na drogi wymaga odpowiedniej infrastruktury oraz regulacji prawnych. Infrastruktura drogowa musi być dostosowana do współpracy z autonomicznymi systemami. Na przykład, sygnalizacja świetlna musi być skomunikowana z samochodami, aby umożliwić płynny ruch i zapobiegać kolizjom.

Ponadto, konieczne jest opracowanie i wdrożenie odpowiednich ram prawnych, które będą regulować użytkowanie autonomicznych samochodów. Należy określić zasady dotyczące

odpowiedzialności za wypadki, ochrony danych, etyki postępowania systemów autonomicznych i inne aspekty związane z ich bezpiecznym i skutecznym funkcjonowaniem.

Autonomiczne samochody to technologia przyszłości, która ma potencjał znacząco zmienić nasze życie. Zalety autonomicznej jazdy, takie jak poprawa bezpieczeństwa drogowego i większa mobilność dla osób niepełnosprawnych i starszych, są nie do przecenienia. Jednak wprowadzenie autonomicznych samochodów wiąże się również z wyzwaniami, takimi jak odpowiednia infrastruktura drogowa i ramy prawne.

8. Jakie mogą być wady korzystania z autonomicznych pojazdów?

Mimo wielu zalet, jakie niesie ze sobą autonomiczna jazda, istnieją również pewne wady i wyzwania związane z korzystaniem z tego rodzaju pojazdów. Poniżej przedstawiam niektóre z potencjalnych wad autonomicznych samochodów:

-Bezpieczeństwo

Mimo że autonomiczne samochody mają potencjał poprawy bezpieczeństwa drogowego, to wciąż istnieje ryzyko awarii technicznych lub błędów systemu, które mogą prowadzić do wypadków. Ponadto, niektóre sytuacje na drodze, takie jak nieprzewidywalne zachowanie innych kierowców lub sytuacje awaryjne, mogą stanowić wyzwanie dla systemów autonomicznych.

-Koszty

Aktualnie, technologia autonomiczna jest kosztowna. Wprowadzenie autonomicznych samochodów na drogi wymaga znacznych inwestycji w badania i rozwój, a także w produkcję i dostosowanie infrastruktury drogowej. Wyższe koszty mogą przekładać się na wyższe ceny samochodów dla konsumentów.

-Odpowiedzialność i ubezpieczenie

W przypadku wypadków lub awarii autonomicznego samochodu, pojawiają się pytania dotyczące odpowiedzialności i ubezpieczenia. Kto ponosi odpowiedzialność za ewentualne szkody? Jakie są ramy prawne i ubezpieczeniowe dla autonomicznych pojazdów? Te kwestie są wciąż przedmiotem dyskusji i wymagają jasnych regulacji.

-Brak zaangażowania kierowcy

Mimo że autonomiczne samochody mają na celu ułatwienie podróży, niektórzy argumentują, że mogą one prowadzić do utraty zaangażowania kierowcy i spadku umiejętności reakcji w sytuacjach awaryjnych. Zbyt duża zależność od systemów autonomicznych może prowadzić do utraty umiejętności samodzielnego prowadzenia pojazdu.

-Prywatność i bezpieczeństwo danych

Autonomiczne samochody gromadzą i przetwarzają duże ilości danych dotyczących naszych podróży, nawyków i preferencji. Istnieje zatem ryzyko naruszenia prywatności i bezpieczeństwa tych danych. Konieczne jest odpowiednie zabezpieczenie danych i przestrzeganie przepisów dotyczących ochrony prywatności.

-Brak kontroli nad pojazdem

Dla niektórych kierowców, korzystanie z autonomicznych samochodów może oznaczać utratę poczucia kontroli nad pojazdem. Niektórzy preferują tradycyjną jazdę, gdzie to oni sami podejmują decyzje i kontrolują wszystkie aspekty podróży. Autonomiczne samochody mogą ograniczać pewne aspekty indywidualnej kontroli, takie jak wybór trasy, tempo jazdy i ogólna interakcja z pojazdem.

-Zatrudnienie, przemysł transportowy i akceptacja społeczna

Wprowadzenie autonomicznych samochodów może mieć wpływ na rynek pracy w branży transportowej. Niektóre stanowiska, takie jak kierowcy zawodowi, mogą być zagrożone automatyzacją. Konieczne będzie dostosowanie się do zmian na rynku pracy i zapewnienie alternatywnych możliwości zatrudnienia dla osób dotkniętych tymi zmianami.

Tak drastyczna zmiana niewątpliwie wymaga akceptacji społecznej. Niektórzy ludzie mogą być nieufni wobec nowej technologii i niechętni przekazaniu kontroli nad swoją podróżą maszynom. Konieczne jest budowanie zaufania i świadomości w społeczeństwie dotyczącej korzyści, bezpieczeństwa i skuteczności autonomicznych samochodów.

9. Wpływ samochodów autonomicznych na społeczeństwo

Rozwój samochodów autonomicznych ma również ogromny wpływ na nasze społeczeństwo. Jednym z najważniejszych aspektów jest zmiana w podejściu do podróżowania. Dzięki samochodom autonomicznym, podróżowanie staje się wygodniejsze i mniej stresujące. Kierowcy nie muszą już martwić się o korki czy szukanie miejsca parkingowego, ponieważ pojazdy same poradzą sobie z tymi problemami. Ponadto, osoby starsze lub niepełnosprawne będą mogły cieszyć się większą niezależnością i swobodą w przemieszczaniu się.

Kolejnym ważnym aspektem jest zmiana w branży transportowej. Wraz z rozwojem samochodów autonomicznych, pojawią się nowe możliwości biznesowe, takie jak usługi car-sharingowe czy dostawy towarów bez udziału kierowców. To może przyczynić się do zmniejszenia liczby pojazdów na drogach oraz poprawy jakości powietrza w miastach. Ponadto, rozwój tej technologii może przyczynić się do zmniejszenia kosztów transportu i poprawy efektywności logistycznej.

10. Wpływ samochodów autonomicznych na branżę ubezpieczeniową

Wprowadzenie samochodów autonomicznych na rynek motoryzacyjny pociąga za sobą szereg zmian w branży ubezpieczeniowej. Tradycyjne modele polis i określanie odpowiedzialności za wypadki muszą być dostosowane do nowej rzeczywistości. Choć samochody autonomiczne mogą przynieść korzyści w postaci redukcji ryzyka wypadków, zmiany te rodzą również wyzwania. Ubezpieczyciele muszą opracować nowe modele ubezpieczeń, uwzględniające m.in. aspekty techniczne i etyczne związane z działaniem autonomicznych pojazdów. Jednocześnie, pojawiają się także nowe dylematy, takie jak określanie odpowiedzialności w przypadku awarii technicznych czy decyzji algorytmów. Współczesna branża ubezpieczeniowa stoi zatem przed koniecznością adaptacji i innowacji, aby sprostać zmieniającym się potrzebom rynku związanym z wprowadzeniem samochodów autonomicznych.

Ciekawostka: Ponad 11 milionów kilometrów i zaledwie 3 wypadki. Waymo radzi sobie zaskakująco dobrze.

Od momentu powstania do końca października 2023 roku, pojazdy Waymo przejechały bez kierowcy około 8,5 miliona kilometrów w Phoenix, prawie 3 miliony kilometrów bez kierowcy w San Francisco i kilka tysięcy km bez kierowcy w Los Angeles. Jak informuje samo Waymo, podczas wszystkich tych przejechanych kilometrów miały miejsce trzy wystarczająco poważne wypadki, aby spowodować obrażenia. Były to:

-W lipcu Waymo w Tempe w Arizonie zahamował, aby uniknąć uderzenia w powaloną gałąź, co doprowadziło do zderzenia trzech samochodów. Pasażer Waymo nie miał zapiętych pasów bezpieczeństwa i odniósł obrażenia, które Waymo określił jako drobne.

-W sierpniu Waymo na skrzyżowaniu „zaczął jechać do przodu", ale następnie „zwolnił, aż się zatrzymał" i został uderzony w tył przez SUV. SUV opuścił miejsce zdarzenia bez wymiany informacji, a pasażer Waymo zgłosił lekkie obrażenia.

-W październiku pojazd Waymo w Chandler w Arizonie jechał lewym pasem, gdy wykrył inny pojazd nadjeżdżający z tyłu z dużą prędkością. Waymo próbował przyspieszyć, aby uniknąć kolizji, ale został uderzony od tyłu. Znów doszło do obrażeń, ale Waymo określił ją jako drobne.

11. Podsumowanie

Rozwój samochodów autonomicznych jest nie tylko rewolucją w świecie motoryzacji, ale również ma ogromny wpływ na nasze społeczeństwo. Dzięki zaawansowanej technologii i

systemom bezpieczeństwa, samochody autonomiczne mogą przyczynić się do zmniejszenia liczby wypadków drogowych oraz poprawy jakości podróżowania. Ponadto, wprowadzenie tej technologii może przyczynić się do zmian w branży transportowej i poprawy jakości życia w miastach. Choć jeszcze wiele wyzwań stoi przed rozwojem samochodów autonomicznych, to już teraz możemy zauważyć ich pozytywny wpływ na nasze społeczeństwo.

Źródła internetowe

1. https://www.link4.pl/blog/czym-tak-wlasciwie-jest-samochod-autonomiczny-odpowiada-ekspert (dostęp: 21.04.2024)
2. https://motofocus.pl/informacje/108645/samochody-autonomiczne-aktualny-etap-rozwoju (dostęp: 21.04.2024)
3. https://mubi.pl/poradniki/samochod-autonomiczny/ (dostęp: 21.04.2024)
4. https://drogowa-pomoc.pl/jak-autonomiczne-samochody-zmienia-nasze-zycie/(dostęp: 21.04.2024)
5. https://correctbiuro.pl/rozwoj-samochodow-autonomicznych-i-ich-wplyw-na-spoleczenstwo/ (dostęp: 21.04.2024)
6. https://antyweb.pl/waymo-av-statystyki-11-mln-km (dostęp: 21.04.2024)

Maja Jaszowska, Filip Skawiński, Piotr Laskowski, Mateusz Fesz, Dominika Fergisz
Koło Naukowe Elektroniki i Technologii Informacyjnych

Dr inż. Bartosz PAWŁOWICZ
Opiekun Koła Naukowego

Robot Linefollower

Streszczenie

Artykuł opisuje wykonanie i sposób działania robota typu linefollower. Przedstawiono, w jaki sposób dane z czujników są przekazywane do mikrokontrolera ESP32 i analizowane przez program napisany w języku C. Zaprezentowano przykładowy sprzęt potrzebny do budowy robota oraz schematy pomocnicze wykorzystywane podczas łączenia poszczególnych komponentów. Pokazano teorię stojącą za algorytmem, dzięki któremu linefollower jest w stanie płynnie i z dużą prędkością poruszać się po wyznaczonej trasie. Opisano kod sterujący silnikami robota, który wykorzystywał dane pobrane z czujników oraz przedstawiony wcześniej algorytm. W artykule przedstawiono wyniki testów dla poszczególnych składowych niezbędnych do poprawnego działania robota. Robot linefollower, na podstawie, którego napisano artykuł, spełnił swoje zadanie i poprawnie pokonał wyznaczone trasy.

Słowa kluczowe: linefollower, robot, trasa, linia, algorytm PID.

1. Wprowadzenie

Linefollower to robot, którego zadaniem jest jazda po wyznaczonej linii. Posiada algorytm, który samodzielnie dostosowuje tor jazdy za pomocą danych popranych z czujników. Sensory analizują ilość odbitego światła, dzięki czemu algorytm jest w stanie odpowiednio zadziałać.

Celem artykułu jest prezentacja działania oraz budowy robota typu linefollower. Poprzez omówienie użytych komponentów, algorytmu PID oraz programu sterującego, artykuł ma na celu przybliżenie tematyki robotyki oraz algorytmiki, które leżą u podstaw działania tego rodzaju robotów.

2. Zawody linefollowerów

W Polsce przez lata odbywały się zawody, w których jedną z kategorii stanowiły konkurencje dedykowane tego typu robotom. Trasa zazwyczaj jest wyznaczona czarnymi liniami na białym lub bardzo jasnym tle, choć zdarzają się też trasy z białymi liniami na czarnym tle. Zarówno trasę, jak i kategorię, dostosowuje się do różnych parametrów robotów, które różnią się wielkością i rodzajem użytych materiałów. Na zawodach można spotkać osobne konkurencje dla robotów elektronicznych oraz tych zbudowanych z klocków Lego. Linefollower, który pokona trasę najszybciej i z największą precyzją, zostaje zwycięzcą.

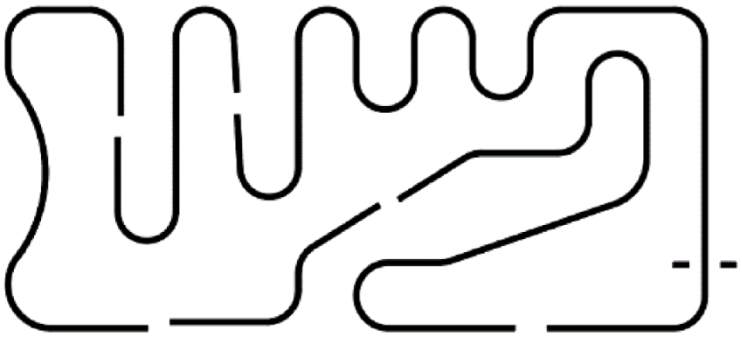
Założenia, które musi spełniać linefollower:

- robot jeździ po trasie autonomicznie, bez pomocy zewnętrznej,
- za pomocą algorytmu dobiera wartości prędkości silników na określonych punktach trasy,
- w przypadku zgubienia czarnej linii potrafi ją samodzielnie odnaleźć,
- poprawnie analizuje dane z czujników, dzięki czemu przejeżdża trasę z dużą prędkością z jak najmniejszą ilością błędów,
- jest dobrze wyważony,
- ma możliwie najmniejszą wagę,

Wszystkie te założenia są kluczowe dla skutecznego działania linefollowera i stanowią fundament podczas projektowania tego typu robotów. Prawidłowe uwzględnienie tych kryteriów zapewnia nie tylko efektywność w działaniu robota, ale także minimalizację błędów.

3. Trasa

Trasa na zawodach jest z góry określona przez organizatorów. W Polsce zazwyczaj można spotkać białą powierzchnię z czarnymi linami o szerokości od 15 do 20 mm. Na rysunku 1. została przedstawiona przykładowa trasa, po której porusza się robot.



Rysunek 1. Przykładowa trasa dla linefollowera
Źródło: http://2gym-samou.sam.sch.gr/autosch/joomla15/images/stories/robotics/samianator_report.pdf (dostęp: 25.05.2024).

Na tak wyznaczonej drodze można spotkać:

- linie proste,
- zakręty prostopadle i łuki,
- skrzyżowania i utrudnienia w postaci kilku linii wychodzących,
- linie przerywane.

Czasami jest dostępna kategoria, gdzie występują trasy na ścianie lub suficie.

4. Wybrany sprzęt

Podczas doboru odpowiednich komponentów kierowano się ich skutecznością, napięciem zasilania, wielkością, wagą oraz trwałością. Czujniki są jednymi z najważniejszych części robota. Wybrano listwę cyfrową z czujnikami odbiciowymi Pololu QTR-8A.

Specyfikacja:

- wymiary 75 x 127 x 4 mm,
- napięcie zasilania 5 V lub 3, 3 V,
- prąd zasilania 100 mA,
- sygnał na wyjściu: cyfrowy (kompatybilny z pinami I/O),
- maksymalna odległość pomiaru: 9, 5 mm.[1]

Mikrokontroler jest kluczową częścią dla robota. ESP32 jest kompatybilny z wybranymi czujnikami oraz zapewnia odpowiednią moc obliczeniową do wykonywania algorytmu śledzenia linii i sterowania silnikami. Wybrano model ESP32-S3-WROOM-1-N16R8.

Specyfikacja:

- układ: ESP32-S3-WROOM-1/1U
- procesor: Dual-Core 32-bit Xtensa LX7
- taktowanie: do 240 MHz
- komunikacja bezprzewodowa: WiFi i Bluetooth
- standard WiFi: IEEE 802.11 b/g/n (802.11n do 150 Mbps)
- pasmo: od 2,412 GHz do 2,484 GHz
- standard Bluetooth: LE 5 Mesh
- pamięć SRAM: 512 kB (w tym 16 kB w pamięci RTC), Pamięć ROM: 384 kB, Pamięć Flash: 8 MB[2]

Sterownik powinien być dopasowany do dobranych silników. Konieczne jest, aby obsługiwał wymagane napięcie oraz prąd potrzebny do poprawnego działania układu. Wybrano więc moduł sterownika Mini MX1508 przeznaczony do silników DC o zakresie napięcia od 2V do 10V i prądzie do 1.5A.

---

[1]https://botland.com.pl/czujniki-odbiciowe/20-listwa-z-czujnikami-odbiciowymi-qtr-8rc-cyfrowa-pololu-961-5903351249287.html (dostęp: 25.05.2024).

[2]https://botland.com.pl/moduly-wifi-i-bt-esp32/20739-esp32-s3-devkitc-1-n8-wifi-bluetooth-plytka-rozwojowa-z-ukladem-esp32-s3-wroom-11u-5904422382353.html?cd=1050025856&ad=51004438223&kd=&gad_source=1&gclid=Cj0KCQjwjLGyBhCYARIsAPqTz19bkTQ6sop-6Y3ugUdsIVEvCBs1hrGb6KdkcGNrm8p2TkwKad-JCNUaAmWWEALw_wcB (dostęp: 25.05.2024).

Specyfikacja:

- moduł umożliwia sterownia dwoma silnikami prądu stałego
- napięcie zasilania: 2 - 10V DC
- napięcie zasilania części logicznej: 5V
- wbudowany regulator napięcia 5V
- maksymalny prąd wyjściowy: 1.5 A na kanał, 0.8A ciągły
- wymiary płytki: 21x25x17 mm
- podając sygnał PWM na którykolwiek z tych pinów wejściowych steruje się kierunkiem i prędkością obrotową silników.[3]

Przetwornica jest niezbędna do odpowiedniego zasilania mikrokontrolera oraz czujników w linefollowerze. Moduł zasilacza obniżającego napięcie DAOKI MINI DC-DC 12-24V TO 5V 3A obsługuje wymagane napięcie wejściowe i wyjściowe.

Specyfikacja:

- napięcie wejściowe: 4,5-24V,
- napięcie wyjściowe: 0,8V do 21V,
- napięcie wyjściowe regulowane z pomocą wbudowanego potencjometru,
- możliwość ustawienia na stałe 6 różnych napięć wyjściowych (1,8V;2,5V;3,3V;5V;8V;12V),
- za pomocą zwarcia pól znajdujących się na płytce,
- maksymalny prąd wyjściowy: 3A,
- wymiary: 11x20,5x6mm.[4]

Silniki powinny mieć odpowiednie wymiary oraz być lekkie. Mogą odegrać istotną rolę w zapewnieniu dobrego wyważenia robota. Zbyt duże silniki mogą utrudniać pracę linefollowera. Ich sterowanie powinno być proste, a jednocześnie posiadają wysoki moment obrotowy. Wybrano silnik DC GA12-N20 6V z przekładnią o prędkości 2000 obr/min.

Specyfikacja:

- napięcie zasilania: od 1V do 6V,
- prąd na biegu jałowym: pon. 0,05A,
- moment obrotowy: 30kg*cm (3Nm),

[3]https://sklep.avt.pl/pl/products/modul-sterownika-mini-mx1508-do-silnikow-dc-podwojny-arduino-187397.html (dostęp: 25.05.2024).
[4]https://rif.sklep.pl/az/regulatory-napiecia/655-mini-przetwornica-step-down-08-21-3a-975.html (dostęp: 25.05.2024).

- wymiary: 25mm dł. x 12mm szer. x 10mm. wys.,
- średnica wału: 3mm,
- długość wału: 9mm.[5]

Koła Solarbotics RW2 Pololu 642 charakteryzują się dobrą przyczepnością, co wspiera robota podczas ostrych skrętów. Doskonale sprawdzają się również podczas gwałtownych zmian prędkości.

Specyfikacja:

- zestaw zawiera oponę, piastę oraz śrubkę mocującą (jedno kompletne koło)
- średnica: 31,2 mm,
- szerokość: 13,2 mm,
- średnica otworu: 3 mm,
- masa: 12 g,
- koła mocowane na zewnątrz.[6]

Akumulatory Li-Po cechują się jedną z najlepszych wydajności energetycznych w porównaniu do innych typów akumulatorów. Mają wysoki stosunek mocy do wagi, co czyni je atrakcyjnym wyborem dla robotów linefollower, które często muszą być lekkie i zwinne. Akumulator Li-Pol Dualsky 520mAh 25C 2S 7,4V dostarcza wysoki prąd w krótkim czasie, co dobrze współgra z całym układem.

Specyfikacja:

- dwa ogniwa (2S) Li-Po,
- napięcie nominalne: 7,4 V,
- pojemność: 520 mAh,
- wymiary: 58 x 18 x 19 mm.[7]

5. Budowa

Głównym celem robota jest poruszanie się po wyznaczonej linii. Podczas konstruowania linefollowera niezbędne jest zaimplementowanie mechanizmu umożliwiającego pobieranie

[5]https://abc-rc.pl/product-pol-9257-Mini-silnik-szczotkowy-GA12-N20-50RPM-wal-9mm-3-6V-z-przekladnicva.html (dostęp: 25.05.2024).
[6]https://botland.com.pl/kola-z-oponami/147-kolo-solarbotics-rw2-mocowanie-zewnetrzne-pololu-642-5903351248150.html (dostęp: 25.05.2024).
[7]https://botland.com.pl/akumulatory-li-pol-2s-74v-/570-pakiet-li-pol-dualsky-520mah-25c-2s-74v-6941047107427.html?cd=19993067448&ad=&kd=&gad_source=1&gclid=Cj0KCQjwjLGyBhCYARIsAPqTz18ktWWt5vxUDfN3VAo-B0eI9fq4_dvstwlujnqk-5hzmR1qZpwnHggaAn-CEALw_wcB (dostęp: 25.05.2024).

danych na temat trasy. Dokładne zbieranie informacji o trasie jest kluczowe dla skutecznego poruszania się robota po linii, umożliwiając mu precyzyjne reakcje na zmiany w otoczeniu.

ESP32-S3-DevKitC-1 w wersji N16R8 posiadający rozszerzoną pamięć FLASH do 16MB oraz PSRAM do 8MB. Ten wydajny mikrokontroler bazuje na 32-bitowym procesorze Xtensa LX7 o dwóch rdzeniach i częstotliwości 240 MHz. Do naszej dyspozycji dostajemy 2 porty USB-C - jeden do połączenia po UART, drugi do komunikacji i zasilania modułu. ESP32-S3 obsługuje połączenia bezprzewodowe WiFi oraz Bluetooth 5.0 LE.[8]

Moduł czujnika odbiciowego QTR-8A jest przeznaczony jako czujnik linii, ale może być używany jako uniwersalny czujnik zbliżeniowy lub odbiciowy. Moduł jest wygodnym nośnikiem dla ośmiu par nadajników i odbiorników IR (fototranzystorów) rozmieszczonych równomiernie w odstępach 0,375" (9,525 mm). Każdy fototranzystor jest podłączony do rezystora podciągającego, aby stworzyć dzielnik napięcia, który generuje analogowe napięcie wyjściowe w zakresie od 0 V do VIN (zazwyczaj 5 V) w zależności od odbitego promieniowania IR. Niższe napięcie wyjściowe oznacza większe odbicie.[9]

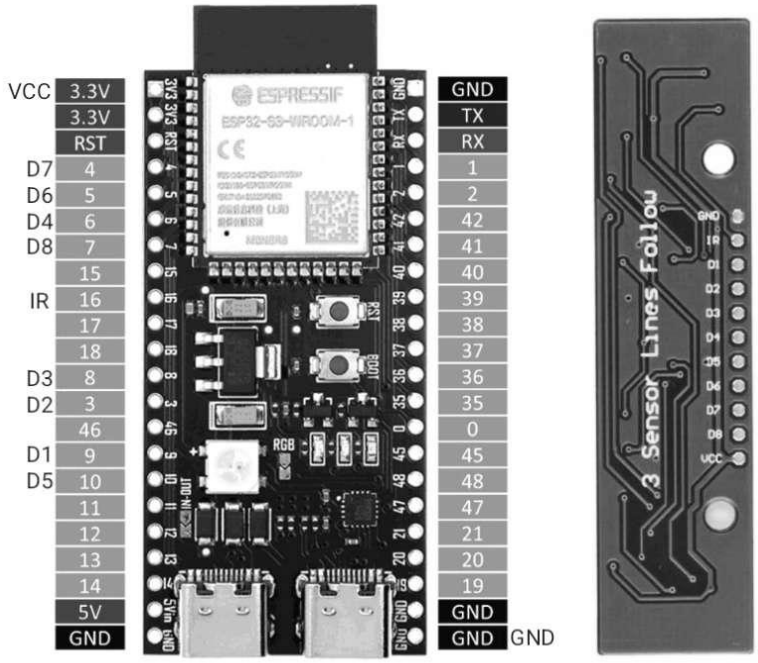
Płytka obsługuje napięcie 5V lub 3,3V, jednak w projekcie wykorzystuje się napięcie 3,3V. Dzięki swoim niewielkim wymiarom, które wynoszą 75 x 127 x 4 mm, linefollower jest dość lekki i kompaktowy, co ułatwia manewry.

Wszystkie wyjścia są niezależne, ale diody LED są ułożone parami, aby zmniejszyć zużycie prądu o połowę. Diody LED są sterowane przez MOSFET z bramką normalnie podciągniętą do góry, co pozwala na wyłączenie diod LED poprzez ustawienie bramki MOSFET na niskie napięcie. Wyłączenie diod LED może być korzystne w celu ograniczenia zużycia energii, gdy czujniki nie są używane lub w celu zmiany efektywnej jasności diod LED poprzez sterowanie PWM.[10]

Podczas montażu bardzo ważne jest, aby diody były w pewnej niewielkiej odległości od powierzchni. W przeciwnym wypadku czujniki będą wykrywać linie w sposób niepoprawny, co prowadzi do problemów z jazdą robota. Przed użyciem listwy należało pamiętać, by zalutować część oznaczoną 3,3V. Jest to wbudowany rezystor, bez którego nie jest możliwe bezpieczne korzystanie z czujników. Na internecie można znaleźć bardzo podobny model, bez wbudowanego rezystora. W takim wypadku trzeba dokupić i dolutować odpowiedni rezystor w wyznaczonym przez producenta miejscu. Moduł QTR8-A ma piny, które należało podłączyć

[8]https://elektroweb.pl/esp32/1312-modul-esp32-s3-devkitc-1-wroom-1-n16r8-16mb-flash-wifi-bluetooth-usb-c-5905523309461.html (dostęp: 25.05.2024).
[9] https://www.pololu.com/product/960 (dostęp: 25.05.2024).
[10]https://www.pololu.com/product/960 (dostęp: 25.05.2024).

do mikrokontrolera. Zarówno piny o numerach D1-D8 jak i te VCC, GND i IR. W przypadku jakiegokolwiek błędu podczas montażu ESP32 nie łączy się z wifi. Rysunek 2. przedstawia schemat łączenia czujników z ESP32.



Rysunek 2. Łączenie mikrokontrolera z czujnikami Pololu QTR-8RC
Źródło: https://ifuturetech.org/product/qtr-8rc-reflectance-sensor-array/ (dostęp: 25.05.2024), https://99tech.com.au/product/esp32-s3-yd-n8r8/ (dostęp: 25.05.2024). Opracowanie własne.

Mikrokontroler należało przylutować do płytki prototypowej za pomocą jednorzędowej listwy kołkowej goldpinów męskich po obu stronach ESP32. Do płytki wymagane było dolutowanie kolejnego zestawu 11 goldpinów męskich jednorzędowych. Tak jak na schemacie wyżej, konieczne jest połączenie 11 goldpinów z odpowiednimi pinami mikrokontrolera posiadającymi funkcję ADC. Podczas przygotowywania czujników wskazane jest zalutowanie oznaczenia 3.3V oraz przylutowanie goldpinów męskich. Łączenie czujników z mikrokontrolerem odbywa się za pomocą zestawu 11 przewodów połączeniowych JustPi żeńsko-żeńskich, gdzie po jednej stronie są wpięte do pinów ESP a po drugiej do pinów czujnika.

Takie rozwiązanie wymaga utworzenia podpory, która zostanie przykręcona za pomocą śrub do płytki prototypowej, oraz czujników, które posiadają odpowiednie wypusty do tego celu w swojej konstrukcji. Istotne jest, aby podpora była odpowiednio długa, co pomaga robotowi w zwiększeniu czasu reakcji na nagłe zmiany kierunku trasy. W przypadku budowy

robota z myślą o zawodach, istotne jest sprawdzenie, jakie wymiary są wymagane przez organizatora. Zazwyczaj są to 210 x 297 mm, czyli rozmiar kartki A4.

Ważne jest pamiętać, że przed rozpoczęciem trasy należy przeprowadzić kalibrację robota. Proces ten zależy od programu i trwa zazwyczaj od 10 do 25 sekund. Kalibracja jest niezbędna, aby czujniki mogły dostosować się do warunków oświetleniowych panujących na powierzchni. Polega ona na powolnym przesuwaniu czujników wzdłuż fragmentu toru, najlepiej pomiędzy czarną linią a białym podłożem.

Moduł sterownika Mini MX1508 to sterownik silników umożliwia sterowanie dwoma silnikami DC. Układem można sterować przy pomocy Arduino, podpinając jego piny pod następujące wejścia układu: IN1, IN2, IN3, IN4. Podając sygnał PWM na którykolwiek z tych pinów wejściowych steruje się kierunkiem i prędkością obrotową silników. [11] Dodatkowym elementem potrzebnym do prawidłowego funkcjonowania robota będzie tzw. mostek H, który pozwala m.in. na zmianę kierunku obrotów silnika oraz sterowanie ich prędkością.[12] Pod pojęciem PWM kryje się anglojęzyczny skrót „Pulse Width Modulation", który oznacza modulację szerokości impulsów. Jest to nic innego jak metoda sterowania układami elektronicznymi przy pomocy manipulacji samym sygnałem sterującym, a konkretniej jego parametrami technicznymi. Wynika to z samej charakterystyki fizycznej sygnału PWM, który składa się ze wspomnianych wcześniej impulsów. Ich podstawowe parametry dotyczą czasu pozostania w stanie wysokim (wartość logiczna 1) oraz częstotliwości przełączania sygnału PWM.[13]

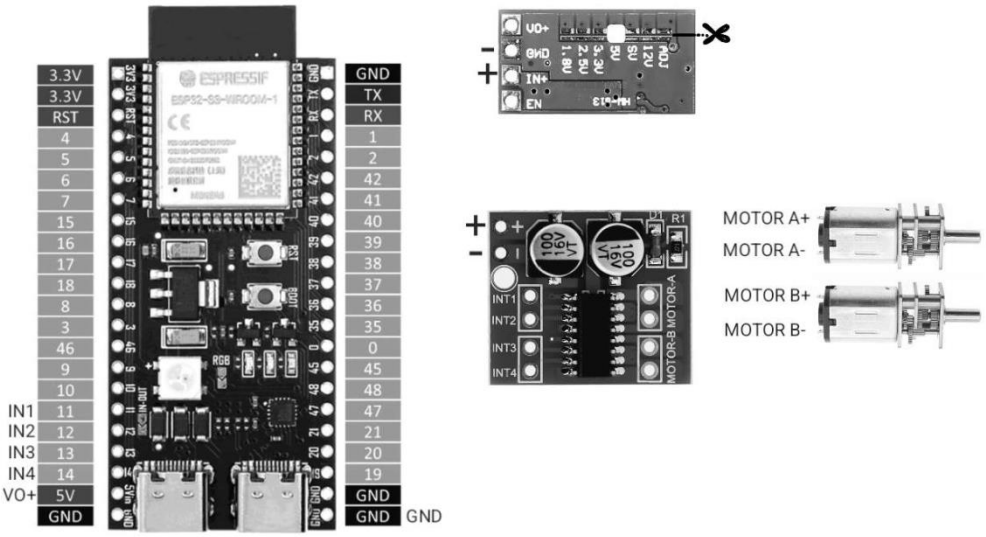
Zasilanie należy podłączyć do pinów GND i VCC. Na sterowniku oznaczono piny dla silnika A i silnika B. Moduł posiada piny o następujących oznaczeniach: IN1, IN2, IN3, IN4. Dla silnika A są dedykowane piny sterujące IN1, IN2, a dla silnika B IN3, IN4. Ta informacja jest istotna, aby zrozumieć sposób sterowania kierunkiem obrotu silników. Sygnał PWM jest wykorzystywany do regulacji prędkości.

Oznacza to, że gdy na obu pinach IN1 i IN2 jest stan wysoki lub niski, silniki zostaną zatrzymane. Natomiast, jeśli na IN1 jest stan wysoki, a na IN2 stan niski (lub odwrotnie), silnik będzie obracał się w określonym kierunku (lub w przeciwnym). Te możliwości pozwalają na bardzo precyzyjną kontrolę prędkości silników, co jest istotne przy sterowaniu robotem liniowym.

---

[11]https://abc-rc.pl/product-pol-17960-Modul-sterownika-Mini-MX1508-do-silnikow-DC-podwojny-Arduino.html (dostęp: 25.05.2024).
[12]https://zpe.gov.pl/a/przeczytaj/DUpr2MS6R (dostęp: 25.05.2024).
[13]https://botland.com.pl/blog/sygnal-pwm-czym-jest/ (dostęp: 25.05.2024).

Przetwornica DAOKI służy do konwersji napięcia wejściowego z zasilacza na niższe napięcie wyjściowe, odpowiednie dla ESP32. Moduł może obsługiwać napięcia wejściowe od 4,5V do 24V, a użytkownik ma możliwość wyboru pożądanego napięcia wyjściowego. Piny IN+ i GND służą do połączenia z akumulatorem, natomiast pin VO+ umożliwia bezpośrednie zasilanie mikrokontrolera ESP32. Rysunek 3. przedstawia schemat łączenia ESP32, przetwornicy i modułu sterowania.

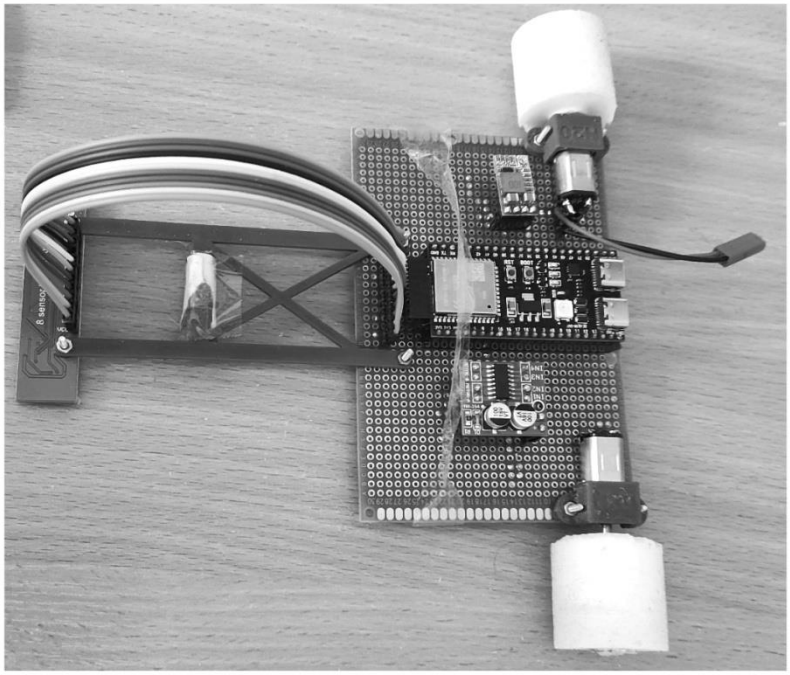


Rysunek 3. Łączenie mikrokontrolera z czujnikami Pololu QTR-8RC
Źródła: https://quartzcomponents.com/products/mini-l298 (dostęp: 25.05.2024), https://99tech.com.au/product/esp32-s3-yd-n8r8/ (dostęp: 25.05.2024), https://www.amazon.ca/DAOKI-Supply-Voltage-Converter-Adjustable/dp/B07X86VKCL (dostęp: 25.05.2024), https://www.amazon.pl/elektryczny-Motoreduktor-wolnoobrotowy-skrzynia-obrotowym/dp/B0C4PH7HS8 (dostęp: 25.05.2024). Opracowanie własne.

Przed zamontowaniem regulatora w układzie konieczne jest wlutowanie na płytce obok zaznaczenia 5V oraz przecięcie linii zaznaczonej na schemacie za pomocą ostrego narzędzia. Tylko wtedy regulator napięcia będzie działał poprawnie. Zarówno przetwornica, jak i sterownik silników należy przylutować do płytki prototypowej przy użyciu goldpinów. Ważne jest, aby połączyć pin przetwornicy VO+ z pinem oznaczonym, jako 5V oraz pin GND z odpowiednim pinem mikrokontrolera. Niezbędne jest podłączenie zasilania do obu części w sposób, który wykluczy możliwość zwarcia. Piny IN1-IN4 muszą zostać połączone z ESP, aby wgrany na nią program mógł sterować silnikami podłączonymi do modułu za pomocą MOTOR-A i MOTOR-B.

Połączenie silników do sterownika to tylko pierwszy krok, by zapewnić stabilność całego systemu. Dla zapewnienia bezpieczeństwa silniki powinny być dodatkowo przymocowane do płytki prototypowej w taki sposób, aby solidnie trzymały się całego linefollowera.

Rysunek 4 przedstawia poglądowy wygląd zbudowanego robota.



Rysunek 4. Zbudowany linefollower.
Źródło: Opracowanie własne.

6. Teoria i algorytm

Zrozumienie działania linefollowera staje się bardziej oczywiste przy użyciu mniejszej ilości czujników. Robot wyposażony w pojedynczy czujnik nie jest jednak efektywny - wykonuje nadmiarowe ruchy, porusza się wolno i łatwo gubi linię, szczególnie na bardziej skomplikowanych trasach. Choć może działać wystarczająco dobrze na prostych i lekko zakrzywionych liniach, ma problemy przy bardziej wymagających zakrętach.

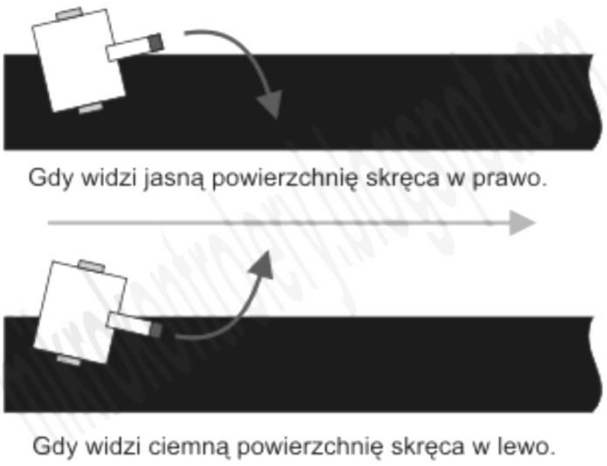
Jego działanie działa na założeniach:

- kiedy czujnik IR widzi czarną linię, jedno koło bota obraca się i bot skręca z dala od linii,
- kiedy czujnik IR widzi białe tło, drugie koło obraca się i bot skręca lekko w kierunku linii.

Powyższe dwa kroki następują po sobie w trybie powtarzającym się bardzo szybko w bardzo krótkim czasie, co daje nam ruch bota podążającego ścieżką wzdłuż czarnej linii. [14]

Rysunek 5. Przedstawia w jaki sposób porusza się linefollower z jednym czujnikiem.

[14]https://www.instructables.com/SIMPLE-LINE-FOLLOWER-ROBOTsingle-Sensor/ (dostęp: 25.05.2024).



Rysunek 5. Robot z jednym czujnikiem, kierunki jazdy.
Źródło: https://mikrokontrolery.blogspot.com/2011/03/Robotyka-Linefollower-Listwa-czujnikow-linii.html (dostęp: 25.05.2024).

Linefollower z dwoma czujnikami jest równie prosty w zrozumieniu co ten z jednym czujnikiem, ale działa bardziej precyzyjnie. Pozwala to na zwiększenie prędkości bez utraty dokładności podążania za linią. Niestety, nadal nie jest to idealne rozwiązanie. Robot wciąż ma niską odporność na błędy, śledzenie zakrzywionych tras nadal nie jest idealne, a nawet najmniejsze zmiany mogą być zbyt gwałtowne przy wyższych prędkościach. Rysunek 6. Przedstawia w jaki sposób porusza się linefollower z dwoma czujnikami.

Rysunek 6. Robot z dwoma czujnikami na trasie.
Źródło: https://forbot.pl/blog/kurs-budowy-robotow-line-follower-czyli-bolid-f1-id19363 (dostęp: 25.05.2024).

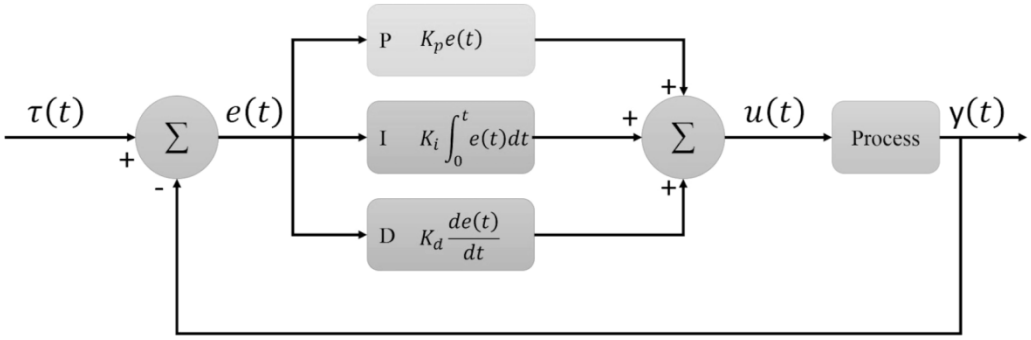
Większa ilość czujników jest najlepszą decyzją przy budowaniu linefollowera. Zbyt mało czujników sprawia, że robot trudniej znajduje małe zmiany w trasie, a za dużo może doprowadzić do zbyt wielu zmiennych, które ułatwiają dezorientacje przy zbieraniu danych przez czujniki.

Najlepiej więc wybrać od 5 do 20, kilkanaście jest najbardziej optymalne.[15]

[15]https://mikrokontrolery.blogspot.com/2011/03/Robotyka-Linefollower-Listwa-czujnikow-linii.html (dostęp: 25.05.2024).

Kod linefollowera korzysta z algorytmu PID.

W prostych słowach, algorytm PID reguluje zmienną procesu poprzez obliczenie sygnału sterującego, który jest sumą trzech składowych: proporcjonalnej, całkującej i różniczkującej. Stąd jego nazwa. W efekcie algorytm może przywrócić zmienną procesu do akceptowalnego zakresu.[16] Rysunek 7. przedstawia graficzne działanie algorytmu PID, Rysunek 8. przedstawia wzór powyżej wspomnianego algorytmu.



Rysunek 7. Wygląd algorytmu PID.
Źródło: https://www.integrasources.com/blog/basics-of-pid-controllers-design-applications/ (dostęp: 25.05.2024).

Część proporcjonalna (P) określa z jak dużą siłą, robot ma skręcać na trasie w zależności od błędu regulacji. Błędem określa się różnicę pomiędzy wartością zadaną a rzeczywistą wartością wyjściową systemu.

Czyli odpowiada za to, jak szybko silniki będą się obracać na podstawie położenia linii. Jeżeli linia będzie po prawej stronie czujnika to robot "wie", że musi skręcić w tym kierunku i zmniejszy szybkość prawego silnika.[17]

Całkująca (I) jest sumą błędów regulacji, które występują w danym czasie, eliminuje to różnicę między wartością zadaną a wartością ustaloną systemu. Nie wykorzystuje się składowej całkującej w programach dla linefollowerów.

Ostatnia składowa różniczkująca (D) przewiduje przyszłe zmiany błędu na podstawie obecnego tempa zmiany. Wygładza to wibracje, które powstają z części proporcjonalnej. Zbyt mała wartość będzie niewystarczająca, a zbyt duża spowoduje próby poprawy przy najmniejszych odchyleń w przypadku nawet prostej trasy. PID jest stosowany w automatyce przemysłowej, napędach, kontrolach procesów i wielu innych dziedzinach. Programy pisane dla linefollowerów pomijają część całkującą, więc można powiedzieć, że korzystają z regulatora PD.

[16]https://www.integrasources.com/blog/basics-of-pid-controllers-design-applications/ (dostęp: 25.05.2024).
[17]https://physics.uwb.edu.pl/wf/fi-bot/?p=1534 (dostęp: 25.05.2024).

$$u(t) = K_p e(t) + K_i \int_0^t e(t)dt + K_d \frac{de(t)}{dt}$$

Rysunek 8. Wzór PID.
Źródło: https://plcynergy.com/pid-controller/ (dostęp: 25.05.2024).

Współczynniki KP i KD określają jaki wpływ mają poszczególne człony na ruch robota i będą się znacząco różnić w zależności od algorytmu wykrywania pozycji linii czy samej budowy robota (np. pozycji czujnika). Właśnie w konkretnych wartościach KP, KD leży sekret działania dobrego regulatora PID.[18]

7. Kod

Podczas tworzenia programu dla linefollowera, najważniejsze jest poprawna implementacja logiki i zasad, które pozwolą robotowi płynnie i bezbłędnie przejeżdżać trasę. Program wgrywany na mikrokontroler został napisany w języku C w środowisku Arduino IDE 2.3.2. Listing 1. Przedstawia implementacje programu dla robota.

```
void loop()
{
  // Odczyt pozycji czarnej linii
  int position = qtr.readLineBlack(sensorValues);

  // Zapamiętywanie ostatniej pozycji linii

if (sensorValues[0] >= 800 && sensorValues[NUM_SENSORS-1] <
800) {
    if (last_sighted != 1 && millis() - last_detection_time >=
100) {
      last_sighted = 1;
      last_detection_time = millis();
      pixels.setPixelColor(0, pixels.Color(0, 255, 0));
// Zielony dla lewej
      pixels.show();
    }
  } else if (sensorValues[0] < 800 && sensorValues[NUM_SENSORS-
1] >= 800) {
    if (last_sighted != 2 && millis() - last_detection_time >=
100) {
      last_sighted = 2;
      last_detection_time = millis();
      pixels.setPixelColor(0, pixels.Color(255, 0, 0));
// Czerwony dla prawej
      pixels.show();
    }
```

[18]http://asq.org/learn-about-quality/cost-of-quality/overview/overview.html (dostęp: 25.05.2024).

```
  }

  // Kontrola orientacji robota, ocena czy linia została
zgubiona

  lost_sensors = 0;
  lost = 0;
  for(int i = 0; i < NUM_SENSORS; i++){
    if(sensorValues[i] <= lost_threshold){
      lost_sensors += 1;
    }
  }
  if(lost_sensors >= (NUM_SENSORS)){
    lost = 1;
  }

 //PID

  int error = position - 3500;

  int motorSpeed = Kp * error + Kd * (error - lastError);
  lastError = error;

  rightMotorSpeed = BaseSpeed + motorSpeed;
  leftMotorSpeed = BaseSpeed - motorSpeed;

  if (rightMotorSpeed > MaxSpeed ) rightMotorSpeed = MaxSpeed;
  if (leftMotorSpeed > MaxSpeed ) leftMotorSpeed = MaxSpeed;
  if (rightMotorSpeed < 0) rightMotorSpeed = 0;
  if (leftMotorSpeed < 0) leftMotorSpeed = 0;


  // Algorytm jazdy

  if(ready == 1){
    if(lost == 1 && last_sighted == 1){
      analogWrite(RIGHT_MOTOR_FORWARD, TurnSpeed);
      analogWrite(LEFT_MOTOR_BACKWARD, TurnSpeed);
      analogWrite(LEFT_MOTOR_FORWARD, 0);
      analogWrite(RIGHT_MOTOR_BACKWARD, 0);
    }
    else if(lost == 1 && last_sighted == 2){
      analogWrite(RIGHT_MOTOR_FORWARD, 0);
      analogWrite(LEFT_MOTOR_BACKWARD, 0);
      delay(10);
      analogWrite(LEFT_MOTOR_FORWARD, TurnSpeed);
      analogWrite(RIGHT_MOTOR_BACKWARD, TurnSpeed);
    }
    else{
      analogWrite(RIGHT_MOTOR_BACKWARD, 0);
      analogWrite(LEFT_MOTOR_BACKWARD, 0);
```

```
      delay(10);
      analogWrite(LEFT_MOTOR_FORWARD, rightMotorSpeed);
      analogWrite(RIGHT_MOTOR_FORWARD, leftMotorSpeed);
    }
  }
  else{
    analogWrite(LEFT_MOTOR_FORWARD, 0);
    analogWrite(RIGHT_MOTOR_FORWARD, 0);
    analogWrite(LEFT_MOTOR_BACKWARD, 0);
    analogWrite(RIGHT_MOTOR_BACKWARD, 0);
  }

}
```

Listing 1. Algorytm poruszania się linefollowera.

Program w pierwszej kolejności odczytuje wartości czujników i zapisuje je w tablicy sensorValues. W zmiennej position zapisano liczbę odpowiadającą położeniu, gdzie robot wykrył linię. Każdy czujnik odczytuje inne wartości, które zależą od położenia toru. Im bliżej czujnika jest czarna linia, tym większą wartość pokaże. Zależność zmiennej position od ułożenia robota jest następująca:

- jeśli wartość wynosi 0, linia jest całkowicie po lewej stronie,
- jeśli wartość wynosi 7000, linia jest po prawej stronie,
- jeśli wartość wynosi 3500, linia jest na środku pod robotem.

Zapamiętywanie ostatniej linii pozwala robotowi na ponowne odnalezienie drogi w przypadku zgubienia się. Program porównuje wartość pierwszego i ostatniego czujnika, i za pomocą diody LED na mikrokontrolerze wyświetla kolor czerwony lub zielony w zależności od tego, po której stronie listwy pojawi się linia. Jeśli wartość sensorValues[0] jest większa lub równa 800, oznacza to, że z lewej strony wykryto linię. Gdy zmienna last_sighted nie jest ustawiona na wartość odpowiadającą pojawieniu się linii z lewej, zmienia się jej wartość na 1. To samo sprawdza się w przypadku, kiedy linia zostanie wykryta przez czujnik z prawej strony.

Zgubienie trasy może prowadzić do całkowitej dezorientacji robota. Zmienna lost przechowuje informacje, czy czujnik zgubił linię (wartość 1) lub jej nie zgubił (wartość 0). lost_sensors zlicza liczbę czujników, które nie wykrywają linii. Pętla przechodzi przez każdy czujnik, aby sprawdzić, który z nich wykrywa białe tło. Jedynie w przypadku, kiedy żaden czujnik nie może określić położenia linii, można stwierdzić, że robot się zgubił. Celem tego fragmentu kodu jest monitorowanie stanu linefollowera. Jeżeli każdy czujnik wskazuje wartość poniżej progu lost_threshold, oznacza to, że zgubiono trasę, i do flagi lost jest przypisana

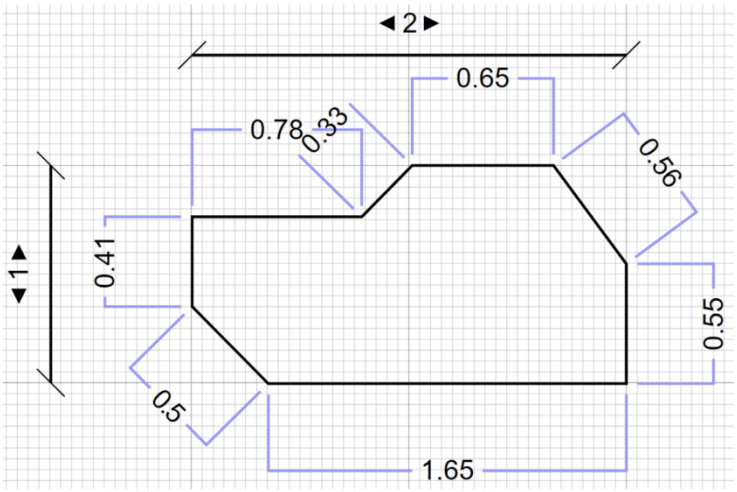
wartość 1. Robot następnie podejmuje działanie, czy należy poszukiwać linii, czy się zatrzymać.

Regulator PID (Proporcjonalno-Różniczkujący) kontroluje prędkość silników, wspomaga poprawne skręcanie i jest kluczowy przy utrzymaniu robota na ścieżce. Error to różnica pomiędzy pozycją środkową (3500) a aktualną pozycją, określa, jakie jest odchylenie robota. Jeśli position jest mniejsze niż 3500, robot jest po lewej stronie linii, a jeśli większe, to po prawej. Inicjalizacja algorytmu PD na zmiennej motorSpeed polega na zsumowaniu iloczynu błędu i wartości Kp oraz iloczynu Kd i różnicy błędu i zmiennej lastError. Cztery linie kodu ograniczają prędkość silników do wartości MaxSpeed, która określa, z jaką prędkością może maksymalnie poruszać się robot na prostej linii.

Ostatni fragment kodu kontroluje ruchem robota na podstawie stanu gotowości i widoczności linii. W przypadku zgubienia robot stara się odnaleźć linię, skręcając w stronę, w której ostatnio była widoczna. Kiedy czujniki wykrywają linię, linefollower jedzie prosto z prędkością obliczoną wcześniej. W przypadku braku gotowości robota, silniki się zatrzymują.

8. Testy

Robot został przetestowany na trasie o szerokości jednego metra i długości dwóch metrów. Testy były przeprowadzane w dobrze oświetlonym pomieszczeniu z użyciem światła sztucznego. Rysunek 9 przedstawia schemat trasy na której testowano robota.



Rysunek 9. Trasa.
Źródło: Opracowanie własne przy użyciu: https://www.smartdraw.com/floor-plan/architecture-software.htm (dostęp: 25.05.2024).

Robot, po przejechaniu trasy dziewięciokrotnie, był wyłączany w celu naładowania baterii. Napięcie na baterii nigdy nie spadło poniżej 6,1 V. Na początku wartość Kd została ustawiona na 0, aby dobrać odpowiednie wartości Kp. Dla każdej wartości Kp linefollower był trzykrotnie

testowany na trasie, zaczynając z tego samego punktu. Tabela 1 przedstawia tabelę wykonaną podczas testów dla robota.

Kp	Kd	Wartość Prędkości Bazowej	1	2	3		
10	0	80	x	31,350	x		31,350
5	0	80	14,842	15,193	13,589		14,541
3	0	80	12,618	11,441	25,024		16,361
2	0	80	12,942	21,024	13,701		15,889
1,7	0	80	12,811	11,608	11,739		12,053
1,5	0	80	12,513	52,935	11,049		25,499
1,3	0	80	18,115	22,105	x		20,110
1	0	80	23,121	12,182	10,882		15,395
0,9	0	80	15,285	11,005	13,924		13,405
0,7	0	80	11,643	12,261	10,070		11,325
0,5	0	80	9,339	11,332	12,683		11,118
0,4	0	80	13,725	15,632	14,952		14,770
0,3	0	80	14,035	13,770	14,275		14,027
0,2	0	80	14,919	15,753	15,924		15,532
0,1	0	80	14,731	14,942	14,325		14,666
						średnia wszystkich wyników	15,335

Rysunek 1. Tabela czasów przejazdu dla wyzerowanego Kd.
Źródło: Opracowanie własne.

Podczas testów zdarzało się, że robot całkowicie gubił trasę i nie był w stanie jej odnaleźć; najczęściej miało to miejsce przy Kp = 10. Niejednokrotnie robot wyjeżdżał poza trasę, ale po pewnym czasie wracał na nią, choć zaczynał z innego punktu. W takich przypadkach czas był mierzony do momentu, aż robot przejechał pełną trasę jednokrotnie. Bateria była ładowana czterokrotnie w trakcie testów. Obliczono wartości średniego czasu dla każdej wartości Kp oraz średnią dla wszystkich czasów. Średnie większe od średniej ogólnej oznaczono czerwonym kolorem, a wartości mniejsze - zielonym. Linefollower najszybciej pokonywał trasę przy Kp 0,5 oraz 0,7. Dla tych wartości testowano następnie wartości Kd. Tabela 2 przedstawia tabelę wyników dla wartości Kp równej 0,7.

Kp	Kd	Wartość Prędkości Bazowej	1	2	3		Średnia
0,7	20	80	26.482	17,937	11,917		14,927
0,7	15	80	14,390	15,131	11,229		13,583
0,7	13	80	11,392	8,630	13,830		11,284
0,7	10	80	11,782	11,482	11,294		11,519
0,7	9	80	11,592	10,362	8,821		10,258
0,7	8	80	8,590	9,241	9,268		9,033
0,7	7	80	10,103	9,759	8,659		9,507
0,7	6	80	8,934	9,370	8,260		8,855
0,7	5	80	10,092	8,927	9,238		9,419
0,7	4	80	12,482	12,012	13,118		12,537
0,7	3	80	14,928	14,382	13,482		14,264
0,7	1	80	17,482	15,392	15,382		16,085
0,7	0.7	80	14,281	15,382	13,382		14,348
0,7	0.5	80	13,927	14,827	18,398		15,717
0,7	0.1	80	14,291	13,193	15,382		14,289
						średnia wszystkich czasów	12,375

Tabela 2. Tabela czasów przejazdu dla pierwszej wartości Kp.
Źródło: Opracowanie własne.

Po opracowaniu wyników wyliczono średnią dla każdego czasu aby wyznaczyć które czasy dla danej wartości Kd są średnio najniższe. Czerwonym kolorem oznaczono wartości większe od średniej ogólnej, a wartości mniejsze oznaczono zielonym. Tabela 3 przedstawia tabelę wyników dla wartości Kp równej 0,5.

Wybór wartości Kp i Kd zależy od prędkości robota. Zaprezentowane wyniki mogą się różnić dla innych wartości prędkości bazowej. Określenie idealnych wartości dla robota jest niemożliwe, można jedynie stworzyć zestaw najkorzystniejszych Kp i Kd dla danej trasy. Dla prędkości 80 i Kp 0,7 najkorzystniejsze są wartości od 5 do 8 dla Kd. Natomiast dla prędkości 80 i Kp 0,5 są to wartości od 3 do 5 dla Kd.

Można więc stwierdzić, że istnieje pewna zależność pomiędzy Kp a Kd. Kd powinno być od 7 do 10 razy większe od wartości Kp, aby robot przebywał trasę w najkorzystniejszym czasie.

Kp	Kd	Wartość Prędkości Bazowej	1	2	3		Średnia
0,5	20	80	16,382	14,382	14,927		15,230
0,5	15	80	14,284	21,482	14,827		16,864
0,5	13	80	14,284	15,382	16,372		15,346
0,5	10	80	29,382	14,382	11,382		18,382
0,5	9	80	13,124	14,582	13,978		13,895
0,5	8	80	14,204	13,283	13,07		13,519
0,5	7	80	14,25	11,238	14,084		13,191
0,5	6	80	13,12	12,149	11,733		12,334
0,5	5	80	10,128	9,535	12,581		10,748
0,5	4	80	9,572	8,489	11,183		9,748
0,5	3	80	9,382	10,285	11,829		10,499
0,5	1	80	12,468	11,482	14,724		12,891
0,5	0.7	80	16,391	17,382	13,294		15,689
0,5	0.5	80	12,482	13,183	21,252		15,639
0,5	0.1	80	12,642	11,525	15,25		13,139
						średnia wszystkich czasów	13,808

Tabela 3. Tabela czasów przejazdu dla drugiej wartości Kp.
Źródło: Opracowanie własne.

9. Podsumowanie

Linefollower to bardzo popularny typ robota, chętnie budowany przez osoby o różnym stopniu zaawansowania i doświadczenia. Jest on obecny w wielu kategoriach na ogólnopolskich i międzynarodowych zawodach, zarówno w Polsce, jak i na całym świecie. Prostota tego robota tkwi w niewielkiej ilości wymaganych komponentów oraz przejrzystych zasadach jego działania.

Omówiono kompleksowo konstrukcję robota, w tym wybór użytego sprzętu, takiego jak mikrokontroler, czujniki, silniki, koła, moduł sterownika oraz moduł zasilacza obniżającego napięcie. Budowa wymagała jedynie podstawowych umiejętności manualnych, a program jest logiczny i nie wymaga wielu zabezpieczeń przed błędami jazdy. Kluczowym elementem było omówienie teorii oraz algorytmu PID, które są zaimplementowane w kodzie w sposób klarowny. Dzięki temu robot dobiera wartości prędkości dla silników w zależności od położenia na trasie.

Przeprowadzono testy w celu oceny wydajności robota. Wyniki wykazały, że robot działa zgodnie z oczekiwaniami i jest w stanie samodzielnie przejechać trasę bez pomocy z zewnątrz, zachowując precyzję i dobrą prędkość. Wyprowadzono wniosek z testów, który stwierdził, że wartość Kd powinna być 7-10 razy większa od wartości Kp. Linefollower stanowi skuteczne i

edukacyjne narzędzie, które może być wykorzystywane nie tylko w zawodach, ale również w celach edukacyjnych. Zaprezentowane wyniki mogą posłużyć jako podstawa do dalszych badań i ulepszania robota.

Literatura

1. Martin E., Joshua N., Jordan H., *Arduino w akcji*, Helion S.A.,2014.
2. . Michael M., Brian J., Nicholas Robert W., *Arduino. Przepisy na rozpoczęcie, rozszerzanie i udoskonalanie projektów*, Helion S.A., Wydanie III, 2021.
3. Udo Brandes, *Mikrocontroller ESP32,* Rheinwerk Verlag GmbH, marzec 2023.
4. Dogan Ibrahim, *Practical Audio DSP Projects with the ESP32*, elektor, Elektor Verlag; Edycja Main, 2023).
5. Agus Kurniawan, *Internet of Things Projects with ESP32*, Packt Publishing, 2019.
6. Simon Monk, *Zabawy z elektroniką. Ilustrowany przewodnik dla wynalazców i pasjonatów,* Helion S. A. 2014.
7. Dahl Nydal Oyvind, *Elektronika dla małych i dużych. Od przewodu do obwodu,* Helion S.A., 2022.
8. Paul S., Simon M., *Practical Electronics for Inventors*, McGraw-Hill Education, Wydanie IV, 2016.
9. Witold Wrotek, *Elektronika bez oporu. Praktyczne przykłady,* Helion S.A., 2022

Źródła internetowe

1. http://asq.org/learn-about-quality/cost-of-quality/overview/overview.html(dostęp: 19.05.2016).
2. https://botland.com.pl/czujniki-odbiciowe/20-listwa-z-czujnikami-odbiciowymi-qtr-8rc-cyfrowa-pololu-961-5903351249287.html (dostęp: 25.05.2024).
3. https://botland.com.pl/moduly-wifi-i-bt-esp32/20739-esp32-s3-devkitc-1-n8-wifi-bluetooth-plytka-rozwojowa-z-ukladem-esp32-s3-wroom-11u-5904422382353.html?cd=1050025856&ad=51004438223&kd=&gad_source=1&gclid=Cj0KCQjwjLGyBhCYARIsAPqTz19bkTQ6sop-6Y3ugUdsIVEvCBs1hrGb6KdkcGNrm8p2TkwKad-JCNUaAmWWEALw_wcB (dostęp: 25.05.2024).

4. https://sklep.avt.pl/pl/products/modul-sterownika-mini-mx1508-do-silnikow-dc-podwojny-arduino-187397.html (dostęp: 25.05.2024).
5. https://rif.sklep.pl/az/regulatory-napiecia/655-mini-przetwornica-step-down-08-21-3a-975.html(dostęp: 25.05.2024).
6. https://abc-rc.pl/product-pol-9257-Mini-silnik-szczotkowy-GA12-N20-50RPM-wal-9mm-3-6V-z-przekladnic cva.html (dostęp: 25.05.2024).
7. https://botland.com.pl/kola-z-oponami/147-kolo-solarbotics-rw2-mocowanie-zewnetrzne-pololu-642-5903351248150.html (dostęp: 25.05.2024).
8. https://botland.com.pl/akumulatory-li-pol-2s-74v-/570-pakiet-li-pol-dualsky-520mah-25c-2s-74v-6941047107427.html?cd=19993067448&ad=&kd=&gad_source=1&gclid=Cj0KCQjwjLGyBhCYARIsAPqTz18ktWWt5vxUDfN3VAo-B0eI9fq4_dvstwlujnqk-5hzmR1qZpwnHggaAn-CEALw_wcB (dostęp: 25.05.2024).
9. https://elektroweb.pl/esp32/1312-modul-esp32-s3-devkitc-1-wroom-1-n16r8-16mb-flash-wifi-bluetooth-usb-c-5905523309461.html (dostęp: 25.05.2024).
10. https://elektroweb.pl/esp32/1312-modul-esp32-s3-devkitc-1-wroom-1-n16r8-16mb-flash-wifi-bluetooth-usb-c-5905523309461.html (dostęp: 25.05.2024).
11. https://botland.com.pl/czujniki-odbiciowe/20-listwa-z-czujnikami-odbiciowymi-qtr-8rc-cyfrowa-pololu-961-5903351249287.html (dostęp: 25.05.2024).
12. https://abc-rc.pl/product-pol-17960-Modul-sterownika-Mini-MX1508-do-silnikow-DC-podwojny-Arduino.html (dostęp: 25.05.2024).
13. https://zpe.gov.pl/a/przeczytaj/DUpr2MS6R (dostęp: 25.05.2024).
14. https://botland.com.pl/blog/sygnal-pwm-czym-jest/ (dostęp: 25.05.2024).
15. https://www.instructables.com/SIMPLE-LINE-FOLLOWER-ROBOTsingle-Sensor/ (dostęp: 25.05.2024).
16. https://mikrokontrolery.blogspot.com/2011/03/Robotyka-Linefollower-Listwa-czujnikow-linii.html (dostęp: 25.05.2024).
17. https://www.integrasources.com/blog/basics-of-pid-controllers-design-applications/ (dostęp: 25.05.2024).
18. https://physics.uwb.edu.pl/wf/fi-bot/?p=1534 (dostęp: 25.05.2024).
19. http://asq.org/learn-about-quality/cost-of-quality/overview/overview.html(dostęp: 25.05.2024).

Sławomir Pareniak, Katarzyna Maternia, Jakub Bocek, Patryk Krupa, Piotr Dubaj
Koło naukowe Elektroniki i Technologii Informacyjnych

dr inż. Bartosz Pawłowicz
Opiekun Koła Naukowego

Wpływ sztucznej inteligencji na mobilne doświadczenia użytkowników: nowe możliwości i wyzwania

Streszczenie

Artykuł analizuje znaczenie sztucznej inteligencji w kontekście urządzeń mobilnych, podkreślając, że zaawansowane algorytmy i systemy uczenia maszynowego rewolucjonizują doświadczenia użytkowników. Poprzez personalizację dostosowaną do indywidualnych potrzeb i preferencji, smartfony stają się bardziej intuicyjne i skuteczne w obsłudze. Automatyzacja zadań, wspierana przez sztuczną inteligencję, umożliwia urządzeniom mobilnym wykonywanie czynności zgodnie z preferencjami użytkownika, co zwiększa wygodę użytkowania.

Rozpoznawanie mowy jest kolejnym kluczowym aspektem omawianym w artykule, gdzie zaawansowane systemy przetwarzania języka naturalnego pozwalają na interakcję z urządzeniami bez konieczności korzystania z klawiatury czy ekranu dotykowego. Ważnym aspektem w naszej przyszłości jest rola sztucznej inteligencji w bezpieczeństwie samochodów oraz w sprzęcie AGD, przynosząc zalety takie jak optymalizacja zużycia energii i zapewnienie większego komfortu użytkowania.

Słowa kluczowe: inteligentny dom, cyberataki, sieci energetyczne, urządzenia mobilne.

1. Wprowadzenie

Wraz z rosnącą cyfryzacją, cyberprzestępczość staje się coraz bardziej złożona i powszechna. Wprowadzenie nowych technologii otwiera drzwi dla różnorodnych cyberataków, wymagając coraz bardziej zaawansowanych strategii bezpieczeństwa cybernetycznego.

W dzisiejszych czasach coraz częściej słyszy się określenie „inteligentny dom” lub z języka angielskiego „smart home”, jednak osoby, które takiego lokum nie posiadają, mogą nie wiedzieć co ta nazwa oznacza. Najprościej rzecz ujmując, jest to wysoko zaawansowany technicznie budynek, który pełni mnóstwo funkcji ułatwiających życie mieszkającym w nim ludziom. Warto jednak poznać dokładne znaczenie tego określania, by wiedzieć czym jest inteligentny dom i jakie korzyści płyną z mieszkania w nim.

Inteligentne sieci energetyczne innymi słowy (Smart grid) umożliwiają monitorowanie oraz zarządzanie dostępnością i wydajnością źródeł energii odnawialnej. Dzięki temu innowacyjnemu rozwiązaniu, możliwe jest efektywne równoważenie wahań w produkcji energii odnawialnej, zależnych od warunków atmosferycznych, co pozwala na lepsze wykorzystanie zasobów oraz minimalizuje to ryzyko nieefektowności sieci. Inteligentne

sterowanie przepływem energii oraz optymalizacja obciążeń sieciowych sprawia że zarządzanie przesyłek energii staje się o wiele sprawniejsze.

2. Co to są inteligentny dom?

Nie istnieje jednoznaczna definicja inteligentnego domu, jednak wszystkie informacje, jakie można znaleźć na jego temat sprowadzają się do jednego stwierdzenia – jest do budynek (dom, blok mieszkalny lub lokal użytkowy, np. hotel, biurowiec), który wyposażony jest w system integrujący wszystkie znajdujące się w budynku instalacje.

System inteligentnego domu jest wyposażony w szereg czujników i detektorów, których działanie jest nadzorowane przez specjalne oprogramowanie. Nowoczesna technologia tworząca BMS ułatwia domownikom korzystanie z wielu funkcji w pomieszczeniach i na zewnątrz, a przede wszystkim zapewnia im bezpieczeństwo. Choć oprogramowanie systemu jest bardzo skomplikowane, to korzystanie z niego jest proste i intuicyjne, a co najważniejsze – mobilne, gdyż pozwala zarządzać domem z każdego miejsca na świecie z dostępem do Internetu (np. za pomocą smartfonu).

Inteligentny dom jest tak nazywany nie tylko ze względu na mnogość przydatnych funkcji, którymi łatwo jest zarządzać i zapewnienie bezpieczeństwa, ale również dlatego, że często „myśli" za domowników. Choć może wydawać się to nieco abstrakcyjne, to jednak faktem jest, że inteligentny dom dba o finanse właścicieli poprzez energooszczędność w użytkowaniu i eksploatacji budynku.

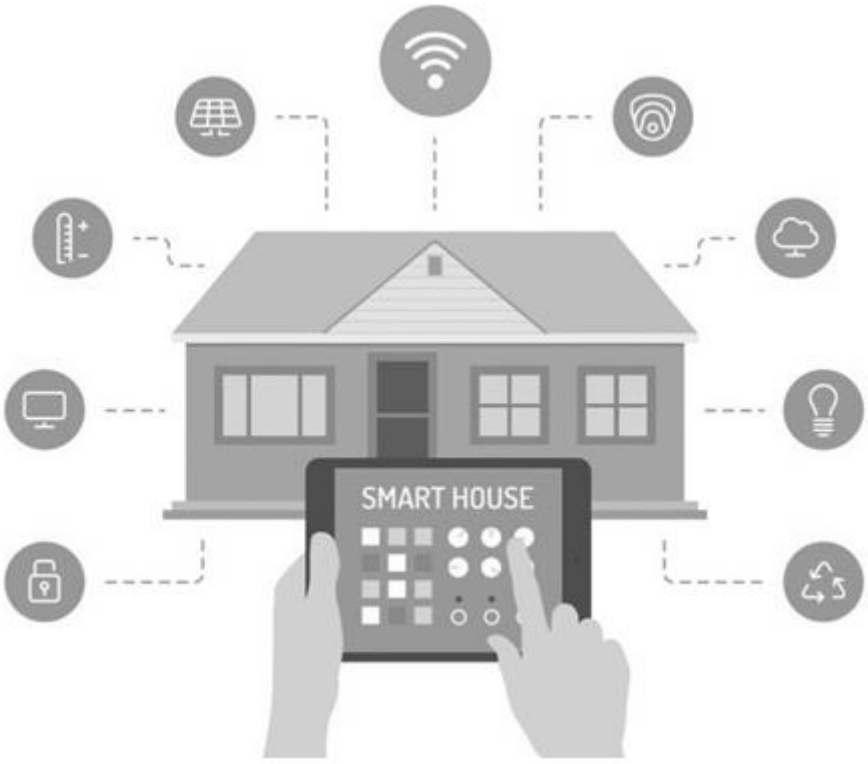
3. Jakie są produkty inteligentnego domu?

Na inteligentny dom mogą składać się dziesiątki różnych produktów, dzielących się na: centralkę oraz komponenty sterujące (do kontrolowania urządzeń) i komponenty funkcjonalne (zapewniające funkcjonalność). Te ostatnie można też podzielić na: multimedialne, klimatyczne i zabezpieczające.

Podstawowe produkty inteligentnego domu:

- **czujniki inteligentne** (wykrywające np. ruch, ulatniający się gaz, zalanie lub pożar itp.),
- inteligentne oświetlenie i ogrzewanie,
- **systemy alarmowe** (na czele z syrenami) i kamery do monitoringu,
- inteligentne gniazdka (umożliwiające zdalne włączanie i wyłączanie urządzeń oraz monitorowanie zużycia energii),
- **panele, przyciski i piloty IR** (służące do sterowania systemem),

- **centralka** (integrująca ze sobą wszystkie te elementy),
- **urządzenie sterujące** – tablet, smartfon lub urządzenie dedykowane dla danego sytemu, za pomocą którego można zarządzać funkcjami domu.



Rysunek 13 Obsługa inteligentnego domu za pomocą urządzenia mobilnego; źródło: https://cdn.benchmark.pl/uploads/backend_img/c/recenzje/2021_05/inteligentny-dom-1.jpg

Jakie funkcje mogą się znaleźć w inteligentnym domu?

- sterowanie klimatyzacją i wentylacją,
- symulacja obecności domowników,
- system alarmowy,
- sterowanie ogrzewaniem,
- system przeciwpożarowy,
- zdalne sterowanie,
- czujniki pogodowe,
- personalizacja.

4. Ile kosztują inteligentne domy?

Zaawansowana technologia sprawia wrażenie bardzo kosztownej, jednak wbrew pozorom ceny za domy czy mieszkania wyposażone w system BMS nie są dużo wyższe niż nieruchomości nie posiadające takich udogodnień. W niektórych miastach (np. Kraków,

Wrocław) inteligentne mieszkania są w tej samej cenie, co standardowe lokale mieszkalne z rynku pierwotnego. Zmiana domu ze „zwykłego" na inteligentny do wydatek rzędu 20-30 tysięcy złotych (dla nieruchomości około 100 lub 120-metrowej).

Inteligentne domy i mieszkania są niewątpliwie o wiele bardziej bezpieczne i samowystarczalne, jednak nie oznacza to, że potrafią same się uchronić przed wszystkimi zdarzeniami losowymi. Można wprawdzie być spokojnym, że nikt się nie włamie do smart home, a pożar praktycznie sam się ugasi, jednak szkody, jakie powstaną na skutek tych sytuacji nie naprawią się same. Dlatego każda nieruchomość – zarówno inteligentna, jak i tradycyjna powinna być ubezpieczona. Ochrona ubezpieczeniowa zapewni wypłatę środków pokrywających straty powstałe na skutek wielu różnych zdarzeń losowych. Szeroki wybór mieszkaniowych polis ubezpieczeniowych znajduje się w ofercie firmy Compero, która od wielu pomaga swoim klientom wybrać ubezpieczenia najbardziej dostosowane do ich potrzeb.

5. Wyzwania stawiane inteligentnym domom

Chociaż istnieje wiele korzyści związanych z ideą smart house, ważne jest, aby uwzględnić również kilka możliwych wyzwań. Po pierwsze: należy liczyć się ze zwiększeniem kosztu budowy, bo zainstalowanie inteligentnego systemu oznacza rzecz jasna dodatkowe wydatki. Po drugie: architekci i kierownicy budów potrzebują wiedzy lub wsparcia ekspertów w zakresie projektowania i wznoszenia inteligentnych domów oraz zastosowanych w nich urządzeń i oprogramowania. Kolejne dwa aspekty to kwestie dotyczące integracji oraz bezpieczeństwa. Korzystanie z inteligentnej technologii zakłada, że powinny być odporne na włamania hakerów, a także kontrolowane przez jeden wspólny system, który dodatkowo będzie potencjalnie otwarty na rozbudowę i przyłączenie nowych urządzeń.

6. Jakie są rodzaje cyberataków

W miarę upływu czasu, świat coraz bardziej jest uzależniony i połączony z technologią. Z każdym dniem, ludzie coraz częściej żyją w Internecie, przez to tworzymy więcej możliwości dla cyberprzestępców, których metody stają się coraz bardziej udoskonalane.

Inżynieria społeczna jest to praktyka manipulowania ludźmi w ujawnianiu wrażliwych i poufnych informacji dla zysku pieniężnego a nawet dostępu do prywatnych danych. Występują też ataki złośliwego oprogramowania, takie jak wirusy, robaki czy oprogramowanie szpiegujące, przez co komputery mogą zostać zarażone tymi niechcianymi atakami. Ransomware jest to znane złośliwe oprogramowanie, które uzyskuje dostęp do plików i blokuje systemy w celu wymuszenia płatności okupu. Obecnie na świecie jest większa liczba urządzeń

IoT niż liczba ludzi na świecie, którzy mają wiele możliwości dla hakerów, ponieważ te urządzenia są podatne na ataki typu manin the medium czy denial of service, złośliwego oprogramowania.

APT to zaawansowane trwałe zagrożenia, które są wieloetapowymi atakami, podczas których hakerzy mają wpływ na sieć niewykrytą i pozostają w niej przez dłuższy czas w celu uzyskania dostępu do danych wrażliwych i prywatnych lub zakłóceń usług o kluczowym znaczeniu. APT są często skierowane do branży posiadającej cenne informacje dla świata jak obrona narodowa, finanse i produkcje.

Rysunek 2 Cyberbezpieczeństwo; źródło: https://executivemagazine.pl/wp-content/uploads/2023/07/internet-security-system-768x531.jpg

7. Jak działa bezpieczeństwo cybernetyczne?

Nie istnieje jedno uniwersalne rozwiązanie w zakresie bezpieczeństwa cybernetycznego dla przedsiębiorstw. Natomiast wiele warstw ochrony współpracuje w celu zabezpieczenia przed zakłóceniami procesów i ważnym dostępem do informacji, za ich zmianę czy zniszczenie i przechowywanie na okup. Ochrona ta jest zmuszona do ciągłego ewoluowania w celu praktycznego przeciwdziałania pojawiających się zagrożeń cybernetycznych. To wszystko jest po to stworzyć jednolitą obronę przed potencjalnymi cyberatakami.

Bezpieczeństwo związane z aplikacjami koncentruje się na bezpieczeństwie by je zwiększyć, gdy aplikacje są na etapie projektowania i po ich wdrożeniu. Typy zabezpieczeń aplikacji obejmują programy antywirusowe, programy szyfrowania i zapory sieciowe.

Trwająca migracja do chmury publicznej, prywatnej i hybrydowej oznacza, że dostawcy muszą ciągle priorytetowo traktować wdrażanie solidnych, aktualnych zabezpieczeń chmury w celu ochrony systemów, dostępności i danych.

Ciągły wzrost popularności Internetu skutkuje wzrost ryzyka. Podczas gdy bezpieczeństwo IoT różni się w zależności od urządzenia i jego zastosowania. Najlepsze praktyki w zakresie IoT bezpieczeństwa to zabezpieczenie urządzeń, które zapewniają bezpieczne aktualizacje i integracje oraz ochrona przed złośliwych oprogramowaniem.

Połączenie rozwiązań sprzętowych i programowych chroniących przed nieautoryzowanym dostępem do sieci, które mogą skutkować przechwytywaniem, zmianą czy kradzieżą informacji to bezpieczeństwo sieci. Tryby jakie obejmują zabezpieczenia sieci to loginy, hasła i zabezpieczenia aplikacji.

Bezpieczeństwo punktu końcowego w tym komputery stacjonarne, laptopy są punktami dostępu dla zagrożeń. Bezpieczeństwo punktu końcowego obejmuję dużą ochronę przed wirusami i złośliwym oprogramowaniem.

8. Działanie inteligentnej sieci energetycznej

Działanie sieci opiera się na kilku elementach. Pierwszym jest system zaawansowanych urządzeń pomiarowych. Umożliwia on dokładne monitorowanie zużycia energii w czasie rzeczywistym, a tym samym precyzyjne analizy oraz szybką reakcje na zmiany w zapotrzebowaniu oraz identyfikację obszarów, w których można osiągnąć spore oszczędności.

Kolejnym elementem działania inteligentnej sieci energetycznej jest jej automatyzacja oraz sterowanie. Dzięki nim można dostosować produkcję i dystrybucję energii w zależności od bieżących warunków i zapotrzebowania. Podczas szczytowego obciążenia, system może automatycznie przekierować energię do obszarów o największym zapotrzebowaniu, co prowadzi do minimalizacji strat i zapewnienia stabilność dostaw.

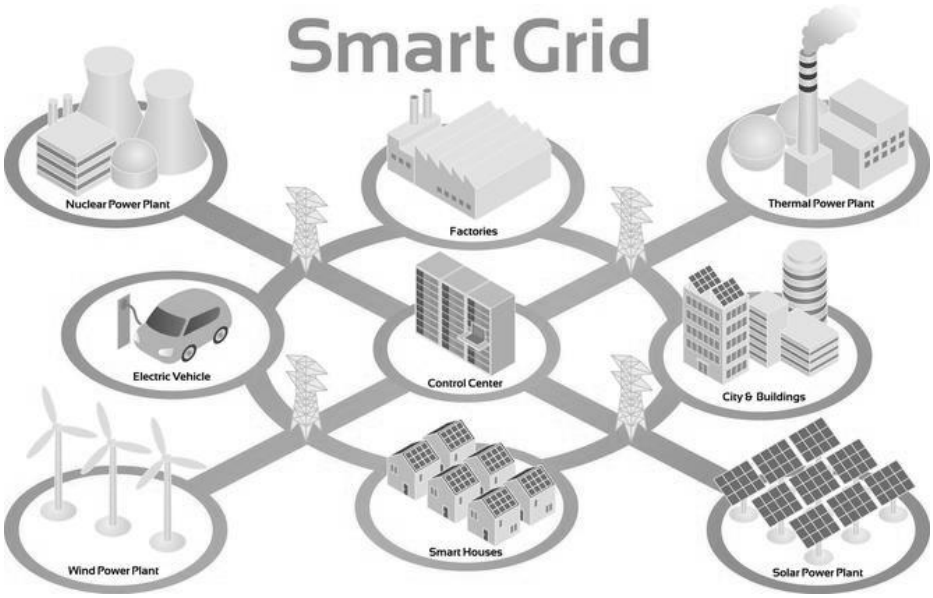
W inteligentnych sieciach energetycznych kluczową rolę odgrywa integracja odnawialnych źródeł energii. Takie systemy umożliwiają efektywną kooperację między różnymi źródłami energii, jak energia słoneczna czy wiatrowa oraz tradycyjnymi źródłami prądu. Dzięki temu można wykorzystać energię ze źródeł odnawialnych w sposób bardziej efektywny i stabilny, zmniejszając wahania w produkcji energii i zwiększając udział energii odnawialnej w całkowitej strukturze energetycznej.

9. Jak zbudowane są inteligentne sieci energetyczne

Skuteczne zarządzanie produkcją i dystrybucją energii wymaga wykorzystania infrastruktury, która umożliwi przetwarzanie nadwyżek. Do tego celu zaliczają się magazyny energii i elektrownie szczytowo-pompowe. Istotnym zadaniem jest również połączenie poszczególnych jednostek wytwórczych w klastry, mikro sieci lub spółdzielnie. To możliwe dzięki wirtualnej infrastrukturze, która obejmuje wirtualne elektrownie (VPP - Virtual Power Plant), wirtualne magazyny energii (VES - Virtual Energy Storage) oraz wirtualne linie elektroenergetyczne (VPL - Virtual Power Lines).

Sprawne działanie inteligentnych sieci wiąże się z użyciem dedykowanych urządzeń a najważniejszymi elementami infrastruktury smart grid są:

- transformatory z przekładnią energoelektroniczną i czujnikami diagnozującymi
- automatyczne wyłączniki służące do przywracania zasilania w przypadku zwarcia
- Elektroniczne liczniki energii elektrycznej umożliwiające zdalną rekonfigurację, odczyt oraz raportowanie
- Urządzenia energetyczno-elektroniczne poprawiające przepustowość systemu FACTS lub FACDS
- Linie przesyłowe prądu stałego
- Bezzałogowe statki powietrzne do monitorowania obiektów i działań niebezpiecznych;
- Systemy automatyki zabezpieczeniowej, restytucyjnej oraz prewencyjnej.



Rysunek 14 Inteligentna sieć elektroenergetyczna źródło: https://cdn.galleries.smcloud.net/t/galleries/gf-9X2x-Y7gB-ikSB_inteligentna-siec-elektroenergetyczna-664x442-nocrop.jpg

10. Sztuczna inteligencja w elektronice

Moduły sztucznej inteligencji stają się powszechnym elementem w urządzeniach elektronicznych poprzez wygodę użytkownika. Urządzenie podobne jest do tradycyjnych procesorów, jednak musi prezentować wysoką moc obliczeniową, przystosowane są one bowiem są do implementacji algorytmów ucznia maszynowego, logiki rozmytej lub sieci neuronowych. Moduły sztucznej inteligencji są coraz bardziej popularne w sprzętach AGD, wyposażone są one w inteligentne algorytmy i są zdolne do wymiany danych z otoczeniem.

Przykładami sztucznej inteligencji w sprzętach AGD jest pralka. Posiada ona usprawnioną możliwość płynnego sterowania pracą silnika, pralka w oparci o uczenie maszynowe wybierze program działania tak oby chroniła materiał naszego ubioru. Moduł sztucznej inteligencji posiada informacje na temat prania, jest to spora dawka danych wprowadzona tak, aby silnik urządzenia pracował optymalnie od konkretnego typu materiału znajdującego się w bębnie. Atutem sprzętów z sztuczna inteligencją jest nie tylko to że pralka sama dostosuje prace silnika ale też będzie wykonywała swoją prace lepiej za każdym razem, będzie dążyła do zmniejszenia zużycia wody jak i energii.

Firmy zajmujące się sprzętami AGD wprowadzają moduły wykrywające swoją usterkę. Firma LG w swoich sprzętach używa program do automatycznej diagnostyki i pomocy użytkownikowi za pomocą sztucznej inteligencji. Urządzenie zwane SmartThinQ powiadomi użytkownika o problemach ze sprzętem oraz powiadomi użytkownika sugerując odpowiednie rozwiązanie problemu. Moduł ten potrafi zaoszczędzić czas, czasami problem może być znikomy, a czas oczekiwania na specjalistę może być nieporównywanie dłuższy niż naprawa urządzenia przez nas samych.

Zalety używania sztucznej inteligencji w sprzętach AGD:

- Wygodniejsza obsługa sprzętu,
- Optymalizacja procesu z uwzględnieniem zużycia energii i wody,
- Brak konieczności wzywania specjalistów do mniejszych usterek.

11. Sztuczna inteligencja w samochodach

Sztuczna inteligencja zrewolucjonizowała rynek samochodów, nie tylko samochody elektryczne ale również spalinowa posiadają w swoich komponentach moduły sztucznej inteligencji. Zaczynając od pomocy sztucznej inteligencji przy projektowaniu samochodów, przy pomocy AI firmy projektują modele samochodów znacznie to przyspiesza proces

produkcji modelu. Samochody zaprojektowane z pomocą sztucznej inteligencji mają lepsza aerodynamikę.

Sztuczna inteligencja działa również w systemach bezpieczeństwa. Bardzo ważnym aspektem podczas jazdy samochodem jest bezpieczeństwo, sztuczna inteligencja ma zminimalizować ryzyko wypadku dlatego w współczesnych pojazdach są zastosowane różne systemy działające na podstawie sztucznej inteligencji, takie jak:

- Asystent hamowania (BAS, BA, MBA, ISA) - pomagający hamować w nagłych wypadkach lub zmniejszać prędkość podczas zbyt dynamicznego zbliżania się od obiektu,
- Elektroniczny system stabilizacji toru jazdy (ESP) - podtrzymuje tor jazdy w przypadku zagapienia się kierowcy,
- System rozpoznawania znaków drogowych (TSR) – wyłapuje znaki i wyświetla je na panelu użytkowania,
- System wykrywania zmęczenia kierowcy (Driver Alert),
- System kontroli ciśnienia w oponach (TPMS) – system wykryje i powiadomi nas o spadku ciśnienia,
- System automatycznego parkowania (Park Assist) – samochód automatycznie zaparkuje samochód na wybrane wolne miejsce parkingowe.

Systemy te oparte są na sztucznej inteligencji, samochód wyposażony jest w spora ilość kamer i czujników ruchu, które są niezbędne do wykonywania tych wszystkich czynności.

12. Sztuczna inteligencja w urządzeniach mobilnych

Sztuczna inteligencja (SI) stanowi kluczowy element współczesnych smartfonów, tabletów oraz wszelakich innych urządzeń elektronicznych, znacząco wpływając na doświadczenia użytkownika na wielu płaszczyznach. Dzięki zaawansowanym algorytmom i systemom uczenia maszynowego, urządzenia mobilne nie tylko stają się bardziej intuicyjne, ale także personalizują się, dostosowując się do indywidualnych potrzeb użytkowników i oferując unikalne rozwiązania.

Jednym z najbardziej widocznych aspektów wpływu sztucznej inteligencji na urządzenia mobilne jest personalizacja. Analizując dane użytkowników, SI potrafi zrozumieć ich nawyki, preferencje oraz potrzeby, co pozwala na dostarczanie spersonalizowanej zawartości i sugestii. Na przykład, algorytmy rekomendujące treści w mediach społecznościowych czy platformach streamingowych wykorzystują sztuczną inteligencję do zrozumienia gustów użytkowników i proponowania im treści odpowiadających ich zainteresowaniom.

Kolejnym istotnym obszarem jest rozpoznawanie mowy, które umożliwia interakcję z urządzeniem bez konieczności użycia klawiatury czy ekranu dotykowego. Dzięki zaawansowanym systemom przetwarzania języka naturalnego, smartfony mogą rozumieć polecenia użytkowników, wykonywać zadania oraz odpowiadać na pytania. To sprawia, że korzystanie z urządzenia staje się bardziej intuicyjne i efektywne, zwłaszcza w sytuacjach, gdy użytkownik jest zajęty lub ma zajęte obie ręce.

Automatyzacja zadań to kolejna zaleta sztucznej inteligencji w smartfonach. Dzięki możliwości uczenia się preferencji użytkowników, urządzenia mobilne mają możliwość automatycznie wykonywać określone czynności zgodniez ustalonymi warunkami. Na przykład, smartfony mogą automatycznie dostosowywać jasność ekranu w zależności od warunków oświetleniowych, planować trasę podróży z uwzględnieniem ruchu drogowego czy nawet regulować ustawienia klimatyzacji w samochodzie na podstawie lokalizacji i preferencji użytkownika.

Wreszcie, sztuczna inteligencja odgrywa istotną rolę w rozwoju funkcji zdrowotnych i monitoringu na przykład w smartfonach. Dzięki zaawansowanym algorytmom analizującym dane z czujników i różnych sensorów, urządzenia mobilne mogą śledzić aktywność fizyczną, monitorować parametry zdrowotne czy nawet wspomagać w diagnozowaniu chorób. Wykorzystując dane związane z pulsometrem, akcelerometrem czy nawet elektronicznym zegarkiem do pomiaru snu, smartfony mogą dostarczać użytkownikom istotnych informacji na temat ich zdrowia i aktywności fizycznej.

W podsumowaniu chciałbym wspomnieć, że obecność sztucznej inteligencji w urządzeniach mobilnych zmienia doświadczenia użytkownika, czyniąc je bardziej personalizowanymi, intuicyjnymi i efektywnymi. Dzięki wykorzystaniu zaawansowanych algorytmów i systemów uczenia maszynowego, smartfony stają się nie tylko narzędziami komunikacji, ale także inteligentnymi asystentami, które aktywnie wspierają użytkowników w codziennych zadaniach. W miarę rozwoju technologii SI możemy spodziewać się dalszego zwiększania się roli sztucznej inteligencji w kształtowaniu przyszłości mobilnych doświadczeń użytkowników. Myślę, że jest to przyszłość do której zmierza rozwój techniczny ludzkiej cywilizacji w najbliższej przyszłości.

13. Podsumowanie

Artykuł podkreśla, że obecność sztucznej inteligencji w mobilnych urządzeniach elektronicznych stanowi istotny krok w rozwoju technologicznym, poprawiając wydajność, bezpieczeństwo i komfort użytkowania. W miarę postępu technologicznego, oczekuje się

dalszego wzrostu roli sztucznej inteligencji w kształtowaniu przyszłości mobilnych doświadczeń użytkowników, otwierając nowe możliwości i stawiając przed nami nowe wyzwania.

Źródła internetowe:

1. Co to jest inteligentny dom? Co się w nim znajduje? - poradnik Compero.pl (dostęp: 28.04.2024).
2. https://android.com.pl/tech/709393-sztuczna-inteligencja-zmienia-smartfony/(dostęp: 28.04.2024).
3. https://cdn.benchmark.pl/uploads/backend_img/c/recenzje/2021_05/inteligentny-dom-1.jpg (dostęp: 28.04.2024).
4. https://gsm24.pl/content/201-sztuczna-inteligencja-w-smartfonach-jakie-funkcje-ai-znajdziemy-w-telefonach (dostęp: 28.04.2024).
5. https://elektromobilni.pl/sztuczna-inteligencja-w-motoryzacji-czyli-przyszlosc-ktora-jest/ (dostęp: 28.04.2024).
6. https://ep.com.pl/rynek/temat-miesiaca/14858-sztuczna-inteligencja-w-praktycznej-elektronice (dostęp: 28.04.2024).
7. https://escola.pl/uslugi/ai-sztuczna-inteligencja/ (dostęp: 28.04.2024).
8. https://lajtmobile.pl/blog/nowinki/sztuczna-inteligencja-w-smartfonie-jakie-sa-perspektywy-rozwoju-i-zastosowan/ (dostęp: 28.04.2024).
9. https://onmedia.com.pl/czym-sa-tzw-inteligentne-sieci-energetyczne-i-jakie-korzysci-wynikaja-z-ich-wprowadzenia/ (dostęp: 28.04.2024).
10. https://www.komputerswiat.pl/artykuly/partnerskie/czy-twoja-pralka-posiada-sztuczna-inteligencje-sprawdz-jak-najnowsze-urzadzenia/5hs4sgm (dostęp: 28.04.2024).
11. https://www.neonet.pl/blog/inteligentna-lodowka-co-to.html (dostęp: 28.04.2024).
12. https://www.sap.com/poland/products/financial-management/what-is-cybersecurity.html (dostęp: 28.04.2024).
13. Inteligentne domy: aktualna sytuacja i prognozy (planradar.com) (dostęp: 28.04.2024).
14. Rozwiązania Smart Home, czyli inteligentny dom - pytania i odpowiedzi (komputronik.pl) (dostęp: 28.04.2024).

Patryk Krupa, Katarzyna Maternia, Sławomir Pareniak, Jakub Bocek, Piotr Dubaj
Koło naukowe Elektroniki i Technologii Informacyjnych

dr inż. Bartosz Pawłowicz
Opiekun Koła Naukowego

Inteligentne systemy sterowania ruchem

Streszczenie

Nasz artykuł omawia rozwiązania technologiczne zastosowane w nowoczesnych systemach automatyki, sterujących ruchem pojazdów oraz prezentuje najważniejsze skutki oraz korzyści ich wykorzystania dla współczesnych i przyszłych użytkowników; co może doprowadzić do wyeliminowania negatywnych konsekwencji związanych z szeroko pojętym transportem. W tekście wspomniane jest o prostych sposobach optymalizowania działania tych systemów oraz perspektywami z tym związanymi na przyszłość. Znajdzie się również informacja na temat najnowszych osiągnięć w dziedzinie inteligentnych systemów zarządzania ruchem, zwracając uwagę na ich zastosowania, korzyści oraz wyzwania. Omówimy również technologie kluczowe dla ITS, takie jak systemy detekcji, systemy komunikacji pojazdów do infrastruktury (V2I) oraz pozostałych systemów komunikacji, a także trendy przyszłościowe, które kształtują rozwój tego obszaru.

Słowa kluczowe: ruch, zarządzanie, sterowanie, ITS, inteligentny, transport.

1. Wprowadzenie

XX wiek był niewątpliwie wiekiem wielkiego boom jeżeli chodzi o rozwój wszelkich technologii m.in. dziedziny którą omówimy poniżej. Zatem z powodu rosnącej urbanizacji i wzrostu liczby pojazdów na drogach, efektywne zarządzanie ruchem staje się kluczowym wyzwaniem dla miast i regionów na całym świecie. W związku z tym powstają inteligentne systemy zarządzania ruchem, które wykorzystują zaawansowane technologie informatyczne i komunikacyjne do optymalizacji przepływu pojazdów, poprawy bezpieczeństwa drogowego oraz redukcji negatywnego wpływu transportu na środowisko.

Inteligentne systemy zarządzania ruchem to kompleksowe rozwiązania, które integrują różnorodne dane, takie jak informacje o ruchu drogowym, dane meteorologiczne, informacje o warunkach drogowych oraz dane z kamer monitorujących. Wykorzystując zaawansowane algorytmy i sztuczną inteligencję, te systemy są w stanie analizować dane w czasie rzeczywistym i podejmować szybkie decyzje w celu optymalizacji przepływu ruchu.

W obliczu stale rosnących potrzeb mobilności miejskiej i regionalnej, inteligentne systemy zarządzania ruchem stają się niezbędnym narzędziem dla miast i regionów dążących do efektywnego, bezpiecznego i zrównoważonego transportu. Przyjrzyjmy się zatem bliżej, jak te

innowacyjne technologie rewolucjonizują nasze sposoby podróżowania i wspierają rozwój inteligentnych miast przyszłości.

2. Ruch miejski w wymienionych sytuacjach:

Dynamiczne dostosowywanie sygnalizacji świetlnej: Systemy sterowania ruchem mających możliwość bierzącego badania ruchu na skrzyżowaniach przy pomocy różnego rodzaju kamer i sensorów oraz dynamicznego dostosowywania czasów cykli świateł, aby zoptymalizować przepływ pojazdów. Krótkie cykle świetlne w godzinach szczytu oraz dłuższe w godzinach o mniejszym natężeniu ruchu mogą zmniejszyć korki w centrach miast.

Optymalizacja tras dla pojazdów awaryjnych: Inteligentne systemy zarządzania ruchem miejskim mogłyby automatycznie identyfikować i priorytetyzować trasy dla pojazdów awaryjnych, takich jak karetka czy straż pożarna, aby umożliwić im szybkie dotarcie do celu bez zakłócania ruchu (zwłaszcza w godzinach szczytu) dla innych użytkowników dróg.

Wykorzystanie inteligentnych systemów nawigacyjnych: Systemy nawigacyjne wyposażone w funkcje ITS mogą informować kierowców o optymalnych trasach, aktualnych warunkach drogowych oraz ewentualnych zagrożeniach na drodze, co pozwala im unikać korków i oszczędzać czas podróży.

Integracja z systemami transportu publicznego: Integracja systemów ITS z systemami transportu publicznego umożliwia lepsze koordynowanie rozkładów jazdy autobusów czy tramwajów miejskich z sygnalizacją świetlną co daje możliwość zwiększenia efektywności i szybkości działania komunikacji miejskiej.

3. Wspomaganie zarządzania transportem w firmie

Oprogramowanie klasy TMS (Transport Management System) ma za zadanie wspomaganie planowania, monitoringu oraz rozliczania transportu zwłaszcza w firmach spedycyjnych.

Nowoczesne oprogramowanie cechuje się budową modułową. Ważnym blokiem funkcjonalnym jest obsługa transportu i spedycji. Istotne pozostaje zarządzanie spedycją, zarówno krajową, jak i międzynarodową. Nie bez znaczenia są również funkcje pozwalające na ofertowanie, tworzenie cenników i generowanie zleceń.

Przydatne rozwiązanie stanowi możliwość automatycznego wysyłania zleceń do kierowcy. Przy wymianie danych bardzo często zastosowanie znajduje wtedy technologia GPS. Mapa cyfrowa, która jest zintegrowana z systemem, stanowi narzędzie do wyznaczania i optymalizowania trasy. Odpowiednie dokumenty, takie jak chociażby potwierdzenia są automatycznie wysyłane do wybranych osób. Nowoczesne programy klasy TMS pozwalają na

zarządzanie flotą oraz kontrolowanie kosztów. Przede wszystkim warto zwrócić uwagę na planowanie i realizację serwisów, przeglądów oraz napraw. Można zarządzać również urlopami i zastępstwami, a także wdrożyć nadzór nad wyposażeniem pojazdu.

Przydatna jest funkcjonalność rozliczania kosztów. Chodzi tutaj o rozliczanie delegacji, zaliczek, diet, polis i szkód. System jest w stanie uwzględniać koszty serwisu łącznie z zarządzaniem akumulatorami i oponami. Z pewnością przyda się również rozliczanie palet i opakowań zwrotnych. Systemy TMS pozwalają na import kosztów z bezgotówkowych kart paliwowych.

W firmach realizujących procesy transportowe z pewnością przyda się funkcjonalność związana z fakturowaniem i zarządzaniem płatnościami. Wiele programów jest w stanie realizować funkcje, które są zarezerwowane dla typowych programów księgowych. Chodzi przede wszystkim o możliwość pracy z różnymi typami faktur takimi jak proforma, zaliczkowe, częściowe i zbiorcze. Istnieje możliwość wystawiana not i korekt księgowych. Oprogramowanie tworzy kompensaty, wezwania do zapłaty oraz dokumenty windykacyjne. Są przy tym uwzględniane kursy walut. Program może współpracować z kasami i drukarkami fiskalnymi. Oczywiście istnieje możliwość wymiany danych z dowolnymi systemami finansowo-księgowymi.

4. Inteligentne systemy transportu wewnątrzzakładowego:

Systemy transportu wewnątrzzakładowego są ważnym elementem logistycznym decydującym o przebiegu realizowanych procesów produkcyjnych w kontekście przemieszczania towarów, poziomu procesów manipulacyjnych, ochrony przed uszkodzeniem oraz utratą wartości użytkowych. Stąd też środki transportu wewnętrznego obejmują przede wszystkim maszyny i urządzenia transportowe, urządzenia do składowania oraz szereg urządzeń pomocniczych.

Podstawowe urządzenia transportowe jakie znajdują zastosowanie w nowoczesnych magazynach, to przede wszystkim wózki widłowe. Zastosowanie znajdują również wózki przegubowe z obrotowym masztem co ułatwi manewrowanie w wąskich korytarzach. Warto również wspomnieć o wózkach bezzałogowych, wózkach systemowych wysokiego składowania oraz wózkach wielokierunkowych.

W wielu fabrykach nie obejdzie się bez dźwignic czyli środkach manipulacji prostej oraz układnic będących urządzeniami poruszającymi się na szynach wzdłuż korytarzy pomiędzy regałami.

Systemy transportu wewnątrzzakładowego to również urządzenia do składowania. Ważne są tutaj urządzenia z automatycznym cyklem pracy (manipulatory) oraz roboty przemysłowe.

Istotną rolę odgrywają regały magazynowe. Stanowią one gwarancję szybkiego dostępu do magazynowanych jednostek przy zapewnieniu bezpiecznego piętrzenia ładunków. Regały magazynowe to przede wszystkim regały grawitacyjne pracujące przy zachowaniu zasady FIFO. Trzeba mieć również na uwadze regały okrężne, które można podzielić na regały o ruchu poziomym i pionowym. Ponadto zastosowanie znajdują regały paletowe gniazdowe umożliwiające swobodny dostęp do palet, umieszczonych w tzw. gniazdach regałowych. Ważne są regały półkowe, przepływowe, przesuwne, windowe, ramowe oraz wspornikowe. Ponadto niejednokrotnie zastosowanie znajdują regały tunelowe.

Wewnątrzzakładowe systemy transportowe nie obejdą się bez wyposażenia regałów i urządzeń pomocniczych takich jak chociażby przenośniki rolkowe, stałe sterownice czy krany. Z kolei urządzenia pomocnicze to urządzenia ułatwiające załadunek środków transportowych – rampy, pomosty ładunkowe i wyrównawcze, rampy ruchowe. Ważne są również urządzenia pomocnicze do składowania i manipulacji towarem, takie jak palety, paletyzery, nadstawki palet, pojemniki, jarzma.

Rysunek 15 Przykład zautomatyzowanego systemu transportowego; Źródło: https://www.press.bmwgroup.com/poland/article/detail/T0347154PL/inteligentna-logistyka-wewn%C4%85trzzak%C5%82adowa:-zautomatyzowane-systemy-transportowe-w-zastosowaniach-na-zewn%C4%85trz?language=pl

5. Dodatkowe korzyści wynikające z wykorzystania ITS:

Poprawa bezpieczeństwa drogowego: Systemy ITS mogą pomagać w identyfikowaniu zagrożeń na drodze, takich jak wypadki drogowe, obiekty na jezdni czy złe warunki

atmosferyczne. Dzięki szybkiemu powiadamianiu kierowców i służb ratowniczych oraz automatycznemu dostosowywaniu sygnalizacji świetlnej, systemy te przyczyniają się do poprawy bezpieczeństwa na drogach.

Redukcja emisji spalin i zanieczyszczeń: Poprzez optymalizację przepływu ruchu, inteligentne systemy zarządzania ruchem mogą zmniejszyć czas spędzany przez pojazdy w korkach i tym samym ograniczyć emisję spalin oraz zanieczyszczeń powietrza. Ponadto, zachęcanie do korzystania z transportu publicznego poprzez integrację z systemami ITS może dodatkowo zmniejszyć liczbę pojazdów na drogach.

Efektywniejsze wykorzystanie infrastruktury drogowej: Dzięki lepszemu zarządzaniu ruchem, inteligentne systemy mogą pomóc w efektywniejszym wykorzystaniu istniejącej infrastruktury drogowej co sprawi, że nie będzie pilnej potrzeby tworzenia nowych dróg a tym samym ograniczona zostanie emisja trujących związków związanych z ich budową do atmosfery.

6. Wyzwania dla ITS:

Koszty wprowadzenia: Jednym z głównych wyzwań związanych z wdrożeniem ITS jest koszt implementacji. Budowa infrastruktury, zakup i instalacja zaawansowanych technologii a także utrzymanie czy też późniejsze aktualizacje systemów mogą być kosztowne dla miast i agencji transportowych.

Standaryzacja systemów: Wiele systemów ITS jest opartych na różnych technologiach i standardach, co może prowadzić do problemów z interoperacyjnością między różnymi systemami. Konieczność zapewnienia kompatybilności pomiędzy różnymi systemami może stanowić wyzwanie techniczne i organizacyjne.

Bezpieczeństwo cybernetyczne: Wraz z rosnącą złożonością systemów ITS, wzrasta ryzyko ataków cybernetycznych na np. firmy transportowe lub system zarządzania ruchem w mieście . Infrastruktura i urządzenia związane z zarządzaniem ruchem mogą być na nie podatne, co może prowadzić do zakłóceń w ruchu oraz zagrożeń dla bezpieczeństwa użytkowników dróg.

Akceptacja społeczna: Wdrożenie nowych technologii w obszarze zarządzania ruchem może spotkać się z oporem społecznym ze strony np. starszych mieszkańców lub użytkowników dróg. Wprowadzenie zmian w organizacji ruchu drogowego może wymagać edukacji społecznej.

Ważna dla możliwości wprowadzania w.w systemów może się okazać utrzymanie dobrej współpracy między sektorem publicznym a prywatnym.

7. W celu zapewnienia odpowiedniego funkcjonowania inteligentnego systemu zarządzania ruchem stosuje się różnego rodzaju systemy detekcji między innymi:

Pętle indukcyjne - Składają się z cienkiego kabla umieszczonego pod powierzchnią jezdni, który tworzy pętlę. Rejestrują one zmiany pola magnetycznego generowanego przez pojazdy, dzięki czemu możliwe jest wykrywanie obecności pojazdów i mierzenie ich prędkości oraz dodatkowo wyłapywanie kierowców przejeżdżających na czerwonym świetle.

Rysunek 16 Pętla indukcyjna zatopiona w jezdni; Źródło: https://przeglad-its.pl/2014/07/25/petlowy-system-detekcji-ruchu-z-wykrywaniem-osi-pojazdow/

Kamery monitorujące (CCTV) - są wykorzystywane do monitorowania na bieżąco ruchu drogowego i identyfikowania pojazdów na drogach. Bardziej zaawansowane technicznie kamery mają funkcję rozpoznawania i analizowania numerów rejestracyjnych oraz technologie analizy obrazu, które pozwalają na zbieranie danych o natężeniu ruchu oraz przestrzeganiu sygnałów świetlnych przez kierowców.

Rysunek 17 Kamera rejestrująca przejazdy pojazdów na czerwonym świetle. Źródło: https://gazetawroclawska.pl/uwaga-kierowcy-nowe-kamery-zrobia-nam-zdjecie-gdy-jedziemy-na-czerwonym-swietle-zobacz/ar/1005459

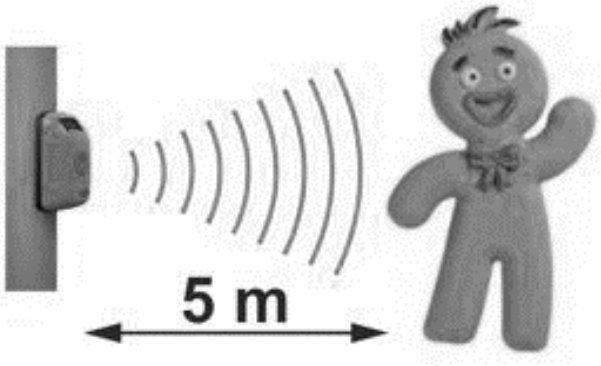
Radar – wykorzystuje on fale radiowe do pomiaru prędkości a także odległości pojazdów. Radar może być stosowany do monitorowania ruchu na drogach, wykrywania kolizji oraz rozpoznawania pojazdów znajdujących się w danym momencie w strefie zasięgu radaru.

Rysunek 18 Radar do pomiaru prędkości pojazdów; Źródło: https://korsokolbuszowskie.pl/wiadomosci/fotoradar-w-widelce-juz-dziala-sa-juz-pierwsze-mandaty/a3bDiUlKlL5mMecBAF3f

Detektory akustyczne - wykorzystują analizę dźwięku wygenerowanego przez ruch pojazdów do wykrywania ich obecności oraz mierzenia prędkości. Systemy te mogą być przydatne w miejscach, gdzie z różnych powodów niemożliwe jest stosowanie innych systemów detekcji, takich jak w strefach centrów miast z dużym wszechobecnym hałasem.

Detektory podczerwieni - wykorzystują promieniowanie podczerwone do wykrywania obiektów poruszających się na drodze. Mogą być stosowane do monitorowania ruchu pojazdów

oraz analizowania obszarów przejść dla pieszych pod kątem ich obecności, w celu zapewnienia im maksymalnego bezpieczeństwa.



Rysunek 19 System wykrywający pieszego przy przejściu dla pieszych; Źródło: https://www.apko.com.pl/przycisk_dla_pieszych_pdp_x_apko.html

8. Systemy komunikacji pojazdów do infrastruktury (V2I):

Systemy sygnalizacji świetlnej (V2I-S): Systemy te umożliwiają komunikację między pojazdami a sygnalizacją świetlną. Dzięki tej technologii pojazdy mają możliwość przekazywać informacje o swoim położeniu i prędkości do sterownika sygnalizacji świetlnej, co co sprawia, że można dynamiczne dostosować cykle świateł do natężenia ruchu i poprawić płynności ruchu na skrzyżowaniach w godzinach szczytu komunikacyjnego.

Systemy ostrzegania o niebezpieczeństwach (V2I-W): Systemy takie umożliwiają infrastrukturze drogowej przekazywanie ostrzeżeń o potencjalnych niebezpieczeństwach na drodze, takich jak wypadki, obiekty na jezdni czy złe warunki atmosferyczne, do pojazdów poruszających się w danym obszarze. Daje to możliwość kierowcy szybciej reagować na zagrożenia i unikać potencjalnych wypadków lub kolizji.

Systemy zarządzania pasami ruchu (V2I-HOV): System taki jest stosowany w pasach ruchu dla pojazdów z wysoką liczbą pasażerów, takich jak pasy HOV lub buspasy. Są one w stanie monitorować liczbę pasażerów w pojazdach poruszających się tymi pasami oraz zarządzać dostępem do nich na podstawie ustalonych kryteriów.

Systemy zarządzania parkingiem (V2I-P): umożliwiają komunikację między pojazdami a infrastrukturą parkingową, taką jak parkingi miejskie czy garaże parkingowe itd. Dzięki temu kierowcy mogą otrzymywać informacje na temat dostępnych miejsc parkingowych, kosztów parkowania oraz innych istotnych danych związanych z parkowaniem swoich pojazdów.

Systemy zarządzania ruchem drogowym (V2I-T): są one stosowane do komunikacji między pojazdami a infrastrukturą zarządzania ruchem drogowym, taką jak centra zarządzania

ruchem czy systemy informacji drogowej. Dzięki tej technologii pojazdy mogą otrzymywać informacje na temat bieżących warunków drogowych, korków, wypadków oraz alternatywnych tras, co pozwala kierowcom pojazdów na podejmowanie świadomych decyzji co do tras swoich podróży.

Systemy zarządzania energią (V2I-EV): Systemy V2I-EV są stosowane w przypadku pojazdów elektrycznych, umożliwiając im komunikację z infrastrukturą ładowania. Dzięki tej technologii pojazdy elektryczne mogą otrzymywać informacje na temat dostępności stacji ładowania, cen energii oraz optymalnego czasu ładowania, co umożliwia efektywne zarządzanie energią i zwiększenie efektywności użytkowania pojazdów elektrycznych.

Wymienione przeze mnie powyżej systemy komunikacji pojazdów do infrastruktury (V2I) są najważniejszymi elementami inteligentnych systemów zarządzania ruchem, umożliwiając poprawę płynności ruchu drogowego zapewniają bezpieczeństwo kierowców oraz pozwalają efektywnie wykorzystać infrastrukturę drogową. Ich zastosowanie pozwala na wspieranie rozwoju zrównoważonej mobilności.

9. Krótki opis pozostałych systemów komunikacji:

V2V - umożliwiają bezpośrednią komunikację między pojazdami poruszającymi się w danej chwili po drodze. Dzięki temu pojazdy mogą wymieniać informacje na temat swojego położenia, prędkości, kierunku jazdy oraz stanu drogi. Systemy te są kluczowe dla poprawy bezpieczeństwa drogowego poprzez ostrzeganie kierowców przez inne pojazdy o potencjalnych zagrożeniach na drodze i zapobieganiu kolizjom.

V2P - umożliwiają komunikację między pojazdami a pieszymi, bądź innymi uczestnikami ruchu niebędącymi kierowcami, takimi jak rowerzyści czy biegacze. Dzięki temu pojazdy mogą ostrzegać pieszych o swojej obecności, zwłaszcza w sytuacjach, gdy pieszy znajduje się w strefie zagrożenia.

V2N - umożliwiają komunikację między pojazdami a siecią telekomunikacyjną. Dzięki temu pojazdy mogą otrzymywać informacje na temat ruchu drogowego, warunków atmosferycznych oraz innych istotnych danych związanych z podróżą z sieci telekomunikacyjnej. Jest to istotne dla usług związanych z nawigacją i informacją drogową.

V2G - umożliwiają komunikację między pojazdami elektrycznymi a siecią energetyczną. Dzięki temu pojazdy elektryczne mogą pełnić rolę zasobu energii odnawialnej, umożliwiając magazynowanie i oddawanie energii do sieci w zależności od potrzeb. Jest to istotne dla zarządzania energią w sieciach elektrycznych oraz zwiększenia udziału energii odnawialnej.

V2B - obejmują wszystkie formy komunikacji między pojazdami, infrastrukturą, pieszymi oraz sieciami telekomunikacyjnymi i energetycznymi. Jest to zintegrowany system komunikacji, który umożliwia kompleksową wymianę informacji między wszystkimi uczestnikami ruchu drogowego oraz infrastrukturą związaną z transportem.

10. Trendy, które mogą ukształtować rozwój tego obszaru w przyszłości:

Autonomiczne pojazdy – są pojazdami zdolnymi do poruszania się po drogach bez ingerencji człowieka (kierowcy). Stanowią one rewolucję transportową a w połączeniu z inteligentnymi systemami zarządzania ruchem będzie w stanie dobierać trasy optymalnie dla oszczędności czasu i pieniędzy właściciela.

Sztuczna inteligencja i uczenie maszynowe - umożliwia tworzenie coraz bardziej zaawansowanych technologicznie algorytmów analizy danych oraz systemów, które są zdolne do podejmowania decyzji w czasie rzeczywistym. Dzięki temu inteligentne systemy zarządzania ruchem będą mogły dokładniej przewidywać zmiany w ruchu drogowym oraz lepiej dostosowywać strategie jego zarządzaniem.

Zrównoważona mobilność - Coraz większe zainteresowanie zrównoważoną mobilnością oraz rosnąca świadomość ekologiczna społeczeństwa skłania do poszukiwania alternatywnych sposobów transportu, takich jak rowery, hulajnogi elektryczne czy transport publiczny. Rozwój inteligentnych systemów zarządzania ruchem będzie więc obejmował także integrację tych środków transportu oraz promowanie rozwiązań sprzyjających zrównoważonej mobilności.

Rozwój miejskich ekosystemów transportowych - Coraz większa liczba miast stawia na rozwój kompleksowych ekosystemów transportowych, które integrują różne środki transportu, usługi transportowe oraz inteligentne systemy zarządzania ruchem. Celem jest stworzenie spójnych, efektywnych i zrównoważonych systemów transportowych, które spełniają potrzeby mieszkańców i minimalizują negatywny wpływ transportu na środowisko.

11. Podsumowanie:

Podsumowując, artykuł ten skupia się na roli inteligentnych systemów zarządzania ruchem w obliczu rosnącej urbanizacji i zwiększonej liczby pojazdów na drogach. Te zaawansowane technologicznie rozwiązania mają za zadanie optymalizować przepływ ruchu, poprawiać bezpieczeństwo drogowe oraz redukować negatywny wpływ transportu na środowisko. Omówione zostały różne aspekty funkcjonowania i korzyści płynące z wykorzystania tych systemów, takie jak dynamiczne dostosowywanie sygnalizacji świetlnej, optymalizacja tras dla pojazdów awaryjnych czy integracja z systemami transportu publicznego.

Ponadto, przeanalizowane zostały tu wyzwania związane z wdrożeniem inteligentnych systemów zarządzania ruchem, takie jak koszty implementacji, konieczność standaryzacji systemów czy zagrożenia związane z cyberbezpieczeństwem. Zaprezentowane zostały również różne systemy detekcji oraz komunikacji pojazdów z infrastrukturą, wraz z ich funkcjami i zastosowaniami.

Na zakończenie, omówiono trendy przyszłościowe, które będą kształtować rozwój tego obszaru, takie jak rosnące znaczenie autonomicznych pojazdów, wykorzystanie sztucznej inteligencji czy rozwój zrównoważonych ekosystemów transportowych. Te tendencje mają na celu budowę bardziej efektywnych, bezpiecznych i zrównoważonych systemów transportowych, które spełniają potrzeby społeczne i minimalizują negatywny wpływ na środowisko.

Źródła internetowe

1. http://yadda.icm.edu.pl/yadda/element/bwmeta1.element.ekon-element-000171562143 (dostęp: 27.04.2024).
2. file:///C:/Users/HP/Downloads/Lewicki.pdf (dostęp: 27.04.2024).
3. https://sprint.pl/pl/uslugi/inteligentne-systemy-transportowe (dostęp: 27.04.2024).
4. https://www.optidata.pl/tms-co-to-jest-i-jak-go-wykorzystac-w-procesach-logistycznych/ (dostęp: 27.04.2024).
5. https://gliwice.eu/aktualnosci/miasto/wszystkie-zalety-its-na-ekranie (dostęp: 27.04.2024).
6. file:///C:/Users/HP/Downloads/TMiR_5_2016_Aleksandrowicz_Sposoby.pdf (dostęp: 27.04.2024).
7. https://orpa.pl/audi-wprowadza-system-v2i-pozwalajacy-na-komunikacje-pojazdu-z-sygnalizacja-swietlna/ (dostęp: 27.04.2024).
8. https://evertiq.pl/news/12593 (dostęp: 27.04.2024).
9. https://mubi.pl/poradniki/technologia-v2x/ (dostęp: 27.04.2024).

Jakub Bocek, Patryk Krupa, Piotr Dubaj, Sławomir Pareniak, Katarzyna Maternia
Koło naukowe Elektroniki i Technologii Informacyjnych

dr inż. Bartosz Pawłowicz
Opiekun Koła Naukowego

Zastosowanie nowych technologii w rolnictwie: rolnictwo precyzyjne, roboty zbierające i pomocne roboty

Streszczenie

Celem prezentacji jest ukazanie na czym polega rolnictwo precyzyjne czyli, pozyskiwanie informacji i czasowej zmienności cech roślin oraz środowiska dla tworzenia analiz komputerowych sporządzających pomocne informacje dla upraw, oraz tworząca się powoli duża popularność w tym temacie. W prezentacji znajdzie się opis Internetu Rzeczy, który w głównej mierze usprawnia prace w rolnictwie, zostanie opisana praca satelity i robotów pomagających w rolnictwie, opisanie ich, pokazanie statystyk i sprawdzenie czy dany robot jest opłacalny w rolnictwie precyzyjnym. Robot urządzenie sterowane przez system komputerowy, który ma wykonać określone działanie. Roboty są tworzony żeby usprawniać i zastępować pracę, która wykonuje człowiek. Na razie większość robotów jest w fazie testów więc i tak trzeba przy nich człowieka, który nadzoruje ich prace ale wszystko idzie w tą stronę aby roboty wykonywały samodzielnie swoje zadania. Niestety koszt takich sprzętów jest ogromny, dlatego większość firm proponuję wypożyczanie robota, aby ten wykonał swoją prace. Firmy dbają o sprawność robotów, a niektóre nawet przydzielą ludzi do obserwowania pracy swojego dzieła.

Słowa kluczowe: rolnictwo precyzyjne, roboty, Internet rzeczy

1. Wprowadzenie

Jednych z najbardziej innowacyjnych i rozwijających się technologii w rolnictwie jest tak zwane rolnictwo precyzyjne, rolnictwo precyzyjne nie obejdzie się bez stosowania dodatkowych przyrządów takich jak satelity. Machine sync jest jedną z technologii rolnictwa precyzyjnego, która pokrótce zostanie opisana w prezentacji.

Roboty zbierające i pomagające w rolnictwie innowacyjnym są niezbędnymi maszynami, przy których farmy nie mogą działać autonomicznie. Przez ostatnie lata została zbudowana niezliczona ilość takich robotów. Roboty zawierają w sobie najnowsze technologie potrzebne do ich prawidłowego funkcjonowania.

2. Rolnictwo precyzyjne

Rolnictwo precyzyjne stanowi kompleksowy system gospodarowania, który dostosowuje poszczególne elementy agrotechniki do zmiennych warunków na konkretnych częściach pola, w zależności od aktualnego stanu rozwoju roślin czy właściwości glebowych. Niezbędne dane pozyskiwane i przetwarzane są przy wykorzystaniu wysoko rozwiniętych technologii

nawigacyjnych i informatycznych. Rolnictwo precyzyjne stanowi pozyskiwanie informacji o przestrzennej i czasowej zmienności cech roślin oraz środowiska. Dane te uzyskiwane są za pomocą nowoczesnego sprzętu sterowanego satelitarnie i komputerowo, a następnie analizowane przez specjalne oprogramowanie. Informacje pozyskiwane są z kilku lub kilkunastu źródeł; pomiary pola, określenie zasobności gleby, praca urządzeń z sensorami, wykorzystanie bezzałogowych obiektów latających ze specjalną kamerą rejestrującą kondycję roślin. Na podstawie tych danych jest tworzona mapa pola tak zwana baza informacji GIS zawierająca informacje o właściwościach gleby i roślin, agrofagach (czyli patogenach, szkodnikach i chwastach), uzyskanym plonie oraz parametrach meteorologicznych. Jest to z kolei podstawą do przygotowania map aplikacji i dokładnych zabiegów agrotechnicznych, ale również służy do tworzenia bazy danych oraz dokumentacji. Najważniejszymi elementami rolnictwa precyzyjnego są siew ze zmienną gęstością oraz używanie środków ochrony roślin i nawozów w dopasowanej dawce dokładnie w miejscach, gdzie jest to wymuszane z dokładnością do kilku centymetrów. W tym celu są używane maszyny rolnicze wyposażone w GPS oraz komputer.

Korzyści stosowania nowych technologii, rolnik ma możliwość maksymalnego użycia potencjału pól i zwiększenia plonów oraz poprawy ich jakości. Przy tym posiada on pełną kontrole nad uprawami. Zróżnicowane dawkowanie agrochemikaliów ma wpływ na oszczędność pieniędzy, która wynika ze zmniejszonego użycia nawozów i oprysków. A użycie nawigacji równoległej z użyciem GPS zmniejsza zużycie paliwa podczas prac polowych oraz poprawia komfort pracy operatora i daję możliwości wykonania operacji w trudnych warunkach pogodowych lub w nocy. W efekcie gospodarowanie w systemie precyzyjnym podnosi efektywność produkcji, co przekłada się na wyższe zarobki. Precyzyjne stosowanie nawozów i środków ochrony roślin jest korzystne dla środowiska i zdrowia człowieka. Nadmiar związków chemicznych nie gromadzi się w plonach i nie wpływa do cieków wodnych oraz gleby. Ponadto zapobiega ubożeniu podłoża w składniki pokarmowe.

Wdrażanie technik rolnictwa precyzyjnego nie musi wiązać się z dużymi kosztami i rewolucją w gospodarstwie. Na start można wprowadzić niektóre elementy. Rolnictwo precyzyjne ma bowiem wiele aspektów. Jednym z pierwszych kroków do wdrożenia rolnictwa precyzyjnego jest szczegółowe określenie powierzchni upraw i zbadanie gleby pod kątem odczynu i zasobności w składanki pokarmowe. Po pomiarach pól oraz zbadaniu próbek z glebą wykonane są mapy bogatości i zmienności glebowej, które umożliwiają dokładne zaplanowanie zabiegów agrotechnicznych w zależności od rodzaju upraw. Nieodłącznym elementem rolnictwa precyzyjnego są maszyny i urządzenia do zmiennego dawkowania nawożenia system

VRA wyposażone w komputery i GPS. Sprzęty te na podstawie utworzonej wcześniej mapy aplikacyjnej umożliwiają aplikację środków w określonej ilości osobno dla każdej wydzielonej strefy. Nowoczesne maszyny rolnicze takie jak opryskiwacze, rozsiewacze wyposażone w komputery i czujniki w trakcie prac na polu mierzą wartości wielu parametrów w czasie rzeczywistym. Oznacza to, że dany zabieg jest robiony na podstawie bieżącej oceny kondycji upraw. Przykładem może być urządzenie N-Sensor do określania deficytu azotu w roślinach. Często wykorzystywany jest kombajn z czujnikiem pomiaru plonów, który daje nam możliwość zbiorów i utworzenia mapy jakości ziarna. Sensor wykonuje pomiar objętościowy oraz pomiar masy, a także określa wilgotność plonu. Zastosowanie nawigacji równoległej z użyciem GPS pozwala na prowadzenie pojazdów po pasach równoległych. System wykonuje maksymalną szerokość roboczą maszyny oraz zapobiega nakładaniu się przejazdów i występowaniu mijaków. W efekcie poprawia się komfort operatora i wydajność pracy, niezależnie od warunków pogodowych oraz pory dnia.

Rolnictwo precyzyjne pozwala na odejście od stosowania nawozów i pestycydów na niektórych polach, na rzecz użycia tylko niewielkiej ilości danego środka jaka jest niezbędna w danych warunkach i na konkretnym obszarze. Takie zastosowanie ma wpływ także na jakość produktów rolnych i obniża ryzyko zanieczyszczenia pozostałościami pestycydów oraz pozwala na ograniczenie negatywnego wpływu działalności rolniczej na środowisko. Zastawanie rozwiązań rolnictwa precyzyjnego przyczynia się również do poprawy efektywności wykorzystania środków produkcji, zwiększenia wydajności pracy ludzi i maszyn oraz usprawnienia zarządzania w gospodarstwie. Pozwala w znacznym stopniu na zmniejszenie kosztów produkcji, jednocześnie zapewniając optymalizację jakości płodów rolnych. W Polsce odpowiedzią na rosnące zapotrzebowanie w zakresie nowych technologii w rolnictwie są instrumenty w ramach Wspólnej Polityki Rolnej. W Planie Strategicznym dla Wspólnej Polityki Rolnej na lata 2023-2027 przewidziano szereg rozwiązań, które przyczynią się do unowocześnienia rolnej gałęzi gospodarki. W ramach interwencji wspierane będą inwestycje wykorzystujące technologie cyfrowe dotyczące stosowania nawozów i środków ochrony roślin, zrównoważonego zarządzania wodą czy zarządzania stadem zwierząt oraz inwestycje dotyczące wspierania procesów podejmowania decyzji w gospodarstwie. Rolnictwo precyzyjne jest dla każdego. Nowoczesne technologie są dopasowanie do gospodarstw o różnej wielkości, w zależności od potencjału upraw i indywidualnych potrzeb rolnika. Rolnictwo precyzyjne wymaga jednak od producentów rolnych odpowiedniej wiedzy i przygotowania.

Machine Sync jest to technologia rolnictwa precyzyjnego przedstawiona przez firme John Deere. Machine Sync łączy bezprzewodowo wiele maszyn takie jak ciągniki, kombajny i

samojezdne sieczkarnie we własnej sieci, jedna maszyna zostaje tak zwanym liderem, który kontroluje prędkość, kierunek i położenie innych maszyn. Technologia ta zapewnia precyzyjne wyładowanie w czasie jazdy maszyn, ustala pierwszeństwo wyładowania kombajnu i jak trzeba zwiększa wydajność zbiorów. System ten pozwala łączyć do 6 maszyn w odizolowanej sieci bezprzewodowej. Operator kombajnu może przemieszczać ciągnik do tyłu, do przodu, aby zapewnić równomierne wyładowywanie ziarna do przyczep. Kombajny w tej samej grupie udostępniają informacje o poziomie napełnienia zbiornika ziarna.

Satelity w rolnictwie precyzyjnym są wykorzystywane aby pozyskać dużo cennych informacji o polu na którym uprawiamy rośliny. Kamery multispektralne które posiadają satelity pobierają informacje o ilości odbitych fal, pokazując różnice w wegetacji roślin w postaci wskaźnika NDVI (wskaźnik pozwalający określić stan rozwojowy oraz kondycje roślinności). Umożliwia to na ukazanie różnic pomiędzy stanem roślinności na jednym polu. Pomiary podczerwonych fal elektromagnetycznych odbijają się od zielonych części roślin pozwalając precyzyjnie określić ilość zgromadzonej tam biomasy, która jest powiązana z potencjałem produkcyjnym. Mapy stworzone przez satelity możemy wykorzystywać do zmiennego wysiewu nawozów, gdy mamy dostęp do obrazów z poprzednich lat możemy określić produkcyjność danego pola. Wtedy rolnik nie będzie dostarczał dużej ilości nawozu tam gdzie nie ma sensu go dostarczać, a tam gdzie rośliny mają większy potencjał rolnik może zmaksymalizować uprawę poprzez dostarczenie większej ilości nawozu.

3. Internet Rzeczy (loT)

Jest to technologiczna koncepcja, w której wiele różnorodnych urządzeń elektronicznych poprzez łącze internetowe jest podpiętych do wspólnej sieci. Dzięki uniwersalnym protokołom komunikacyjnym urządzenia te mogą wymieniać dane, bez wkładu człowieka. Dane te mogą być pozyskiwane z rozproszonych terytorialnie urządzeń, a użytkownik posiadający upoważnienie do wglądu może sprawdzić dane wszędzie tam gdzie jest Internet. Z poziomu programu można też uruchamiać instalacje lub urządzenia.

Dzięki Innowacyjnym technologiom możemy przełożyć się na zwiększenie produktywności roślin, poprawę dobrostanu zwierząt i zarządzanie ryzykiem. Najważniejsze że możemy zwiększyć rentowność rolnictwa dzięki wyższemu popytowi na produkt o wyżej klasie jakości. Internet Rzeczy daje możliwość:

- Zastosowania sensorów w uprawie, hodowli i ochronie roślin, hodowli zwierząt, nadzór nad fazami wzrostu,
- Śledzenie stanów magazynowych,

- Zadawania paszy,
- Zarządzanie i monitoring gospodarki pasiecznej,
- Sprawdzanie warunków pogodowych na polu i mikroklimatyczne w szklarniach,
- Zarządzanie używaniem nawozów i środków ochrony roślin,
- Zarządzanie zużyciem wody,
- Optymalizacja nawożenia,
- Sterowanie parametrami technologicznymi w chłodniach, przechowalniach i suszarniach,

Zastosowanie Internetu Rzeczy (IoT). Internet Rzeczy jest wykorzystywane zarówno w konwencjonalnym, zrównoważonym jak i rolnictwie ekologicznym. W tej technologii podpięte mogą być też stacje pogodowe. Poprzez łącze internetowe rolnik może sterować procesami produkcyjnymi w telefonie. Może On zdalnie uruchomić deszczownię czy wentylację w szklarni. Podstawą funkcjonowania rolnictwa cyfrowego są stacje pogodowe oraz różnego rodzaju sensory i czujniki dostarczające szereg danych w czasie rzeczywistym. Urządzenia te dzięki swojej rozbudowanej strukturze mierzą szereg potrzebnych w produkcji roślinnej informacji. Ilość opadów, wilgotność, temperatura powietrza oraz gleby, przewodność elektryczną gleby w różnych częściach powierzchni i pH gleby. Za pomocą Internetu Rzeczy dane te są przekazywane do narzędzi analitycznych znajdujących się w chmurzę obliczeniowej w celu analizy, poprzez które rolnik może doprowadzić do zwiększonej produkcji lub poprawić jej efektywność. Mianowicie za pomocą rozmieszczonych na gospodarstwie stacji pogodowych i czujników rolnicy mogą rozpoznać jakość gleby i określić warunki klimatyczne w obrębie danego terenu, w tym samym może on dobrać odpowiednią odmianę rośliny i wykonywać przemyślane działania w celu zwiększenia wydajności produkcji. Również poprzez przedmioty takie jak czujniki i stację pogodowe, które reagują na wszelkie anomalie rolnicy mogą monitorować uprawy i podejmować stosowne działania na podstawie zebranych informacji np. w zapobieganiu rozprzestrzeniania się chorób roślinnych. Rolnicy mając informację ile plonów mogą zebrać, będą mogli zaplanować ilość sprzedaży płodów rolnych. Rolnik również będzie mógł zaplanować koszty przez algorytmy wchodzące w skład programów do zarządzania gospodarstwem, dokonuje on analizy kosztów produkcji poszczególnych upraw na danych polach. Poprzez IoT możliwie jest też zwiększenie kontroli nad procesem produkcyjnym, jak i utrzymanie wysokich standardów produkcji. Po przez aplikację podłączone do systemu możliwe jest łatwiejsze rejestrowanie zbiorów z wykorzystaniem kodów Qr albo kodów kreskowych. Poprzez kody oznacza się pracowników wykonujących swoją pracę przy zbiorach

roślin, pojemniki, do których zbierają produkt oznacza się kodem i po odczytaniu znacznika informację o owocach czy warzywach trafia do bazy danych. Dzięki tym technologią, w których zastosowanie mają czytniki ułatwiony jest dostęp do informacji o produkcie i ułatwione jest rozliczenie z pracownikami. Za pomocą Qr można również kontrolować w sposób automatyczny procesy zachodzące w szklarni czy magazynie. Czytniki tych znaczników mają wgląd do danych rozpoczęcia/zakończenia danego procesu produkcyjnego jaki i przepływów materiałów.

Oczywiście dane interpretuję rolnik i ostateczna decyzja zależy od niego ale jest on wspierany przez narzędzia analityczne, które dostarczają mu masę informacji by mógł podjąć dobrą decyzję. Tym bardziej że dostęp od informacji rolnik ma również poza swoim gospodarstwem. Również ważnym aspektem jest jakość, rodzaj i odpowiedni dobór czujników do urządzeń, maszyn i miejsc w których mają one pracować. W tej chwili w Polsce nie funkcjonuje zbyt wiele gospodarstw, które pozwoliły by na zaprezentowanie możliwości Internetu Rzeczy, co przekłada się na niską świadomość branży w zakresie technologii Internetu Rzeczy (Bazując na danych z 2021).

4. Roboty zbierające w rolnictwie

Robot do zbierania jabłek z Hauward w Kalifornii. Abubdabt Ribitucs porusza na podwoziu kołowym robot jest wyposażony w ramiona w postaci elastycznego przewodu. Działanie maszyny przypomina odkurzacz, robot po tym jak namierzy na drzewie dojrzałe jabłko zrywa go zasysając niczym odkurzacz.

Rysunek 20 Zdjęcie: Robot do zbierania jabłek, element robota odpowiedzialny za zbiory
Źródło:https://kobietawsadzie.pl/sam-zrywa-i-pakuje-jablka-zobacz-robota-przyszlosci/

Jabłka rozpoznawanie są za pomocą systemu analizy obrazu wykorzystującego sztuczną inteligencję. Użytkownik robota musi najpierw określić poziom dojrzałości owoców oraz ich barwę jako kryterium jakim będzie posługiwał się robot podczas zbierania. Następnie owoce przenoszone są po ścieżce wyścielonych rurek i trafiają do pojemnika znajdującego się na maszynie. Robot porusza się pomiędzy drzewami wykorzystując technologię radarową. Według licznych szacunków maszyna może uzyskać dostęp od 50% do 90% owoców na drzewach, w zależności od zarządzania nasadzeniem. W rzędzie maszyna sama się porusza ale pracownik musi doprowadzić robota do rzędu za pomocą pilota. Pracownicy zabierają pełne pojemniki jabłek i ładują puste pojemniki na robota, ale konstruktorzy twierdzą że będzie on miał możliwość automatycznej zamiany pojemników. Ze względu na cenę i okres wykorzystywania robot ma być wynajmowany.

Technologię zasysania owoców z drzewa wykorzystuję robot skonstruowany przez australijską firmę Ripe Robotics. Chcąc zautomatyzować zbiory owoców firma stworzyła robota który zbiera jabłka, śliwki, grusze i pomarańcze. Robot Eve jest wyposażony w zaawansowaną sztuczną inteligencję i jest połączony z chmurą, dzięki czemu możesz śledzić jego postępy, gdziekolwiek jesteś. Robot wyposażono w czujnik i kamery, które podczas przemieszczania się maszyny w rzędzie rozpoznają, gdzie znajduję się dobry owoc, zgodny z jego normami. Następnie używa miękkiego systemu ssącego, aby zebrać owoc. System ssący

minimalizuję uszkodzenia owoców na drzewie. Maszyna jest w fazie testów a do produkcji ma być gotowa do końca 2023.

Rysunek 21 Zdjęcie: Robot do zbierania jabłek, element robota odpowiedzialny za zbiory Źródło:https://www.sadyogrody.pl/logistyka_i_opakowania/107/roboty_do_zbioru_owocow_sprawdzamy_nowe_technologie,31205.html

Hiszpańska firma Agrobot opracowała automatyczny kombajn Agrobot SW 6010 przeznaczony do zbierania truskawek. Robot jest czterokołowcem wyposażonym w dwadzieścia cztery ruchome ramiona, które mogą identyfikować i zbierać dojrzałe truskawki. Ramiona wyposażono w dwa cienkie i ostre jak brzytwa noże, które odcinają szypułkę od dojrzałej truskawki. Tak zerwane owoce spadają do koszyczka z gumowymi rolkami skąd transportują się na taśmę przenośnika transportującego truskawki do strefy ich pakowania. Znajdujący się na platformie operatorzy mogą od razu sprawdzać i oceniać zebrane przez robota truskawki i pakować je na tace. W konstrukcji robota zastosowano kamery i system wizyjny, który analizuje indywidualnie każdy owoc, sprawdza jego kształt i kolor, a następnie, gdy stwierdza że truskawka jest dojrzała wydaje polecenie do wykonania precyzyjnego cięcia. Robota wyposażono jeszcze w czujniki ultradźwiękowe, które w sposób ciągły wykrywają odległość pomiędzy kołami a rzędem pola truskawek, utrzymując zadany tor jazdy maszyny i nie dopuszczający do uszkodzenia owoców.

Rysunek 22 Zdjęcie: Robot do zbierania truskawek. Zdjęcie pokazuje jego wygląd i budowę Źródło: https://www.sadyogrody.pl/agrotechnika/103/au

Autonomiczną maszynę do zbioru truskawek zaproponowała brytyjska firma Dogtooth Technologies, twórca inteligentnych robotów. Robot ma zastosowanie w uprawach stołowych. Maszyna składa się z platformy na podwoziu gąsienicowym z którą zintegrowano ramiona robota. Robot wykorzystuje algorytmy widzenia maszynowego i planowania ruchu do rozpoznawania i lokalizowania dojrzałych owoców, które mają zostać zerwane. W celu identyfikacji czy owoc jest dojrzały zastosowano kilka kamer, które wykonują ruchy w górę i w dół, aby uzyskać szczegółowy widok uprawy. Do poruszania się w tunelu foliowym stonuję się nawigacje GPS.

Rysunek 23 Zdjęcie: Robot do zbierania truskawek. Zdjęcie pokazuje jego wygląd i budowę Źródło: https://www.sadyogrody.pl/agrotechnika/103/au

Harvest CROO (Computerized Robotic Optimized Obtainer) Robotics to amerykański start-up, który opracował autonomiczną maszynę do zbioru truskawek, sposobem tym zamienili pracę 30 osób. Konstrukcję maszyny stanowi kołowe podwozie pod którym znajduje się szesnaście ramion z chwytakami do zrywania owoców. Maszynę wyposażono w zestaw kamer i system wizyjny z algorytmem sztucznej inteligencji, który skanuje każdą truskawkę i określa, czy jest ona gotowa do zbioru. Zebrane truskawki są następnie poddawane dalszej kontroli a system wizyjny określa czy nadają się do sprzedaży bezpośredniej, przetwarzania lub odrzucenia. Autonomię poruszania się robota zapewnia system LIDAR i nawigacja GPS. Robot jest zrobiony do pracowania także w porach nocnych.

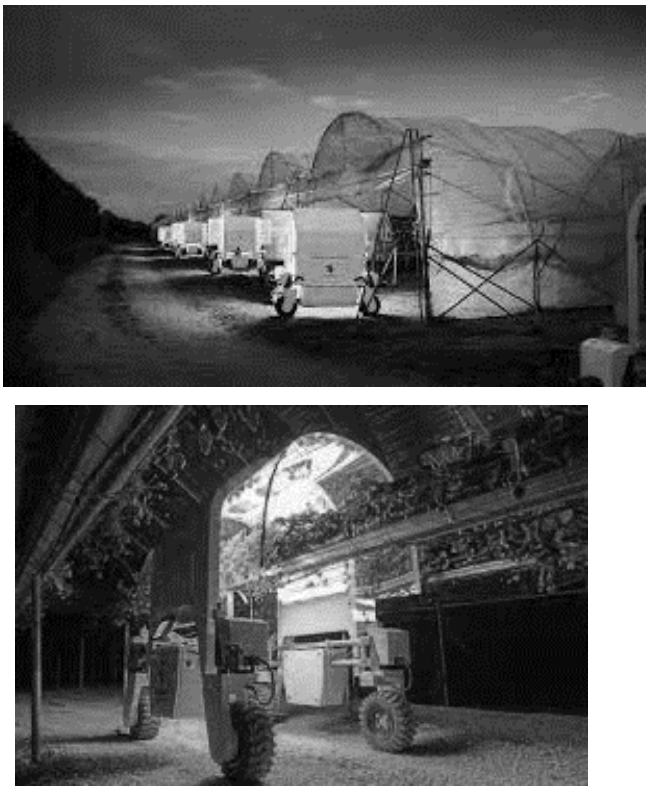
Rysunek 24 Zdjęcie: Maszyna do zbierania truskawek. Zdjęcie pokazuje jego wygląd i budowę Źródło: https://www.harvestcroorobotics.com/

Autonomiczna maszynę do zbioru truskawek opracowała również belgijska firma Octinion. Robot Rubin jest wyposażony w ramię ze specjalnym uchwytem wykonanym z tworzywa sztucznego. Uchwyt dopasowuje się do kształtu tryskawki i nie powoduje jej uszkodzenia. Ramię robota może poruszać się w wielu płaszczyznach. Zbiera ono owoce i układa w koszyku z wagą. Zarówno do poruszania się robota jak i zbioru z selekcją truskawek zastosowano system wizyjny wykorzystujący sztuczną inteligencję.

Rysunek 25 Zdjęcie: Uchwyt robota do zbierania upraw, plastikowy element łapiący owoce Źródło: https://www.sadyogrody.pl/agrotechnika/103/aut

5. Roboty pomagające w rolnictwie

Saga Robotics stworzyła roboty o nazwie Thorvald naświetlające światłem UV-C, które zwalczają choroby grzybowe między innymi mączniaka prawdziwego i eliminują potrzebę stosowania środków chemicznych. Zabiegi robota okazały się na tyle skuteczne, że firma, która zajmuję się wypożyczeniem robotów odnotowała znaczne zwiększenie się listy oczekujących na usługę. Robot jest przystosowany do działania nocą, dzięki czemu nie powstaję kolizja z pracownikami zbierającymi owoce. Saga Robotics nie sprzedaję robota, ale pracę, którą wykonuje, Firma pobiera opłatę za hektar, która obejmuję działanie robota. Robot waży około 200kg i jest w pełni elektryczny, można go ładować ze gniazdka elektrycznego. Firma planuję liczbę robotów i liczbę akumulatorów potrzebnych do pracy na danym polu, więc hodowca nie musi się martwić o ładowanie robota. Robot nawiguje za pomocą zaawansowanych czujników lokalizacji. Jest to wykorzystywane jako dane wejściowe do oprogramowania nawigacyjnego, które jest opracowane wewnętrznie poprzez zespół Saga Robotics, i to umożliwia robotowi poruszanie za rzędami oraz jazdę między rzędami i tunelami całkowicie autonomicznie. Robot działa całkowicie autonomicznie, ale jest zdalnie monitorowany przez zespół oprogramowania firmy Saga. Wielkość robota jest dostosowana do jego otoczenia. Przykładowo wersja robota pracująca przy stołowej produkcji truskawek w tunelach ma kształt łuku, który umożliwia mu przejeżdżanie po stołach, całkowicie zanurzając rośliny światłem UV-C. Prędkość robota nie przekracza 5km/h, ale również zależy od wykonywanej pracy.



Rysunek 26 Zdjęcie: Robot do naświetlania zwalczający choroby. Zdjęcie pokazuje jego prace Źródło:https://www.farmer.pl/techniŹródło:https://www.farmer.pl/technika-rolnicza/maszyny-rolnicze/leczenie-swiatlem-zamiast-oprysku-robot-thorvald-saga-robotics,118057.html

Przykład stosowania robota Thorvald: Robot wykonał zabieg UV-C w celu ochrony truskawek na około 10,6 hektarach ziemi Clock House i 2 hektarach w Hugh Lowe, oba w hrabstwie Kent w Wielkiej Brytani, od marca do października 2021r. W tym okresie żadne z gospodarst nie musiało opryskiwać swoich truskawek środkami chemicznymi zawalającymi mączniaka prawdziwego. Dyrektor generalny Saga Robotics powiedział: „Przez cały sezon nasze roboty skutecznie przetworzyły ponad 7300 km liniowych truskawek, zapewniając całkowicie skuteczną obróbkę, 100 % niezawodność obsługi robotów i brak awarii. Te wyniki tylko zwiększyły i tak już spore zainteresowanie naszą usługą.”. „Żadna substancja chemiczna nie była potrzebna do ochrony tych roślin przed mączniakiem przez cały sezon, a to świetna wiadomość dla hodowców, ich klientów i konsumentów” dodał Dan Sargent, szef działu nauki o roślinach w Saga Robotics.

Robot jako pies w Rolnictwie. Wynalazkiem, który ma pomóc przy pracy w gospodarstwie może stać się niewielki robot w kształcie psa, o nazwie Spot. Robo pies został opracowany przez amerykańską firmę trudniącą się w tworzeniu robotów Boston Dynamics. Wynalazek był

testowany w Nowej Zelandii. Jak dotąd z powodzeniem porusza się po wymagającej trasie fermy. Docelowo Spot ma pomagać w zganianiu i nadzorowaniu owiec, a również w kontrolowaniu stanu uprawy. Dyrektor generalny Rocos, nowozelandzkiej firmy, która zajmowała się przeprowadzaniem testów Spota, wyjaśnia że robot może być bardzo przydatnym urządzeniem dla rolników. Zgłaszający się klienci coraz częściej chcieliby zautomatyzować wiele fizycznych czynności na fermach, które są czasochłonne, nudne a czasami nawet niebezpieczne. Zespół z takich autonomicznie poruszających się robotów mógłby zapewnić rolnikom dostęp do bardzo dokładnych i aktualnych szacunków dotyczących upraw w dużych gospodarstwach ale jak na razie nie wydaję się aby Spot, czyli robo pies odesłał na emeryturę psy pasterskie. Robot jest wyposażony w LIDAR, czyli metodę pomiaru odległości poprzez oświetlanie światłem laserowym, kamerę 360 stopni i czujniki loT. Robo pies posiada cztery nogi i porusza się płynnym przejściem, który potrafi omijać przeszkody, ładnie pokonuje nierówne powierzchnie potrafi schodzić po schodach i potrafi transportować aż 13,6 kg ładunku. Spot może służyć jako jednostka dozorująca lub monitorująca, która zbiera dane.

Rysunek 27 Zdjęcie: Robot spot. Zdjęcie pokazuje jego budowę w odniesieniu do człowieka Źródło:https://tech.wp.pl/robot-boston-dynamics-w-nieŹródło:https://tech.wp.pl/robot-boston-dynamics-w-niezwyklej-roli-sprawdzil-sie-jako-pies-pasterski-wideo,6512883446225025a

Ciekawostka o Spocie. W Polsce Hyundai Motor Poland promował wizję przyszłości przy wykorzystaniu nowoczesnego roboto psa firmy Boston Dynamic. Prezentację robota odbyły się tylko w salonach z samochodami elektrycznymi i hybrydowymi marki Hyundai. W trójmieście taki pokaz odbył się w Gdańsku 14 stycznia 2022r. Cena takiego robo psa wynosiła 74 500 dolarów, a firma sprzedała około 400 jednostek do tamtego czasu. Robot wyposażony w ramie potrafi otwierać drzwi, podnosić przedmioty z ziemi i odkręcać butelki. Spot nie tylko nadaję się w rolnictwie co odzwierciedla akcja, gdzie Spot pod koniec 2021 roku został umieszczony w Czarnobylu na Ukrainie i zmierzył on poziom promieniowania i pola elektromagnetycznego w całym obiekcie i wygenerował mapę 3D.

Robot wykonujący zabiegi pielęgnacji gleby. Roboty Oz i Dino skonstruowane przez francuski start-up Naio Technologies mają za zadanie wspierać rolników, głównie w walce z chwastami.

Robot Oz został stworzony z myślą o niedużych gospodarstwach zajmującym się produkcją warzyw, również pod osłonami. Robot posiada napęd elektryczny i sterowany jest z wykorzystaniem nawigacji RTK GPS oraz czujników, które rozpoznają położenie upraw oraz umożliwiają manewrowanie. Robot jest stworzony do odchwaszczania.

Rysunek 28 Zdjęcie. Robot do pielęgnacji gleby Źródło: https://agtecher.com/pl/produkt/naio-oz/

Robot Dino został zaproponowany dla większych obszarowo farm. Maszyna ma 2,5 metra długości, wszystkie koła ma skrętne, co zapewnia mu znakomitą manewrowość. Wydajność dzienna konstrukcji wynosi 3 do 5 hektara. Robot jest sterowany z użyciem nawigacji RTK GPS oraz kamery wideo z analizą obrazu. Robot Dino jest wielofunkcyjny, w zależności do przymocowanego narzędzia może odchwaszczać, bronować, siać, pielić, sadzić i wykonywać bruzdy.

Rysunek 29 Zdjęcie. Maszyna wielofunkcyjna Dino Źródło: https://agrihandler.pl/naio/

Szwajcarska firma Ecorobotix opracowała autonomicznego robota do pielenia. Ten robot jest laureatem prestiżowego konkursu SIMA Innovations Awards, który nagradza najbardziej innowacyjne technologie w danym roku.

AVO posiada specjalną kamerę do wykrywania roślin oraz układ sensoryczny, który służy mu do nawigacji po polu. Można sterować nim przy pomocy telefonu, jednak człowiek nie jest potrzebny zbyt często, poniewasz robot radzi sobie sam bez najmniejszych porblemów. Robota AVO odrużnia od innych tego typu maszyn to, że jest w pełni zależny od energii słonecznej, dzięki panelom fotowoltanicznym zamontowanym na dachu robot jest w stanie pracować do 12 godzin dziennie bez przerwy i wyplewić nawet 10 hektarów pola. Posiada on też baterię, dzięki którym może on pracować równiesz i w nocy.

Rysunek 30 Zdjęcie. Robot do pielęgnacji gleby Źródło: https://agtecher.com/pl/produkt/avo-by-ecorobotix/

6. Podsumowanie

Artykuł prezentuje nowoczesne technologie wykorzystywane w rolnictwie, koncentrując się na rolnictwie precyzyjnym oraz roli robotów w pielęgnacji i zbiorze roślin. Dzięki narzędziom takim jak GPS, satelitarne monitorowanie roślinności oraz roboty, możliwe jest dokładne dostosowanie zabiegów agrotechnicznych do indywidualnych potrzeb obszaru uprawnego. Automatyzacja prac rolniczych za pomocą robotów, wyposażonych w sztuczną inteligencję i zaawansowane systemy wizyjne, przyczynia się do zwiększenia wydajności produkcji oraz redukcji kosztów. Ponadto, artykuł omawia wykorzystanie Internetu Rzeczy (IoT) w rolnictwie, co pozwala na monitorowanie warunków upraw i optymalizację procesów produkcyjnych. Wnioski wskazują na potencjał tych technologii do poprawy efektywności produkcji rolniczej oraz ograniczenia negatywnego wpływu na środowisko, jednak ich wdrożenie wymaga odpowiedniej wiedzy i inwestycji.

Źródła internetowe

1. https://www.sadyogrody.pl/ (dostęp: 28.04.2024).
2. https://www.sadyogrody.pl/logistyka_i_opakowania/107/roboty_do_zbioru_owocow_sprawdzamy_nowe_technologie,31205.html (dostęp: 28.04.2024).
3. https://www.sadyogrody.pl/logistyka_i_opakowania/107/internet_rzeczy_w_rolnictwie_jakie_ma_zastosowanie_w_rolnictwie,31485.html (dostęp: 28.04.2024).
4. https://www.abc.net.au/news/rural/2022-06-16/robot-fruit-picker-eve-farm-worker-shortage/101150514 (dostęp: 28.04.2024).
5. https://www.smh.com.au/technology/could-robots-be-coming-to-the-places-that-grow-your-food-20220919-p5bjcn.html (dostęp: 28.04.2024).
6. https://www.riperobotics.com/ (dostęp: 28.04.2024).
7. https://www.sadyogrody.pl/logistyka_i_opakowania/107/internet_rzeczy_jakie_ma_zastosowanie_w_rolnictwie_cz_ii,31863.html (dostęp: 28.04.2024).
8. https://www.tvmaster.eu/artykul-4media/14531,innowacyjne-technologie-w-rolnictwie
9. https://www.cenyrolnicze.pl/ (dostęp: 28.04.2024).
10. https://mrjagrotim.pl/pl/aktualnosci/34/rolnictwo-precyzyjne-czym-jest-i-jakie-daje-korzy (dostęp: 28.04.2024).
11. https://tygodnik.pl/pl/639_materialy-partnera/9213_innowacyjne-technologie-w-rolnictwie.html (dostęp: 28.04.2024).
12. https://sagarobotics.com/ (dostęp: 28.04.2024).

13. https://www.futurefarming.com/tech-in-focus/field-robots/robot-combats-powdery-mildew-in-strawberries-using-uv-c-treatment/ (dostęp: 28.04.2024).
14. https://kobietawsadzie.pl/sam-zrywa-i-pakuje-jablka-zobacz-robota-przyszlosci/(dostęp: 28.04.2024).
15. https://www.deere.pl/pl/rozwi%C4%85zania-rolnictwa-precyzyjnego/systemy-prowadzenia-i-automatycznego-sterowania/machinesync/ (dostęp: 28.04.2024).
16. https://www.cenyrolnicze.pl/wiadomosci/technika-rolnicza/29420-rolnictwo-precyzyjne-wiedza-to-zysk (dostęp: 28.04.2024).
17. https://mrjagrotim.pl/pl/aktualnosci/34/rolnictwo-precyzyjne-czym-jest-i-jakie-daje-korzy (dostęp: 28.04.2024).

Katarzyna Maternia, Sławomir Pareniak, Jakub Bocek, Patryk Krupa, Piotr Dubaj
Koło naukowe Elektroniki i Technologii Informacyjnych

dr inż. Bartosz Pawłowicz
Opiekun Koła Naukowego

Technologie bezprzewodowe Bluetooth i Wi-Fi: wpływ, zastosowania i przyszłość

Streszczenie

Tekst omawia dwie popularne technologie komunikacji bezprzewodowej: Bluetooth i Wi-Fi. Bluetooth jest przedstawiony jako standardowa technologia do krótkich połączeń bezprzewodowych, umożliwiając szybkie i wygodne przesyłanie danych między różnymi urządzeniami na niewielkie odległości. Z kolei Wi-Fi to sieć bezprzewodowa, która zapewnia szybsze połączenia na większe odległości i jest częściej stosowana do dostępu do Internetu oraz sieci lokalnych.

W tekście poruszono historię i działanie technologii Bluetooth, omówiono zalety i zastosowania, w tym różnice między Bluetooth klasycznym a Bluetooth Low Energy. Przedstawiono również rozwój standardów sieci Wi-Fi, począwszy od 802.11b aż do najnowszej generacji, czyli Wi-Fi 7. Opisano również zabezpieczenia sieci Wi-Fi oraz porównano Bluetooth i Wi-Fi pod względem zasięgu, prędkości transferu, zastosowań i efektywności energetycznej.

Słowa kluczowe: Wi-Fi, Bluetooth, sieć, bezpieczeństwo, technologie bezprzewodowe

1. Wprowadzenie

Jedna z powszechnych technologii w dzisiejszym świecie jest Bluetooth, dzięki niemu możemy łączyć urządzenia bez konieczności uciążliwego stosowania kabli. Bluetooth jest technologią bezprzewodowej komunikacji krótkiego zasięgu pomiędzy różnymi urządzeniami elektronicznymi, takimi jak klawiatura, laptop czy komputer. Systemy Bluetooth pracują w paśmie o częstotliwości 2,4GHz, które są przydzielone do wykorzystania przez systemy przemysłowe, medyczne i naukowe.

Bluetooth jest protokołem komunikacji bezprzewodowej małej mocy, który funkcjonuje na falach radiowych. Bluetooth jest dziś standardowym elementem wyposażenia większości urządzeń RTV oferowanych na rynku. Komunikacja Bluetooth jest również wykorzystywana w telefonach komórkowych aby posłuchać muzyki na słuchawkach lub zewnętrznych głośnikach, można przesyłać dźwięk bezprzewodowo, bez konieczności sięgania po przewód audio z ewentualnymi przejściówkami.

WiFi (wireless fidelity) to grupa standardów radiowej sieci bezprzewodowej, które zostały stworzone przez firmę Wi-Fi Alliance. Powszechnie Wi-Fi jest to technologia WLAN

umożliwiająca założenie lokalnej sieci, poprzez połączenie urządzeń bez zastosowania kabli . Wi-Fi obecnie jest najpopularniejszym z dostępnych na rynku produktów WLAN.

2. Czym jest, historia i jak działa technologia Bluetooth

Bluetooth jest standardem komunikacji bezprzewodowej, który umożliwia wymianę danych miedzy urządzeniami znajdującymi się w bliskim zasiągu. Ta technologia pozwala nam na łączenie naszych telefonów, laptopów, słuchawek, głośników i wiele więcej różnych urządzeń, by szybko i wygodnie przesłać informacje.

Technologia Bluetooth powstała w latach 90. XX wieku, gdyż firma Ericsson zaczynała prace nad stworzeniem standardu bezprzewodowego łączenia urządzeń. Nazwa technologii Bluetooth pochodzi od imienia Haralda Blåtanda, który był królem Dani i Norwegii w X wieku. Zjednoczył on te dwa królestwa podczas swojego panowania, dlatego jest technologia Bluetooth, która tak samo ma jednoczyć różne urządzenia w jednym standardzie komunikacji.

Do komunikacji miedzy urządzeniami technologia Bluetooth wykorzystuje fale radiowe. Każde urządzenie posiadające technologie Bluetooth posiada specjalny chip, który umożliwia mu nadanie i odbieranie sygnałów radiowych w określonym zakresie częstotliwości. Kiedy dwa różne urządzenia chcą się połączyć, muszą mieć wspólny klucz szyfrowania, przez co gwarantuje to bezpieczną komunikację.

3. Zalety Bluetooth

Technologia Bluetooth ma dość dużo zalet, dzięki którym stała się tak popularna. Oto przykładowe z nich:

- Możliwość komunikacji miedzy urządzeniami bez potrzeby używania kabli.
- Standard Bluetooth jest powszechny i wspierany przez wiele urządzeń, zarówno na rynku elektroniki przemysłowej, jak i użytkowej.
- Technologia Bluetooth jest zoptymalizowana pod kątem oszczędzania energii, co jest szczególnie ważne w odniesieniu do urządzeń mobilnych.
- Parowanie urządzeń Bluetooth jest dość prostym i szybkim procesem połączenia urządzeń.
- Urządzenia Bluetooth są wielofunkcyjne, umożliwiają rożnego rodzaju połączenia takie jak transmisja dźwięku, przesyłanie danych czy sterowanie zdalne.

4. Rodzaje i zastosowania technologii Bluetooth

Bluetooth klasyczne (Bluetooth 1.0 – 3.0) to pierwsza generacja technologii, która umożliwia przesyłanie danych w niewielkich ilościach. Mimo dużych ograniczeń, te wersje

Bluetooth odegrały kluczową rolę w rozwoju bezprzewodowych słuchawek i głośników. Bluetooth Classic, znany także jako Bluetooth Basic Rate/Enhanced Data Rate, to technologia radiowa o niskim zużyciu energii, która przesyła dane przez 79 kanałów w nielicencjonowanym paśmie częstotliwości 2,4GHz ISM. Głównie jest wykorzystywany do bezprzewodowego przesyłania dźwięku, co staje się standardem dla bezprzewodowych słuchawek, głośników oraz dla systemów rozrywki w samochodach. Bluetooth klasyczny umożliwia przesyłanie danych w zastosowaniach takich jak drukowanie mobilne.

Bluetooth Low Energy (Bluetooth 4.0+) jest również znane jako Bluetooth Smart, został stworzony z myślą o urządzeniach niskoprądowych. Idealnie nadaje się do monitorowania zdrowia, noszonych urządzeń i innych zastosowań, gdzie dość ważna jest oszczędność energii. Obsługuje wiele topologii komunikacyjnych, od jednego do drugiego punktu, przez broadcast, po najnowszą mesh, co pozwala na tworzenie niezawodnych, dużych sieci urządzeń. Bluetooth LE jest dość szeroko stosowany również jako technologia pozycjonowania urządzeń, dzięki czemu zaspokaja rosnące zapotrzebowanie na usługi lokalizacji wewnętrznej z wysoką dokładnością.

Tabela 11: Tabela porównawcza

Cecha	**Bluetooth Low Energy (LE)**	**Bluetooth Classic**
Pasmo Częstotliwości	2.4GHz ISM Band (2.402 – 2.480 GHz)	2.4GHz ISM Band (2.402 – 2.480 GHz)
Kanały	40 kanałów z odstępem 2 MHz (3 reklamowe/37 danych)	79 kanałów z odstępem 1 MHz
Wykorzystanie Kanałów	FHSS	FHSS
Modulacja	GFSK	GFSK, π/4 DQPSK, 8DPSK
Szybkość Transmisji Danych	Do 2 Mb/s	Do 3 Mb/s (EDR)
Moc Nadajnika	≤ 100 mW (+20 dBm)	≤ 100 mW (+20 dBm)
Czułość Odbiornika	Do ≤-82 dBm	≤-70 dBm
Transporty Danych	Różne, w tym asynchroniczne i izochroniczne	Asynchroniczne i synchroniczne, orientowane na połączenie
Topologie Komunikacji	Punkt do punktu, broadcast, mesh	Punkt do punktu (w tym piconet)

Źródło:https://goodaudio.pl/blog/bluetooth-wszystko-co-warto-wiedziec/

Bluetooth 5.0 wprowadziło znaczne ulepszenia w zakresie zasięgu, prędkości przesyłania danych i pojemności jaka jest na łączu. Pozwala to na bardziej stabilne połączenia i obsługę większej liczby urządzeń jednocześnie.

Możemy zauważyć, że technologia Bluetooth znalazła zastosowania w różnych dziedzinach takich jak:

- Słuchawki bezprzewodowe i zestawy głośnomówiące dla bezprzewodowego słuchania muzyki oraz prowadzenia rozmów telefonicznych.
- Głośniki przenośne Bluetooth dzięki nim można odtwarzać bezprzewodowo muzykę z telefonów czy innych urządzeń.
- Smartfony i telefony komórkowe, używają Bluetooth do połączenia się z innymi urządzeniami, przesyłaniem danych czy udostępnianiem Internetu.
- Samochodowe systemy audio używają tej technologii do połączeń z telefonem i bezprzewodowego przesyłania informacji o trasie.
- Myszki i klawiatury bezprzewodowe dla dużo bardziej wygodniejszej pracy bez konieczności używania kabli.
- Smartwatche i opaski fitness do synchronizacji danych w telefonie i monitorowanie aktywności fizycznej.
- Kontrolery do gier do bezprzerwowego sterowania grami na różnych komputerach i konsolach.
- Drukarki i skanery umożliwiają bezprzewodowy druk i skanowanie dokumentów.
- Urządzenia inteligentnego domu takie jak termostaty, oświetlenia czy systemy alarmowe, które dzięki Bluetooth można kontrolować zdalnie.

5. Wpływ Bluetooth na nasze życie codzienne i bezpieczeństwo

Bluetooth stał się naszym codziennym życiem, częścią integralną. Technologia Bluetooth ułatwia prowadzenie rozmów telefonicznych w samochodzie, pozwala na słuchanie ulubionej muzyki bezprzewodowo oraz umożliwia monitorowanie aktywności fizycznej bez noszenia dodatkowych urządzeń. Technologie Bluetooth mają istotny aspekt bezpieczeństwa. Dzięki mechanizmowi szyfrowania i autoryzacji, ryzyko nieuprawnionego dostępu jest dużo mniejsze wręcz minimalne. Jednakże warto zachować ostrożność i unikać łączenia się z nieznanymi urządzeniami. Im wyższy standard przesyłania danych, tym jest łatwiej i szybciej wyłapać urządzenia, które chcą się połączyć z naszym sprzętem. Warto pamiętać, że zawsze można

zabezpieczyć się przed niechcianym bezprzewodowym kontaktem, poprzez ustawienie naszych urządzeń jako niewidoczne poza zaufanym kręgiem.

6. Porównanie technologii Bluetooth z NFC

NFC ma bardzo ograniczony zasięg, który ma zaledwie kilka centymetrów, w przeciwieństwie do technologii Bluetooth. Łączność połączeń, dzięki NFC umożliwia szybkie i łatwe parowanie urządzeń przez dotknięcie, kiedy Bluetooth wymaga kilku kroków do sprawowania z innym urządzeniem. Zastosowanie NFC zazwyczaj jest stosowane w płatnościach mobilnych i identyfikacji, natomiast Bluetooth służy do przesyłania danych i łączności urządzeń.

7. Rozwój standardów cyfrowej komunikacji bezprzewodowej Wifi

Standard sieci wifi 802.11 wprowadzony został w latach 90 ubiegłego wieku. Pomysł przeniesienia komunikacji sieciowej wydawał się trudny ze względu na ograniczenia techniczne. Ówczesna elektronika miała trudności z wydajną pracą na samych skrętkach a sama technologia przesyłania drogą radiową dużych ilości danych wydawała się nie możliwa. Mimo to inżynierowie dążyli do wizji przyszłości bez okablowania a przesyłania informacji drogą radiową. Między innymi ich koncepcji narodził się standard sieci IEEE 802.11b wprowadzony w 1999 roku czyli innymi słowy Wi-Fi. Była to pierwsza popularna wersja radiowej sieci LAN która teoretycznie umożliwiała maksymalną prędkość transmisji na poziomie 11 Mb/s lecz w praktyce było to nie możliwe. Ówczesne Wi-Fi korzystało jedynie z częstotliwości 2,4 Ghz i modulacji DSSS. Niestety w tamtym okresie częstotliwość 5Ghz nie zdobyła popularności poza niektórymi specyficznymi zastosowaniami w wersji a.

Kolejną wersją która zyskała dużą popularność była sieć IEEE 802.11g która była kompatybilna wstecz z poprzednią wersją b. Swój przełom uzyskała na początku 2003 roku w momencie popularyzacji Internetu i komputerów przenośnych. Ten standard pozwalał już na wyższą szybkość transmisji do 54 Mb/s. Jednakże rzeczywiste wartości wynosiły do kilkunastu Mb/s z kolei czego sieć bezprzewodowa była jedynie 8-krotnie wolniejszą siecią od popularnego w tym czasie Ethernetu przewodowego. Co zaspokajało ówczesnych użytkowników laptopów.

Następna wersja Wi-Fi ukazała się w 2009 roku była to sieć IEEE 802.11n wprowadzała ona wiele nowości przez co zyskała duży sukces i popularność. Nowy wariant sieci umożliwiał transfer na poziomie do 300 Mb/s lecz w rzeczywistości realne przesyłanie przy dobrych warunkach uzyskiwano do 100 Mb/s. Nowa technologia zwiększała zasięg transmisji, dzięki

czemu komunikacja radiowa w sieci lokalnej stała się dużo bardziej niezawodny. Producenci elektroniki wiedzieli już że przyszłość komunikacji będzie związana z technologią bezprzewodową. Gdy technologia była wprowadzana na rynku było już wiele telefonów komórkowych które posiadały wsparcie Wi-Fi przez co sieć IEEE 802.11n była bardzo wyczekiwana i popularna w swoim złotym okresie.

Standard sieci 802.11ac bezprzewodowej technologii sieciowej wprowadzony został w roku 2013 przewidywał większe prędkości od poprzedników oraz obsługiwał sieć 5 Ghz. Nowa technologia pozwalała na transfer niemalże 3,5 Gb/s ale jak w poprzednich przypadkach nie były to realnie możliwe wartości do uzyskania w normalnych warunkach. Nowy standard pozwalał już graczom oraz na bez problemowe transmitowanie swoich rozgrywek w Internecie co jeszcze bardziej spopularyzowało nową technologie

Wersja sieci 802.11ax została wprowadzona w 2019 roku oferowała i oferuje teoretyczną przepustowość przekraczającą 10 Gb/s. Standard sieci jest najnowszym etapem rozwoju tej technologii. Bazuje on na technologii 802.11ac czyli Wi-Fi 5 oferując wzrost prędkości i przepustowości sieci.

Rozwój sieci na przestrzeni lat			
Rok	Generacja sieci	Nazwa sieci	Maksymalna przepustowość
1999	802.11b	Wi-Fi 1	11 Mb/s
1999	802.11a	Wi-Fi 2	54 Mb/s
2003	802.11g	Wi-Fi 3	54 Mb/s
2009	802.11n	Wi-Fi 4	600 Mb/s
2014	802.11ac	Wi-Fi 5	1 Gb/s
2019	802.11ax	Wi-Fi 6	10 Gb/s
----	802.11be	Wi-Fi 7	46 Gb/s

Rysunek 31 Historia standardów Wi-Fi, opracowanie własne

8. Specyfikacja Wi-Fi 6

To standard generacji 6 oparty na grupie bezprzewodowych protokołów 802.11 ax. Jest to sieć bezprzewodowa stanowiąca w dzisiejszych czasach podstawowe łącze z którego z którego korzysta bardzo wielu użytkowników na swoich urządzeniach. Częstotliwość sieci jest skupiona na dwóch pasmach 2,4 GHz która jest odpowiedzialna za lepszy zasięg Wi-Fi w naszym domu lub biurze oraz pasmo 5,2 GHz oferująca wyższą szybkość i stabilność na naszych urządzeniach. Maksymalna rzeczywista prędkość połączenia dla Wi-Fi 6 osiąga 6Gb/s gdy osiągi poprzednika graniczą zaledwie 1,3 Gb/s. Jest to poprawa maksymalnych osiągów

względem Wi-Fi 5 (sieci 802.11ac) o ok. 360%. Możliwości technologii MIMO zostały zwiększone do 8, co umożliwia lepsze wykorzystanie przestrzeni radiowej. Dodatkowo, wykorzystanie techniki beamformingu pozwala na bardziej precyzyjne dostosowanie sygnału radiowego do warunków danego połączenia. Dodatkowo warto wspomnieć o wprowadzeniu mechanizmów zmniejszających zakłócenia w sieci, co daje nam większą swobodę z korzystania z urządzeń. Więc podsumowując główne zalety i innowacje wprowadzone dla Wi-Fi 6 względem poprzednika to:

- ulepszenie fizycznej przepływności medium, dzięki zastosowaniu modulacji OFDMA,
- ulepszenie mechanizmów umożliwiających sterowanie mocą sygnału,
- połączenie jest znacznie bezpieczniejsze,
- zastosowanie MU-MIMO.

9. Wi-Fi 7 Przyszłość technologii

Jest to następna przyszła generacja sieci bezprzewodowych która zakłada bardzo wysoką przepustowości nazywana standardem sieci IEEE 802.11be. Na ten moment planowane jest udostępnienie nowego zakresu częstotliwościowego dla najnowszej technologii przez Federalną Komisje Łączności FCC. Przewiduje się zapewnienie spektrum pomiędzy 5.925 i 7.125 GHz oczywiście wliczając powszechnie używane 2.4 GHz oraz 5 GHz, z których aktualnie korzystamy. Nowe pasmo zapewni przyszłej innowacji 1200 MHz dodatkowej szerokości a praca w wielu zakresach częstotliwości da możliwość obsługi większej liczby użytkowników i większego zagęszczenia punktów dostępowych. Kolejną ważną zaletą i przewagą nad poprzednikami będzie 5 krotnie szybsza prędkość połączenia która ma osiągnąć 46 Gb/s. Przyszła technologia ma zapewnić 100x mniejsze opóźnienie co zauważalnie poprawi jakość w czasie rzeczywistym wideokonferencji oraz gier video. Dla dalszego zwiększenia maksymalnych prędkości, Wi-Fi 7 wprowadza zaawansowany schemat modulacji o nazwie 4096-QAM. Ten nowy schemat pozwala na przenoszenie 12 bitów w każdym symbolu, w przeciwieństwie do 10 bitów, co z kolei przekłada się na teoretyczny wzrost przepustowości o 20% w porównaniu do Wi-Fi 6 z 1024-QAM. Dzięki temu użytkownicy będą mogli zauważyć większą wydajnością w przesyłaniu danych. W niedługim czasie użytkownicy będą mieli możliwość płynnego oglądania filmów w rozdzielczości 4K/8K, czy też przesyłania transmisji na żywo w bardzo wysokiej rozdzielczości.

10. Zabezpieczenia sieci Wi-Fi

-Rozgłaszanie sieci

Pierwszym pomysłem na zabezpieczenie sieci Wi-Fi było ukrycie SSID poprzez wyłączenie jego rozgłaszania przez punkt dostępowy (AP). Głównym celem było utrudnienie wykrycia sieci, co miało uniemożliwić atakującemu podłączenie się do niej bez znajomości SSID. Niestety metoda jest całkowicie nieskuteczna ponieważ SSID i tak jest rozgłaszane w odpowiedzi na żądanie dowolnego klienta lub może być przechwycone przy użyciu snifferów sieciowych.

-Filtrowanie dostępu

Kolejnym zabezpieczeniem sieci Wifi jest filtrowanie urządzeń, które mają dostęp do komunikacji z AP poprzez sprawdzanie tylko wybranych adresów MAC. Jednak problem powraca do charakterystyki medium transmisyjnego, gdzie adresy MAC są widoczne publicznie w najniższej warstwie komunikacji sieciowej. W rezultacie to zabezpieczenie jest łatwo pokonywane poprzez podsłuchanie transmisji i identyfikację adresów MAC urządzeń, które były w komunikacji z punktem dostępowym (AP).

-Izolacja

Kolejnym rozwiązaniem zabezpieczenia sieci jest izolacja użytkowników sieci pomiędzy sobą. Dzięki temu podłączone urządzenia do punktu dostępowego AP, nie mogą komunikować się ze sobą bez jej zgody. Skutkuje to tym, iż nie ma możliwości podsłuchiwania transmisji. Niestety metoda ta też nie jest w 100% skuteczna gdyż wystarczy przestawić swoją kartę bezprzewodową tak aby odbierała wszystkie dostępne pakiety i uzyskuje się możliwość podglądania transmisji innego użytkownika.

-Szyfrowanie

Bardziej skuteczną metodą zabezpieczenia sieci od pozostałych jest szyfrowanie punktu dostępowego AP i komunikacji sieciowej co znacząco poprawia bezpieczeństwo przed osobami niepożądanymi. Jest to zabezpieczenie wymagające podania hasła w celu połączenia się do sieci, a wszystkie dane są następnie szyfrowane. Pierwszym tego typu zabezpieczeniem był mechanizm WEP który posiadał niestety poważne wady kryptograficzne. Co w kolejnych latach zostało rozwiązane przez wprowadzenie nowego mechanizmu WPA mające naprawić poprzednie błędy. Niestety nadal sieci nie są 100 procentowo bezpieczne i do tej pory istnieją różnego rodzaju ataki na sieci zabezpieczone przy pomocy WPA.

-Rozwiązanie Enterprise

Jest to utworzenie w pełni wyizolowanej od LAN, sieci Wi-Fi z otwartym dostępem. Przez co każdy użytkownik chcący uzyskać dostęp powinien zalogować się poprzez stronę www, na której wprowadza login i hasło oraz łączy się z Internetem po czym uzyskuje dostęp do zasobów sieci LAN poprzez VPN. W tym przypadku sieć jest całkowicie odizolowana i nie ma możliwości przejścia z jednej do drugiej sieci. VPN w tym przypadku sprawia, że osoby chcące zaatakować naszą sieć nie będą w stanie jej zdekodować

11. Bluetooth a Wi-Fi, porównanie

Bluetooth i Wi-Fi to dwie różne technologie komunikacji bezprzewodowej. Bluetooth bardziej odpowiedni jest do krótszych dystansów i urządzeń osobistych, natomiast Wi-Fi zapewnia szybsze i lepsze połączenie na większą odległość. Zasięg Wi-Fi oferuje większy zasięg zwykle do 100 metrów w porównaniu do Bluetooth, który ma zwykle od 10 do 30 metrów. Natomiast prędkość transferu Wi-Fi zapewnia wyższą prędkość transmisji danych, która jest idealna do przesyłania dużych plików i streamingów wideo. Podczas gdy technologia Bluetooth jest bardziej używana do mniejszej ilości danych, takich jak audio. Bluetooth, zwłaszcza w wydaniu Low Energy, jest dużo bardziej energooszczędny niż technologia Wi-Fi. Zastosowanie Wi-Fi jest zazwyczaj preferowane do dostępu do Internetu i sieci lokalnych, podczas gdy Bluetooth jest znacznie lepszy do łączenia urządzeń peryferyjnych i przesyłania danych w krótkich dystansie czasowym.

Choć zarówno Bluetooth, jak i Wi-Fi do komunikacji bezprzewodowej wykorzystują fale radiowe, to te technologie bardzo się od siebie różnią. W przypadku sieci Wi-Fi mamy zazwyczaj do czynienia z hostami a dużej mocy, dzięki której mogą się łączyć urządzenia zewnętrzne i przekazywać między sobą duże ilości danych. Ponadto łączność Wi-Fi zapewnia większą przepustowość danych, jest bezpieczniejsza oraz ma większy zasięg fizyczny i może być stosowana do większej ilości użytkowników. Natomiast połączenia Bluetooth potrzebują niższego poboru mocy. Są wolniejsze i bardziej używane do szybkiego, prostszego i bezpośredniego parowania urządzeń zamiast aplikacji a stałej lokalizacji.

12. Podsumowanie

Artykuł pokazuje, jak Bluetooth i Wi-Fi zmieniły sposób, w jaki łączymy się i komunikujemy z naszymi urządzeniami. Bluetooth stał się nieodłączną częścią naszego codziennego życia, umożliwiając bezprzewodowe słuchanie muzyki, prowadzenie rozmów telefonicznych oraz kontrolowanie urządzeń elektronicznych. Z kolei sieci Wi-Fi zapewniają

nam dostęp do Internetu i sieci lokalnych w domu, biurze i innych miejscach publicznych, umożliwiając szybkie przesyłanie danych i strumieniowanie treści.

Obie technologie mają swoje zalety i zastosowania, które sprawiają, że są niezwykle wartościowe w dzisiejszym świecie cyfrowym. Ich ciągły rozwój i udoskonalenia, takie jak Bluetooth Low Energy czy standardy Wi-Fi 6, pozwalają nam cieszyć się coraz lepszymi połączeniami bezprzewodowymi o większej prędkości, zasięgu i bezpieczeństwie.

Wraz z postępem technologicznym możemy oczekiwać, że zarówno Bluetooth, jak i Wi-Fi będą kontynuować swoją ewolucję, zapewniając nam jeszcze lepsze doświadczenia komunikacyjne i połączenia bezprzewodowe w przyszłości.

Źródła internetowe:

1. https://botland.com.pl/blog/jak-dziala-bluetooth-zasieg-roznice/ (dostęp: 20.04.2024).
2. https://ep.com.pl/technologia/12501-ewolucja-standardow-cyfrowej-komunikacji-bezprzewodowej (dostęp: 20.04.2024).
3. https://lanster.com/wi-fi-5-wi-fi-6-czy-wi-fi-7-analiza-porownawcza/(dostęp: 20.04.2024).
4. https://sekurak.pl/bezpieczenstwo-sieci-wi-fi-czesc-1/ (dostęp: 20.04.2024).
5. https://www.alget.pl/blog/aktualnosci/co-to-jest-bluetooth-wszystko-co-musisz-wiedziec (dostęp: 20.04.2024).
6. https://www.morele.net/wiadomosc/co-to-jest-wifi-6-najwazniejsze-informacje-i-szczegoly-techniczne/16378/ (dostęp: 20.04.2024).
7. https://www.netgear.com/pl/home/discover/wifi7/ (dostęp: 20.04.2024).
8. https://www.tophifi.pl/blog/post/jakie-sa-standardy-bluetooth.html (dostęp: 20.04.2024).

Piotr Laskowski, Maja Jaszowska, Dominika Fergisz
Koło Naukowe Elektroniki i Technologii Informacyjnych

dr. inż. Bartosz Pawłowicz
Opiekun Koła Naukowego

Adaptacja Anycubic Mega X na frezarkę CNC

Artykuł opisuje proces przekształcenia drukarki 3D Anycubic Mega X w frezarkę CNC do wyrobu płytek PCB, z wykorzystaniem firmware'u Klipper. Anycubic Mega X, ceniona za solidną konstrukcję i dużą przestrzeń roboczą, została wybrana do tej modyfikacji. Projekt obejmuje zmiany mechaniczne, takie jak wymiana głowicy drukującej na frezującą, oraz modyfikacje elektroniczne, w tym instalację nowego sterownika i firmware'u. Opisane etapy prac uwzględniają dostosowanie oprogramowania sterującego oraz wyzwania napotkane podczas realizacji projektu. Artykuł zawiera praktyczne wskazówki dotyczące wyboru narzędzi i materiałów, niezbędnych do przeprowadzenia modyfikacji. Przekształcenie Anycubic Mega X w frezarkę CNC umożliwia precyzyjne tworzenie płytek PCB, co otwiera nowe możliwości dla hobbystów i profesjonalistów zajmujących się elektroniką. Artykuł stanowi wartościowe źródło informacji dla każdego, kto chce zwiększyć funkcjonalność swojej drukarki 3D za pomocą nowoczesnego firmware'u Klipper.

Słowa kluczowe: Frezarka CNC, Drukarka 3D, Precyzyjne frezowanie

1. Wprowadzenie

W dzisiejszym erze dynamicznego rozwoju technologicznego, drukarki 3D stały się nieodłącznym elementem świata techniki, zarówno dla hobbystów, jak i profesjonalistów. Jednakże, ich potencjał można jeszcze bardziej poszerzyć poprzez modyfikacje, które przekształcają je w zaawansowane narzędzia. W niniejszym artykule skupimy się na przerobieniu drukarki 3D Anycubic Mega X na frezarkę CNC, specjalnie przystosowaną do produkcji płytek PCB (Printed Circuit Boards). Naszym celem jest przedstawienie procesu adaptacji, głównych kroków oraz korzyści, jakie niesie za sobą ta transformacja.

Nasze badania koncentrują się na metodach mechanicznych i elektronicznych, które umożliwiają przekształcenie drukarki 3D w zaawansowane narzędzie do produkcji PCB. Poprzez wykorzystanie innowacyjnego firmware'u Klipper oraz precyzyjnych modyfikacji mechanicznych, nasza modyfikacja stawia sobie za cel zapewnienie użytkownikom możliwości tworzenia wysokiej jakości płytek PCB w domowym zaciszu. Artykuł zawiera praktyczne wskazówki, które mogą być przydatne dla osób zainteresowanych rozwojem w dziedzinie elektroniki oraz produkcji DIY.

2. Budowa Anycubic Mega X: Kluczowe Komponenty

Anycubic Mega X to zaawansowana drukarka 3D, która wyróżnia się solidną konstrukcją i dużą przestrzenią roboczą. Aby przekształcić tę drukarkę w frezarkę CNC do wyrobu płytek PCB, konieczne jest zrozumienie jej budowy oraz kluczowych komponentów.

Kluczowe komponenty Anycubic Mega X:

- Rama i Konstrukcja: Anycubic Mega X posiada mocną i stabilną ramę wykonaną z aluminium, co zapewnia precyzję i wytrzymałość podczas pracy. Sztywna konstrukcja jest kluczowa dla utrzymania dokładności zarówno w druku 3D, jak i w obróbce CNC.
- Ekstruder i Głowica Drukująca: Ekstruder to mechanizm, który podaje filament do głowicy drukującej. Głowica drukująca odpowiedzialna jest za topienie i nakładanie filamentu warstwa po warstwie. W przypadku konwersji na frezarkę CNC, zarówno ekstruder, jak i głowica drukująca zostaną usunięte i zastąpione przez wrzeciono frezujące, które będzie wykorzystywane do obróbki materiału.
- Stół Roboczy: Oryginalny stół roboczy drukarki, który służy do podtrzymywania drukowanych obiektów, zostanie zamieniony na stół frezarski. Nowy stół musi być solidny i dostosowany do mocowania materiałów, które będą frezowane.
- Elektronika i Sterowanie: Drukarka Anycubic Mega X jest wyposażona w zaawansowany system sterowania, który zarządza ruchem osi oraz temperaturą głowicy. W przypadku konwersji na frezarkę CNC, konieczne będzie dostosowanie elektroniki i zainstalowanie nowego firmware'u, takiego jak Klipper, który umożliwi precyzyjne sterowanie wrzecionem i innymi komponentami CNC.
- Oprogramowanie: Anycubic Mega X korzysta z dedykowanego oprogramowania do przygotowywania modeli do druku 3D. Po przekształceniu w frezarkę CNC, konieczne będzie użycie innego oprogramowania, takiego jak GRBL lub innego kompatybilnego z firmware'em Klipper, które pozwoli na sterowanie procesem frezowania.

3. Podobieństwa i różnice między drukarką 3D a frezarką CNC

Drukarki 3D i frezarki CNC to narzędzia, które rewolucjonizują produkcję prototypów i małoseryjną produkcję w dziedzinie technologii. Pomimo że oba narzędzia są używane do obróbki materiałów, istnieją fundamentalne różnice między nimi, które należy uwzględnić podczas adaptacji drukarki 3D na frezarkę CNC do wyrobu płytek PCB. W tym rozdziale omówimy cechy wspólne oraz różnice między drukarką 3D a frezarką CNC.

Cechy Wspólne:

Obróbka materiałów: Zarówno drukarki 3D, jak i frezarki CNC, są używane do obróbki materiałów, ale w różny sposób. Drukarki 3D budują obiekty warstwa po warstwie, podczas gdy frezarki CNC usuwają materiał z bloku, tworząc pożądany kształt.

Sterowanie numeryczne: Obie maszyny wykorzystują technologię sterowania numerycznego (CNC), która umożliwia precyzyjne wykonywanie złożonych operacji obróbczych z wykorzystaniem komputerowego programu sterującego.

Automatyzacja procesu: Zarówno drukarki 3D, jak i frezarki CNC, oferują wysoki poziom automatyzacji procesu produkcyjnego, co prowadzi do zwiększenia efektywności i powtarzalności.

Różnice:

Metoda Produkcji: Drukarki 3D stosują metodę addytywną, gdzie materiał jest dodawany warstwa po warstwie, podczas gdy frezarki CNC stosują metodę subtraktywną, gdzie materiał jest usuwany z bloku.

Zastosowanie Materiałów: Drukarki 3D mogą wykorzystywać różnorodne materiały, takie jak plastik, metal, czy materiały biologiczne, podczas gdy frezarki CNC są zazwyczaj ograniczone do obróbki materiałów twardych, takich jak aluminium, stal, czy tworzywa sztuczne.

Zakres Zastosowań: Drukarki 3D są często stosowane do tworzenia prototypów, modeli i elementów złożonych, podczas gdy frezarki CNC są wykorzystywane do produkcji precyzyjnych detali, form, narzędzi oraz elementów maszyn.

Stopień Skomplikowania: Drukarki 3D są zazwyczaj prostsze w obsłudze i bardziej przyjazne dla początkujących, podczas gdy frezarki CNC wymagają większej wiedzy technicznej i doświadczenia w programowaniu i obsłudze.

4. Przygotowania drukarki Anycubic Mega X do konwersji

W tym rozdziale omówimy teoretyczne modyfikacje, które muszą zostać przeprowadzone, aby przekształcić drukarkę 3D Anycubic Mega X w frezarkę CNC. Skupimy się na krokach niezbędnych do usunięcia zbędnych elementów drukarki, dodania nowych komponentów oraz dostosowania elektroniki i oprogramowania do nowego zastosowania. Przedstawimy również ogólną listę narzędzi i materiałów potrzebnych do przeprowadzenia tych modyfikacji, podkreślając kluczowe aspekty każdej zmiany, aby osiągnąć założone cele projektowe.

4.1.Narzędzia i materiały

Przed przystąpieniem do modyfikacji drukarki Anycubic Mega X na frezarkę CNC, niezbędne jest przygotowanie odpowiednich narzędzi i materiałów. Właściwe wyposażenie i zasoby są kluczowe dla zapewnienia płynnego przebiegu konwersji oraz osiągnięcia zamierzonych rezultatów. Poniżej przedstawiono listę narzędzi i materiałów, które będą potrzebne:

Narzędzia:

Zestaw śrubokrętów: Różne rodzaje i rozmiary śrubokrętów (krzyżakowe, płaskie, torx) do demontażu i montażu komponentów.

Klucze imbusowe: Klucze o różnych rozmiarach do pracy z śrubami imbusowymi, które są często używane w konstrukcjach drukarek 3D.

Szczypce i kombinerki: Przydatne do trzymania i manipulacji małymi elementami oraz do cięcia przewodów.

Lutownica: Do lutowania przewodów i połączeń elektronicznych, które mogą wymagać modyfikacji.

Multimetr: Do pomiarów elektrycznych, weryfikacji połączeń oraz diagnozowania problemów elektronicznych.

Wkrętarka elektryczna: Ułatwia szybki montaż i demontaż licznych śrub.

Imadło i uchwyty: Do bezpiecznego trzymania komponentów podczas modyfikacji.

Materiały:

Wrzeciono CNC: Kluczowy element, który zastąpi głowicę drukującą, umożliwiający precyzyjną obróbkę materiału.

Stół frezarski: Nowa powierzchnia robocza, stabilna i odpowiednio przygotowana do mocowania obrabianych materiałów.

Elementy montażowe: Śruby, nakrętki, podkładki i inne elementy montażowe niezbędne do zamocowania nowych komponentów.

Przewody i złącza: Do połączeń elektrycznych pomiędzy nowymi komponentami a elektroniką drukarki.

Osłona ochronna: Materiały do budowy obudowy zabezpieczającej, chroniącej przed odłamkami i hałasem.

Odkurzacz do trocin: System odsysania trocin i pyłu, który powstaje podczas frezowania.

Firmware (Klipper): Oprogramowanie kontrolujące pracę frezarki CNC, które trzeba zainstalować i skonfigurować.

Przygotowanie odpowiednich narzędzi i materiałów jest kluczowe dla zapewnienia sprawnego przebiegu konwersji oraz osiągnięcia wysokiej jakości wyników. W następnych podrozdziałach omówimy szczegółowo poszczególne etapy modyfikacji, zaczynając od demontażu zbędnych komponentów drukarki, przez instalację nowych elementów, aż po dostosowanie elektroniki i oprogramowania.

4.2.Usuwanie oryginalnych komponentów

Następnym krokiem w procesie konwersji drukarki Anycubic Mega X na frezarkę CNC jest usunięcie zbędnych komponentów, które nie będą już potrzebne lub mogą zakłócać prawidłowe działanie urządzenia w nowym zastosowaniu. Poniżej przedstawiono listę komponentów, które należy usunąć:

Ekstruder: Głowica drukująca oraz elementy związane z ekstruderem, takie jak prowadnice i prowadnice filamentu, nie będą już potrzebne do frezowania.

Stół drukujący: Stół, na którym normalnie umieszcza się materiał drukowany, również będzie zbędny w przypadku frezowania i należy go usunąć.

Elektronika drukarki: Płyta główna pozostaje bez zmian, podobnie jak silniki krokowe i inne elementy sterujące, które są niezbędne do prawidłowego funkcjonowania urządzenia. Jednak elementy związane bezpośrednio z procesem drukowania, takie jak czujniki położenia i ogrzewanie stołu, należy usunąć.

Usuwanie tych komponentów wymaga ostrożności i precyzji, aby uniknąć uszkodzenia innych części drukarki. Należy również pamiętać o zachowaniu wszystkich elementów, które mogą być przydatne w przyszłości lub mogą być wykorzystane do innych projektów. Po dokładnym usunięciu zbędnych komponentów drukarka będzie gotowa do instalacji nowych elementów, niezbędnych do przekształcenia jej w frezarkę CNC.

4.3.Montaż nowych komponentów

Po usunięciu zbędnych elementów z drukarki Anycubic Mega X, kolejnym krokiem jest instalacja nowych komponentów niezbędnych do przekształcenia jej w frezarkę CNC. Poniżej przedstawiono listę komponentów, które należy zainstalować:

Wrzeciono CNC: Główny element roboczy frezarki, który zastępuje głowicę drukującą. Wrzeciono powinno być zamocowane w miejscu, gdzie wcześniej znajdował się ekstruder, zapewniając stabilność i precyzję obróbki.

Stół frezarski: Nowa powierzchnia robocza, stabilna i odpowiednio przygotowana do mocowania obrabianych materiałów. Stół musi być solidny i dobrze zamocowany, aby wytrzymać siły generowane podczas frezowania.

Elementy montażowe: Śruby, nakrętki, podkładki i inne elementy montażowe niezbędne do zamocowania nowych komponentów. Ważne jest, aby używać elementów montażowych o odpowiednich rozmiarach i wytrzymałości, aby zapewnić stabilność całej konstrukcji.

Przewody i złącza: Nowe przewody i złącza do połączeń elektrycznych pomiędzy wrzecionem a elektroniką drukarki. Należy upewnić się, że wszystkie połączenia są solidne i bezpieczne, aby zapobiec przerwom w zasilaniu lub sygnałach sterujących.

Osłona ochronna: Materiały do budowy obudowy zabezpieczającej, chroniącej przed odłamkami i hałasem. Osłona ochronna jest ważnym elementem bezpieczeństwa, który zapobiega wydostawaniu się trocin i pyłu oraz chroni użytkownika przed urazami.

Odkurzacz do trocin: System odsysania trocin i pyłu, który powstaje podczas frezowania. Odkurzacz powinien być odpowiednio zamocowany i skonfigurowany, aby efektywnie usuwać trociny z obszaru roboczego.

Firmware (Klipper): Oprogramowanie kontrolujące pracę frezarki CNC. Klipper musi zostać zainstalowany i skonfigurowany na płycie głównej drukarki, aby umożliwić precyzyjne sterowanie wrzecionem i innymi komponentami.

Instalacja tych komponentów wymaga staranności i dokładności, aby zapewnić prawidłowe działanie frezarki CNC. Po zakończeniu instalacji nowych elementów należy przeprowadzić dokładne testy, aby upewnić się, że wszystkie komponenty są prawidłowo zamocowane i działają zgodnie z oczekiwaniami. W kolejnych podrozdziałach omówimy szczegółowo procedury konfiguracji i kalibracji frezarki CNC.

4.4.Dostosowanie elektroniki i instalacja firmware’u

W procesie przekształcania drukarki 3D Anycubic Mega X na frezarkę CNC, kluczowe jest odpowiednie dostosowanie elektroniki oraz instalacja firmware'u, który będzie zarządzał nową konfiguracją urządzenia. Poniżej przedstawiono teoretyczne kroki, które należy wykonać, aby osiągnąć zamierzony cel:

Przygotowanie elektroniki: Przed rozpoczęciem pracy należy upewnić się, że płyta główna drukarki i silniki krokowe są w dobrym stanie. Wszystkie połączenia elektryczne muszą być solidne i bezpieczne.

Instalacja firmware: Kluczowym elementem jest instalacja odpowiedniego firmware'u, który będzie kontrolować pracę frezarki CNC. W tym projekcie rekomenduje się użycie firmware'u Klipper, który oferuje zaawansowane funkcje sterowania i umożliwia precyzyjne dostosowanie parametrów pracy urządzenia.

Konfiguracja plików konfiguracyjnych: Po zainstalowaniu Klippera, konieczne jest skonfigurowanie plików konfiguracyjnych. Parametry takie jak prędkości, przyspieszenia, limity pracy osi X, Y i Z oraz specyfikacje wrzeciona muszą być dokładnie określone, aby zapewnić prawidłowe działanie frezarki.

Podłączenie przewodów: Nowe komponenty, takie jak wrzeciono CNC, wymagają odpowiedniego podłączenia do elektroniki drukarki. Należy zastosować odpowiednie przewody i złącza, aby zapewnić niezawodne połączenia elektryczne.

Testowanie połączeń: Przed uruchomieniem urządzenia konieczne jest sprawdzenie wszystkich połączeń elektrycznych pod kątem zwarć i przerw. Miernik do sprawdzenia ciągłości obwodów może być przydatnym narzędziem w tym procesie.

Kalibracja elektroniki: Sterowniki silników krokowych muszą być skalibrowane, aby działały prawidłowo z nowymi komponentami. Ustawienia mikro-kroków i prądów silników powinny być dostosowane do wymagań nowej konfiguracji.

Integracja systemu odsysania trocin: Jeśli planowany jest system odsysania trocin, musi on być odpowiednio zintegrowany z elektroniką drukarki. Synchronizacja jego działania z frezarką jest kluczowa dla efektywnego usuwania odpadów z obszaru roboczego.

Sprawdzenie bezpieczeństwa: Wszystkie elementy elektroniczne muszą być odpowiednio zabezpieczone. Zainstalowanie osłon i obudów zabezpieczających przed uszkodzeniem i zanieczyszczeniami jest niezbędne dla bezpiecznej eksploatacji urządzenia.

Przeprowadzenie testów: Po zakończeniu konfiguracji elektroniki i instalacji firmware'u, należy przeprowadzić testy w celu weryfikacji poprawności działania wszystkich komponentów. Testy wrzeciona, silników krokowych oraz systemu odsysania trocin pomogą upewnić się, że urządzenie działa zgodnie z oczekiwaniami.

Dostosowanie elektroniki i instalacja firmware'u są kluczowymi krokami w przekształcaniu drukarki 3D w frezarkę CNC. Chociaż projekt jest w trakcie realizacji, teoretyczne omówienie tych procesów pozwala na lepsze zrozumienie wymagań technicznych i przygotowanie się do praktycznej realizacji. W kolejnych podrozdziałach zostaną omówione techniki frezowania oraz zastosowania frezarki CNC do wyrobu płytek PCB.

4.5.Testowanie i kalibracja wstępna

Przed przystąpieniem do pełnej konfiguracji i uruchomienia frezarki CNC, niezbędne jest przeprowadzenie testów oraz kalibracji wstępnej. Proces ten ma na celu upewnienie się, że wszystkie komponenty działają poprawnie i są gotowe do dalszej pracy. W tej części omówimy kroki niezbędne do testowania i kalibracji urządzenia przed właściwym rozpoczęciem obróbki.

Testowanie ruchu osi: Po uruchomieniu urządzenia w trybie testowym należy dokładnie obserwować ruch każdej osi. Warto sprawdzić, czy każda z osi porusza się płynnie, bez szarpnięć lub opóźnień, co może świadczyć o problemach z mechanicznym przesuwem lub konfiguracją elektroniczną.

Testowanie wrzeciona: Ważnym krokiem jest przetestowanie działania wrzeciona CNC pod kątem stabilności obrotów i precyzji pracy. W trakcie testów warto zwrócić uwagę na ewentualne wibracje

lub niepokojące dźwięki, które mogą wskazywać na konieczność dokładniejszej regulacji lub konserwacji wrzeciona.

Kalibracja prędkości i przyspieszeń: Dostosowanie parametrów prędkości i przyspieszeń ruchu osi do konkretnego zadania obróbki jest kluczowe dla uzyskania optymalnych wyników. W trakcie kalibracji należy uwzględnić zarówno wymagania materiału, jak i zdolności techniczne samego urządzenia.

Kalibracja wysokości wrzeciona: Poprawne ustawienie wysokości wrzeciona nad powierzchnią roboczą ma istotny wpływ na jakość i dokładność wykonanej obróbki. Dokładna kalibracja wysokości zapewnia minimalizację ryzyka kolizji oraz optymalne warunki pracy narzędzia.

Testowanie systemu odsysania trocin: Efektywny system odsysania trocin jest kluczowy dla utrzymania czystości w obszarze roboczym i zapobiegania gromadzeniu się odpadów. Przeprowadzenie testów pozwala ocenić skuteczność działania systemu i ewentualne konieczności jego regulacji lub ulepszeń.

Kalibracja mikro-kroków silników: Precyzyjne dostrojenie mikro-kroków silników krokowych pozwala na uzyskanie płynnego i precyzyjnego ruchu osi. W trakcie kalibracji warto zadbać o odpowiednie dobranie parametrów, aby zapewnić optymalną wydajność urządzenia przy minimalnym zużyciu energii.

Testy dokładności pozycjonowania: Przeprowadzenie testów dokładności pozycjonowania osi X, Y i Z pozwala ocenić precyzję i powtarzalność ruchów frezarki CNC. Dokładne sprawdzenie tej cechy jest kluczowe dla zapewnienia wysokiej jakości obróbki materiałów oraz uniknięcia błędów w produkowanych elementach.

Po przeprowadzeniu testów i kalibracji wstępnej, urządzenie powinno być gotowe do dalszej pracy. Zapewnienie poprawnego działania wszystkich komponentów oraz precyzyjne skalibrowanie parametrów ruchu i obróbki jest kluczowe dla osiągnięcia wysokiej jakości wyników. Kolejnym krokiem będzie już właściwa obróbka materiałów przy użyciu frezarki CNC.

5. Konfiguracja Oprogramowania i Parametrów Obróbki

W procesie przygotowania frezarki CNC do pracy istotną rolę odgrywa konfiguracja oprogramowania sterującego oraz dostosowanie parametrów obróbki. W tym rozdziale omówimy podstawowe kroki niezbędne do konfiguracji oprogramowania oraz doboru parametrów obróbki, które są istotne dla teoretycznego projektu.

5.1.Wybór i Konfiguracja Oprogramowania Sterującego.

Wybór odpowiedniego oprogramowania sterującego stanowi kluczowy krok podczas konfiguracji frezarki CNC. Oprogramowanie to pełni rolę interfejsu między użytkownikiem a maszyną, umożliwiając kontrolę nad ruchem osi, prędkością wrzeciona, oraz innymi parametrami obróbki. W tej sekcji omówimy proces wyboru oraz podstawową konfigurację oprogramowania sterującego, kierując się ogólnymi kryteriami i wymaganiami projektu.

Kryteria wyboru oprogramowania sterującego:

a) Oprogramowanie powinno oferować szeroki zakres zaawansowanych funkcji i możliwości konfiguracyjnych, umożliwiających precyzyjne sterowanie frezarką CNC oraz dostosowanie jej pracy do specyfiki projektu,

b) Istotne jest, aby oprogramowanie było elastyczne i skalowalne, co pozwoli na łatwe dostosowanie do różnych typów maszyn oraz rozbudowę o dodatkowe funkcje w miarę potrzeb,

c) wysoka wydajność i stabilność działania oprogramowania są kluczowe dla zapewnienia płynnej pracy frezarki CNC oraz uniknięcia awarii w trakcie obróbki materiałów.

d) istotne jest, aby oprogramowanie cieszyło się wsparciem aktywnej społeczności użytkowników oraz deweloperów, co zapewni regularne aktualizacje oraz wsparcie techniczne w przypadku problemów.

e) oprogramowanie powinno być kompatybilne z różnymi rodzajami kontrolerów oraz innymi urządzeniami, co umożliwi jego łatwe dostosowanie do konkretnego sprzętu oraz integrację z innymi systemami.

Oprogramowanie użyte w projekcie:

W ramach tego projektu zdecydowano się na wykorzystanie oprogramowania Klipper dla urządzenia docelowego, z hostem działającym na Klippy i Mainsail. Pomimo dostępności innych opcji oprogramowania, takich jak GRBL czy LinuxCNC, wybór Klippera został podyktowany kilkoma czynnikami. Klipper oferuje zaawansowane funkcje, wysoką elastyczność oraz skalowalność, co sprawia, że jest idealnym rozwiązaniem dla bardziej zaawansowanych projektów. Dodatkowo, Klipper cechuje się wydajnością i stabilnością działania, co jest kluczowe dla zapewnienia płynnej pracy frezarki CNC. Wykorzystanie Klippy i Mainsail jako hosta pozwala na łatwą integrację z Klipperem oraz zapewnia intuicyjny interfejs użytkownika, co ułatwia obsługę urządzenia nawet dla mniej doświadczonych użytkowników. W rezultacie, wybór oprogramowania Klipper został dokonany z myślą o zapewnieniu efektywnej i niezawodnej pracy frezarki CNC w ramach tego projektu.

Klipper wyróżnia się również aktywną społecznością użytkowników oraz regularnymi aktualizacjami, co zapewnia wsparcie oraz rozwój oprogramowania na przyszłość. Ponadto, architektura Klippera oparta na separacji oprogramowania sterującego od hosta pozwala na lepsze wykorzystanie zasobów sprzętowych oraz osiągnięcie wyższej wydajności w porównaniu do niektórych innych rozwiązań.

Warto również zauważyć, że Klipper jest dostosowany do pracy z różnymi kontrolerami drukarek 3D, co sprawia, że jest elastyczny i skalowalny. Dzięki temu można łatwo dostosować go do konkretnych potrzeb projektu oraz wykorzystać go w innych aplikacjach, nie tylko w frezarce CNC.

Podsumowując, wybór oprogramowania Klipper dla urządzenia docelowego, z hostem Klippy i Mainsail, został dokonany po dokładnej analizie dostępnych opcji oraz uwzględnieniu wymagań projektu. Klipper oferuje zaawansowane funkcje, wydajność oraz stabilność, co sprawia, że jest idealnym rozwiązaniem dla frezarki CNC.

5.2.Konfiguracja Parametrów Obróbki.

Konfiguracja parametrów obróbki jest kluczowym etapem przygotowania frezarki CNC do pracy. Odpowiednie dostosowanie tych parametrów ma bezpośredni wpływ na efektywność, jakość oraz bezpieczeństwo procesu obróbki materiałów. W tej sekcji omówimy szczegółowo proces konfiguracji najważniejszych parametrów obróbki, takich jak prędkość wrzeciona, posuw materiału, głębokość skrawania oraz chłodzenie i smarowanie, kierując się ogólnymi zasadami i praktykami branżowymi.

1. Prędkość Wrzeciona:

a) prędkość wrzeciona jest jednym z kluczowych parametrów obróbki, który ma bezpośredni wpływ na wydajność oraz jakość frezowania.
b) wartość prędkości wrzeciona należy dostosować do rodzaju materiału oraz rodzaju narzędzia używanego do obróbki.
c) przy zbyt wysokiej prędkości wrzeciona istnieje ryzyko przegrzania narzędzia i materiału, natomiast przy zbyt niskiej prędkości może dojść do niewystarczającego usuwania materiału i pogorszenia jakości obróbki.

2. Posuw Materiału:

a) posuw materiału określa szybkość poruszania się narzędzia w stosunku do obrabianego materiału i ma wpływ na tempo oraz jakość obróbki.
b) optymalna wartość posuwu materiału zależy od rodzaju materiału, rodzaju narzędzia oraz głębokości skrawania.
c) zbyt niski posuw może prowadzić do nadmiernego zużycia narzędzia oraz zwiększonego czasu obróbki, natomiast zbyt wysoki posuw może prowadzić do uszkodzenia narzędzia oraz pogorszenia jakości obróbki.

3. Głębokość Skrawania:

a) głębokość skrawania określa głębokość, na jaką narzędzie wnika w obrabiany materiał podczas jednego przejścia.
b) wartość głębokości skrawania należy dostosować do twardości i wytrzymałości materiału oraz możliwości narzędzia.
c) zbyt duża głębokość skrawania może prowadzić do przeciążenia narzędzia oraz zwiększonego ryzyka uszkodzenia obrabianego materiału, natomiast zbyt mała głębokość skrawania może prowadzić do wydłużenia czasu obróbki oraz pogorszenia jakości wykończenia.

4. Chłodzenie i Smarowanie:

a) w niektórych przypadkach, szczególnie podczas obróbki metali, zaleca się stosowanie chłodzenia i smarowania, aby zmniejszyć temperaturę obrabianego materiału oraz narzędzia.
b) optymalny rodzaj i ilość chłodziwa lub smaru oraz parametry chłodzenia należy dostosować do rodzaju materiału i rodzaju obróbki.
c) stosowanie odpowiednich chłodziw i smarów oraz dostosowanie parametrów chłodzenia do konkretnych warunków obróbki ma kluczowe znaczenie dla efektywności i bezpieczeństwa procesu obróbki.
d) Konfiguracja parametrów obróbki jest kluczowym elementem przygotowania frezarki CNC do pracy. Poprzez odpowiednie dostosowanie prędkości wrzeciona, posuwu materiału, głębokości skrawania

oraz chłodzenia i smarowania możliwe jest osiągnięcie optymalnych wyników obróbki materiałów. Wartość tych parametrów należy dostosować do specyfiki obrabianego materiału oraz wymagań projektu, co pozwoli na uzyskanie optymalnej wydajności i jakości obróbki.

5.3.Procedury Bezpieczeństwa i Awaryjne.

Konfiguracja oprogramowania oraz odpowiednich parametrów obróbki są istotnym elementem procesu przygotowania frezarki CNC do pracy. Poprzez właściwe dostosowanie tych elementów możliwe jest osiągnięcie optymalnych wyników obróbki materiałów. Kolejnym etapem po przeprowadzeniu konfiguracji jest właściwa obróbka materiałów, która będzie realizowana w ramach projektu.

6. Wnioski

Realizacja projektu przekształcenia drukarki 3D Anycubic Mega X w frezarkę CNC do wyrobu płytek PCB dostarczyła cennych doświadczeń i wniosków. Przede wszystkim, projekt pokazał, że nawet relatywnie niedrogie i popularne drukarki 3D mogą zostać z powodzeniem przekształcone w bardziej zaawansowane urządzenia, takie jak frezarki CNC, przy odpowiedniej modyfikacji i dostosowaniu ich komponentów oraz oprogramowania.

Proces ten wymagał szczegółowego zrozumienia budowy drukarki 3D oraz zasad działania frezarek CNC. Kluczowe okazało się usunięcie elementów, takich jak ekstruder i stół, które nie są potrzebne w nowej konfiguracji, oraz dodanie komponentów specyficznych dla frezarki CNC, takich jak wrzeciono, stół do frezowania i system odprowadzania trocin.

Wybór odpowiedniego oprogramowania sterującego był krytyczny dla powodzenia projektu. Po dokładnej analizie dostępnych opcji, zdecydowano się na oprogramowanie Klipper, wspierane przez Klippy i Mainsail. Kluczowe kryteria wyboru obejmowały zaawansowane funkcje, elastyczność, wydajność, stabilność oraz aktywne wsparcie społecznościowe. Klipper okazał się optymalnym wyborem, zapewniającym nie tylko efektywne sterowanie frezarką, ale również możliwość dalszego rozwoju i adaptacji systemu.

Dostosowanie parametrów obróbki, takich jak prędkość wrzeciona, posuw materiału, głębokość skrawania oraz chłodzenie i smarowanie, było kluczowe dla uzyskania wysokiej jakości rezultatów. Parametry te musiały być precyzyjnie dostosowane do specyfiki obrabianego materiału oraz wymagań projektu, co pozwoliło na optymalizację procesu obróbki.

Projekt podkreślił również znaczenie testowania i kalibracji wstępnej. Dokładne testy i kalibracje były niezbędne do zapewnienia, że wszystkie komponenty działają zgodnie z oczekiwaniami oraz że konfiguracja parametrów obróbki jest optymalna. Dzięki temu możliwe było uniknięcie potencjalnych problemów i zapewnienie wysokiej precyzji oraz jakości obróbki.

Wnioski płynące z tego projektu wskazują na ogromny potencjał modyfikacji istniejących urządzeń w celu poszerzenia ich funkcjonalności. Dostosowanie drukarki 3D do pełnienia funkcji frezarki CNC może być efektywnym i ekonomicznym rozwiązaniem, umożliwiającym realizację bardziej złożonych zadań w domowych czy akademickich warsztatach. Projekt ten może stanowić inspirację dla dalszych prac badawczo-rozwojowych oraz być podstawą dla kolejnych innowacyjnych projektów w dziedzinie technologii wytwarzania.

7. Podsumowanie

Projekt przekształcenia drukarki 3D Anycubic Mega X w frezarkę CNC do wyrobu płytek PCB był ambitnym i wymagającym przedsięwzięciem, które dostarczyło wielu cennych doświadczeń i wniosków. Proces ten pokazał, że przy odpowiedniej modyfikacji i adaptacji możliwe jest wykorzystanie popularnych

drukarek 3D do realizacji zaawansowanych zadań obróbkowych, co otwiera nowe możliwości w dziedzinie amatorskiej i półprofesjonalnej produkcji.

Przede wszystkim, projekt wymagał szczegółowego zrozumienia budowy i zasad działania zarówno drukarki 3D, jak i frezarki CNC. Kluczowe było usunięcie niepotrzebnych elementów drukarki oraz dodanie nowych komponentów, takich jak wrzeciono, stół do frezowania, system odprowadzania trocin oraz odpowiednia obudowa. Ważnym krokiem było również dostosowanie elektroniki, która zarządza nowymi funkcjami urządzenia, oraz instalacja i konfiguracja zaawansowanego oprogramowania sterującego.

Wybór oprogramowania Klipper, wspieranego przez Klippy i Mainsail, okazał się trafny, zapewniając zaawansowane funkcje, elastyczność oraz wydajność niezbędną do sterowania frezarką CNC. Kluczowe kryteria wyboru oprogramowania obejmowały zaawansowane możliwości konfiguracyjne, skalowalność, wsparcie społecznościowe oraz stabilność działania, co pozwoliło na uzyskanie optymalnych wyników.

Konfiguracja parametrów obróbki, takich jak prędkość wrzeciona, posuw materiału, głębokość skrawania oraz chłodzenie i smarowanie, była istotnym etapem projektu. Poprawne dostosowanie tych parametrów do specyfiki obrabianego materiału i wymagań projektu było kluczowe dla osiągnięcia wysokiej jakości obróbki i efektywności pracy urządzenia.

Ważnym aspektem projektu było również przeprowadzenie dokładnych testów i wstępnych kalibracji, które zapewniły poprawne działanie wszystkich komponentów oraz optymalną konfigurację parametrów obróbki. Dzięki temu możliwe było uniknięcie potencjalnych problemów oraz zapewnienie wysokiej precyzji i jakości wykonywanych zadań.

Projekt pokazał, że modyfikacja istniejących urządzeń, takich jak drukarki 3D, w celu poszerzenia ich funkcjonalności jest nie tylko możliwa, ale również efektywna i ekonomiczna. Tego typu projekty mogą stanowić inspirację dla dalszych badań i rozwoju w dziedzinie technologii wytwarzania oraz być podstawą dla kolejnych innowacyjnych rozwiązań. Przekształcenie drukarki 3D w frezarkę CNC to doskonały przykład kreatywnego podejścia do wykorzystania dostępnych zasobów i technologii, które może przynieść wymierne korzyści w wielu dziedzinach, zarówno w środowisku domowym, jak i akademickim.

KOŁO

NAUKOWE

RACHUNKOWOŚCI

„ASSETS"

Joanna Chruściel
Studenckie Koło Naukowe Rachunkowości „ASSETS”

Prof. Grzegorz Lew[1]
Opiekun Koła Naukowego „ASSETS”

Wyzwania dotyczące prowadzenia rachunkowości w szpitalach

Streszczenie

Rachunkowość w szpitalu jest dość rozbudowana, przez co może powodować wyzwania związane z jej prowadzeniem. W artykule przedstawiono znaczenie, istotę i funkcje rachunkowości w podmiotach leczniczych. Skupiono się także na problematyce prowadzenia rachunku kosztów w szpitalach. Przedstawiono jednocześnie aspekty związane z zarządzaniem finansami w podmiotach leczniczych. Celem badania jest zrozumienie, jak rachunkowość funkcjonuje w szpitalach oraz jakie wyzwania przed nimi stawia. Jako metodę badawczą wykorzystano ścisły przegląd literatury. Rachunkowość odgrywa kluczową rolę w dostarczaniu niezbędnych informacji do zarządzania czy kontrolowania wyników szpitala. Wyzwaniem w rachunkowości jest identyfikacja i rozliczenie kosztów w sposób dostosowany do specyfiki działalności medycznej. Zarządzanie finansami jest utrudnione ze względu na podział na cele medyczne i ekonomiczne prowadzenia działalności.

Słowa kluczowe: podmioty lecznicze, rachunek kosztów, zarządzanie finansami, rachunkowość.

1. Wprowadzenie

Polski system ochrony zdrowia obejmuje wiele podmiotów leczniczych, które w nim funkcjonują. Wśród różnic między tymi podmiotami można wyróżnić m.in. zakres, wielkość oraz skala prowadzonej działalności, a także rodzaj świadczonych usług medycznych, forma własności czy forma organizacyjno-prawna. Wszystkie te podmioty pełnią rolę świadczeniodawców usług zdrowotnych w celu pomocy pacjentom. Finansowanie tych świadczeń, realizowanych przez świadczeniodawców, jest obowiązkiem płatnika. Płatnikami mogą być m.in. Narodowy Fundusz Zdrowia, organy rządowe (takie jak Ministerstwo Zdrowia), organy samorządowe, inne podmioty lecznicze, prywatne firmy ubezpieczeniowe oraz indywidualni pacjenci. Podmiot leczniczy to jednostka organizacyjna posiadająca zespół majątkowy, który obejmuje m.in. majątek ruchomy (takie jak sprzęt, aparatura medyczna, wyposażenie) oraz majątek nieruchomy (przykładowo budynki i inne obiekty). Przeznaczenie tego majątku służy realizacji określonej działalności leczniczej. Działalność lecznicza obejmuje udzielanie świadczeń zdrowotnych dla potrzebujących oraz propagowanie zdrowia. Dodatkowo, obejmuje ona również działalność dydaktyczną i badawczą, prowadzoną równocześnie z udzielaniem świadczeń zdrowotnych. Działalność ta wspiera wprowadzanie

[1] Prof. Grzegorz Lew Opiekun Studenckiego Koła Naukowego Assets.

nowoczesnych technologii medycznych oraz metod leczenia, a także uczestniczy w procesie kształcenia osób przygotowujących się do zawodu medycznego oraz osób już pracujących w tym sektorze. Szpital pełni wyjątkowy charakter, ponieważ jego nadrzędnym celem jest wypełnianie misji społecznej w postaci udzielania wysokiego poziomu świadczeń zdrowotnych. W przypadku pozostałej części jednostek gospodarczych celem jest maksymalizacja i osiąganie zysku[2].

Aby efektywnie prowadzić działalność leczniczą, konieczne jest zatrudnienie odpowiednio wykwalifikowanej kadry, w tym personelu medycznego (takiego jak lekarze, pielęgniarki, technicy medyczni), personelu administracyjno-finansowego (w tym kadra kierownicza) oraz personelu zajmującego się eksploatacją i techniką[3].

Celem badania jest zrozumienie, jak rachunkowość funkcjonuje w szpitalach oraz jakie wyzwania przed nimi stawia. Jako metodę badawczą wykorzystano ścisły przegląd literatury.

2. Znaczenie, istota i funkcje rachunkowości w podmiotach leczniczych

Rachunkowość może być definiowana jako proces, w którym identyfikuje się, mierzy czy przekazuje wiadomości ekonomiczne, mając na celu świadome podejmowanie wyboru poprzez osoby używające tych danych. Prowadzenie rachunkowości w polskich podmiotach leczniczych odbywało się nawet w czasach średniowiecza. W XV w. był już sporządzany inwentarz dóbr szpitala czy rejestr dotyczący dochodów oraz wydatków szpitala. Przede wszystkim rachunkowość dostarcza informacji niezbędnych do kontroli wyników uzyskiwanych w danej działalności. W początkowych latach 80. XX wieku zwrócono uwagę na rachunkowość zarządczą w podmiotach leczniczych, gdyż pozwalała ona na zarządzanie instytucją leczniczą w całości jak i w poszczególnych jej częściach. Pomagała także kadrze kierowniczej przy wyborze właściwych decyzji, planowaniu oraz kontrolowaniu wyników szpitala. W przypadku rachunkowości finansowej, czyli jednego z członu rachunkowości, użytkownikami tych informacji są odbiorcy zewnętrzni w przeciwieństwie do rachunkowości zarządczej, gdzie odbiorcami są wewnętrzni użytkownicy. Sporządzone sprawozdania finansowe, obejmujące w postaci rozszerzonej m.in. bilans, rachunek zysków i strat, sprawozdanie z przepływu środków pieniężnych, zestawienie zmian w kapitale (funduszu) własnym oraz informację dodatkową, są wynikiem prowadzonej rachunkowości finansowej.

[2] R. Orliński, M. Niestrata-Ortiz, *Wynik na podstawowej działalności operacyjnej szpitala*, Zeszyty Naukowe Uniwersytetu Ekonomicznego w Katowicach, nr 285, UEK, Katowice 2016, s. 172.
[3] M. Hass-Symotiuk, M. Kludacz-Alessandri, M. Cygańska, *Rachunkowość podmiotów leczniczych I jej potencjał informacyjny. Dobre praktyki w ochronie zdrowia*, Wydawnictwo Wolters Kluwer, Warszawa 2020, s. 23-24.

System rachunkowości obejmuje też rachunek kosztów, który w szpitalu powoduje duże wyzwanie. Rozumie się przez niego system, który pozwala na identyfikację, pomiar czy przetwarzanie wiadomości dotyczących koszów w celach rachunkowości zarządczej oraz finansowej. Rachunek kosztów to część składowa systemu informacyjnego w rachunkowości. W jego zakres wchodzi planowanie, ewidencja, kalkulacja oraz analiza kosztów. Można wyróżnić rachunek kosztów m.in. retrospektywny czyli działający w celach rachunkowości finansowej oraz prospektywny, który funkcjonuje w ramach rachunkowości zarządczej. Rachunek kosztów rozwinął się w latach 80. XX w. w Stanach Zjednoczonych ze względu na zmianę systemu finansowania świadczeń zdrowotnych. W podmiotach leczniczych oczekuje się, aby rachunek kosztów zaspokajał potrzeby informacyjne odnośnie przygotowania sprawozdań[4].

Przy prowadzeniu dokumentacji w szpitalach bierze się pod uwagę przepisy zawarte zwłaszcza w ustawie o rachunkowości[5], ale także w ustawie o finansach publicznych[6] oraz w ustawie o działalności leczniczej[7]. Rachunkowość w podmiotach leczniczych odgrywa określone funkcje. Rozumie się przez to cele czy zadania następujące na skutek poznania czy używania rachunkowości w działalności. Do głównej funkcji rachunkowości należy funkcja informacyjna, która generuje wiele wiadomości odnośnie zjawisk czy procesów gospodarczych w szpitalach. Obejmuje także dopasowanie ich do potrzeby informacyjnej kadr zarządzających czy podmiotów zewnętrznych np. organy założycielskie, banki, urzędy statystyczne czy Ministerstwo Zdrowia. W funkcji informacyjnej można wyszczególnić m.in. funkcje przebiegu tych informacji, czyli funkcję dowodową, klasyfikacyjną, rejestracyjną czy sprawozdawczą, oraz funkcje wykorzystania pozyskanych informacji czyli funkcję analityczną, motywacyjną, rozliczeniową, kontrolną i optymalizacyjną. Funkcja dowodowa oznacza wymóg zapisywania danych w postaci dokumentów, przetrzymywaniu ich w jednostce oraz zabezpieczaniu. Funkcja rejestracyjna obejmuje rejestrowanie zużycia zasobów szpitala i ewidencję skutków zdarzeń w zakresie gospodarczym powiązanych z użyciem tych zasobów. Z kolei funkcja klasyfikacyjna umożliwia podzielenie uzyskanych danych na kategorie czy podkategorie. Sporządzanie sprawozdań finansowych wchodzi w zakres funkcji sprawozdawczej. Z drugiego typu szczegółowych funkcji można wyjaśnić funkcję analityczną jako działanie pozwalające na

[4] M. Cygańska, *Integracja informacji finansowych i klinicznych na potrzeby zarządzania operacyjnego szpitalem*, Wydawnictwo Uniwersytetu Warmińsko-Mazurskiego w Olsztynie, Olsztyn 2018, s. 55-57; A. Dyhdalewicz, I. Rutkowska, *Specyfika ustalania wyniku finansowego podmiotów leczniczych – studium przypadku*, Akademia Zarządzania, nr 3(4) 2019, Białystok 2019, s. 52-54.
5 Ustawa z dnia 29 września 1994 r. o rachunkowości (Dz.U. z 2023 r., poz. 120 z późn.zm.).
[6] Ustawa z dnia 27 sierpnia 2009 r. o finansach publicznych (Dz.U. z 2023 r., poz. 1270 z późn.zm.).
[7] Ustawa z dnia 15 kwietnia 2011 r. o działalności leczniczej (Dz.U. z 2024 r., poz. 799 z późn.zm.).

badanie, a następnie interpretację uzyskanych wyników finansowych. Spełnienie tej funkcji jest problematyczne, gdyż w szpitalu panuje większa zmienność oraz niski poziom standaryzacji leczenia pacjentów. Z kolei funkcja optymalizacyjna dostarcza danych umożliwiających dokonanie decyzji odnośnie najlepszych metod działania. W rachunkowości jest także funkcja kontrolna, która obejmuje pomiar poziomu realizacji ustalonych celów, ustalenie odchyleń oraz ich oceny. Szczególnym obszarem, w którym dokonuje się kontroli kosztów jest rachunek kosztów dostarczający wiele problemów. Zostanie on dokładniej omówiony ze względu na wyzwania, jakie powoduje. Działa on jako składnik systemu budżetowania, gdzie prowadzi się analizę w zakresie odchyleń między kosztami rzeczywistymi a założeniami budżetowymi. W szpitalach często jest brana pod uwagę kontrola dotycząca chociażby kosztu jednostkowego leczenia pacjenta z daną jednostką choroby, kosztu przyjęcia chorego do szpitala albo koszt co do utrzymania jednego łóżka na danym oddziale. Inną funkcją rachunkowości jest funkcja rozliczeniowa, gdzie można ją rozumieć jako rozliczenie zarządu w zakresie efektywnego zarządzania czy właściwego korzystania z majątku i jego zwiększania[8].

3. Charakterystyka rachunku kosztów – wyzwania w rachunkowości podmiotu leczniczego

Występuje wiele definicji odnośnie rachunku kosztów. Można go przedstawiać jako ogólne działania w rachunkowości m.in. takie jak mierzenie, podział na grupy, przetwarzanie, przedstawianie, interpretowanie czy podejmowanie analiz określonych wartościowo oraz ilościowo rezultatów procesu wykorzystania zasobów majątkowych firmy, które zachodzą na skutek prowadzenia jednostki gospodarczej. W przypadku szpitali rachunek kosztów jest określony jako zespół ogólnych działań, który ma na celu wyodrębnienie kosztów dotyczących procedur medycznych wynikających z działalności podmiotu leczniczego. Dokonuje się tego w celu zdobycia wiadomości w zakresie kosztów niezbędnych do określenia wyniku finansowego, wygenerowania sprawozdań finansowych czy ustalania podjętych wyborów zarządczych. Wśród celów dotyczących prowadzenia rachunku kosztów w podmiocie leczniczym można wyróżnić m.in.[9]:

- określenie czynników determinujących poziom oraz strukturę kosztów,

[8] *Ibidem*, s. 58-60; R. Wawrowski, *Specyfika polityki rachunkowości samodzielnych publicznych zakładów opieki zdrowotnej* [w:] *Polityka rachunkowości a kształtowanie wyniku finansowego*, nr 201, UEK, Katowice 2014, s. 396.

[9] M. Talarska, *Rachunek kosztów w szpitalu publicznym*, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu, nr 182, Wrocław 2011, s.525-526.

- zdobycie wiadomości o koszcie jednostkowym oraz całkowitym wykonanych świadczeń zdrowotnych,
- pomóc w podejmowaniu właściwych wyborów powiązanych z np. określeniem ceny rozliczeniowej na świadczenie wewnętrzne, wykonaniem planu prowadzonego podmiotu leczniczego,
- wykonanie wyniku jednostki gospodarczej danego miejsca powstawania kosztów oraz wskazanie na udział w całym wyniku finansowym,
- śledzenie ponoszonych kosztów oraz osiąganych rezultatów w zakresie działalności wyodrębnionych komórek organizacyjnych czy typów działalności,
- kontrolowanie, analizowanie wyników czy kosztów firmy czasowo i przestrzennie,
- zbadanie czy ocenianie stopnia prawidłowego użycia zasobów, które są w dyspozycji firmy celem polepszenia efektywności działania podmiotu,
- zwiększenie wiedzy ekonomicznej pracowników.

W szpitalach publicznych należy używać rachunku kosztów popartego rachunkiem kosztów pełnych, lecz musi być on przystosowany do charakteru prowadzenia szpitali. Na początku koszty działalności operacyjnej podmiotu leczniczego ujmuje się na rzecz kosztów rodzajowych. Kolejno określa się koszty dotyczące obecnego okresu, a następnie rozlicza się je na poszczególne podmioty czyli ośrodki kosztów. Koszty, które obejmują przyszłe okresy umieszcza się na koncie rozliczeń międzyokresowych kosztów. W dalszej kolejności rozlicza się koszty wewnętrznych świadczeń czyli takich, gdzie jeden ośrodek kosztów pełni świadczenia dla drugiego ośrodka. Po tej czynności rozlicza się koszty wydziałowe według nośników kosztów oraz ustala się rzeczywiste koszty w zakresie usług medycznych i pozostałych nośników[10].

Szpitale często dokonują wiele różnych typów działalności medycznej ze względu na ich rozmiar prowadzonej działalności. To powoduje, że jednostka musi prowadzić układ podmiotowy czy przedmiotowy kosztów. Powodem, przez który szpitale ponoszą koszty jest przygotowanie do udzielenia pomocy w zakresie świadczeń zdrowotnych. Koszty te są większe, gdy zwiększa się niepewność i ryzyko prowadzonej działalności. Często niepewność może powodować ograniczenia w procesie planowania kosztów. Wyzwaniem jest zarządzanie szpitalem publicznym tak, aby minimalizować koszty w dłuższym okresie i jednocześnie nie powodować poważnych zakłóceń w działaniu szpitalu. Zasadniczym warunkiem, aby szpital

[10] *Ibidem*, s. 526-527; E. Rabiej, *Rachunek kosztów leczenia w aspekcie wyceny świadczeń zdrowotnych*, Wydawnictwo Uniwersytetu Rzeszowskiego, Rzeszów 2013, s.43-45.

dobrze funkcjonował jest dobry wybór produktu działalności, którym mogą być usługi zdrowotne, gdzie układane jest to według potrzeb zgłaszanych przez pacjentów. Aspekt finansowy wpływa często na popyt czyli dostęp do usług spoza świadczeń oferowanych przez Narodowy Fundusz Zdrowia[11].

4. Zarządzanie podmiotem leczniczym

Finanse określonego podmiotu są powiązane z procesami, w których występuje ruch pieniądza. Procesy te mogą dotyczyć dwóch sytuacji m.in. gdzie pieniądz wkracza do podmiotu i wydostaje się. W pierwszym przypadku jest to konieczne, aby działalność funkcjonowała, gdyż pozwala to na zakup oraz użycie różnych aktywów do celów działalności. Zarządzanie finansami odnośnie szpitala obejmuje trzy obszary czyli zarządzanie posiadanym kapitałem, zastosowanie go i finansowanie. W przypadku zarządzania kapitałem należy rozumieć czynności, w których wyodrębnia się zysk i nadwyżkę finansową oraz wpływy z wydatkami. Przewaga wpływów nad wydatkami mówi o stopniu utrzymania płynności finansowej tej działalności. Finansowanie to gromadzenie potrzebnych środków pieniężnych, które przeznacza się na bieżące potrzeby i na rozwój jednostki. Może się to przejawiać w postaci finansowania własnej jednostki w długim czy krótkim okresie albo gromadzenia nadwyżek w innych działalnościach w formie np. lokat bankowych. Cele działania podmiotów leczniczych nadają wyjątkowy charakter w przypadku zarządzania finansami. Wynika to w większości z podwójnego celu działania takich szpitali czyli medycznego i ekonomicznego. Należy brać pod uwagę czy podmiot leczniczy działa w postaci własności publicznej czy prywatnej. To wyróżnienie ma wpływ na określanie czy zrozumienie szczegółowych celów w takich podmiotach. Jednym z celów jest zapewnienie społeczeństwu wyleczenia czyli ma to charakter medyczny. Może być on wykonywany w różnym zakresie. W przypadku publicznych zakładów leczniczych wykonują one zadania celem realizacji prawa do ochrony zdrowia według konstytucji Rzeczpospolitej Polskiej. Gwarantują one wykonywanie świadczeń zdrowotnych zapewnianych przez państwo czyli nie skupiają się na wybranych świadczeniach. Z kolei prywatne podmioty w zakresie leczenia świadczą usługi według uznanej przez założycieli strategii działania. Działają tylko w tym obszarze, który określili. Zaś celem ekonomicznym jest chęć maksymalizacji wartości danego podmiotu leczniczego. W niepublicznych zakładach cel ten nie może być osiągany bez nadwyżki finansowej. Natomiast w publicznych zakładach

[11] J. Chluska, *Problemy analizy kosztów w zarządzaniu SPOZ,* Zeszyty Naukowe Uniwersytetu Szczecińskiego, nr 679, Częstochowa 2011, s. 9-10; R. Orliński, *Rachunek kosztów pacjenta na przykładzie szpitala* [w:] *Wyzwania w zarządzaniu kosztami i dokonaniami,* Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu, nr 442, Wrocław 2016, s. 361-363.

maksymalizacja taka może dotyczyć przykładowo unikania strat, zmniejszenia zadłużenia przez jednostki czy zwiększenia efektywności w maksymalny sposób użycia posiadanych zasobów. Przy analizie różnych obszarów zarządzania finansami oraz stosowaniu różnorodnych narzędzi zarządzania finansami w konkretnym podmiocie leczniczym, należy brać pod uwagę jego specyfikę, dostosowując wnioski i interpretację do charakteru jego działalności[12].

5. Podsumowanie

Celem artykułu było zrozumienie, jak rachunkowość funkcjonuje w szpitalach oraz jakie wyzwania przed nimi stawia. Jako metodę badawczą wykorzystano ścisły przegląd literatury. Rachunkowość odgrywa kluczową rolę w dostarczaniu niezbędnych informacji do zarządzania czy kontrolowania wyników działalności podmiotów leczniczych. Stosowanie rachunkowości finansowej jak i rachunkowości zarządczej powoduje, że występują specyficzne funkcje przy każdej z nich, co pokazuje złożoność systemu rachunkowości w szpitalach. Rachunkowość w szpitalach odgrywa wiele funkcji, wśród których można wyodrębnić funkcję informacyjną, która przedstawiana jest jako nadrzędna. Generuje ona wiele informacji na temat zjawisk czy procesów gospodarczych w szpitalach. Dopasowuje je także do potrzeb informacyjnych kadr zarządzających czy podmiotów zewnętrznych np. urzędów statystycznych. Funkcja informacyjna wyodrębnia podział na inne funkcje składowe. Wynika z nich to, że jednostka musi zapisywać dane w postaci dokumentów, przechowywać je i zabezpieczać. Obejmuje także obserwację zużycia zasobów szpitala oraz ewidencjonowanie skutków zdarzeń w zakresie gospodarczym powiązanych z użyciem tych zasobów. Funkcja klasyfikacyjna pozwala na podział danych na kategorie czy podkategorie. Sporządzanie sprawozdań finansowych wchodzi w zakres funkcji sprawozdawczej. Wśród funkcji wyzwanie powoduje funkcja analityczna w podmiotach leczniczych, gdyż oznacza badanie oraz interpretację osiągniętych wyników finansowych. W szpitalach panuje większa zmienność oraz niski poziom standaryzacji leczenia pacjentów, dlatego spełnienie tej funkcji może być problematyczne. Poprzez funkcję optymalizacyjną możliwe jest wybranie najlepszych metod działania szpitala. Rachunkowość w szpitalach pełni także funkcje kontrolne, to znaczy, że mierzy stopień realizacji ustalonych celów. Najwięcej problemów dostarcza rachunek kosztów. Problematyczna jest identyfikacja i rozliczenie kosztów w sposób dostosowany do specyfiki działalności medycznej. Wyzwaniem jest takie

[12] K. Stańczak-Strumiłło, R. Kotapski, *Zarzadzanie finansami w podmiotach leczniczych,* Wydawnictwo Uniwersytetu Gdańskiego, Gdańsk 2021, s. 23-26; A. Wiercińska, *Przydatność danych pochodzących z rachunkowości w analizach benchmarkingowych szpitala* [w:] *Finanse, Rynki Finansowe Ubezpieczenia,* nr 2/2026 (80), Uniwersytet Gdański, Gdańsk 2016, s. 412-414.

zarządzanie szpitalem publicznym, aby minimalizować koszty w dłuższym okresie, ale jednocześnie też nie powodować poważnych zakłóceń w działaniu szpitala. Głównym warunkiem, aby szpital dobrze działał jest odpowiedni dobór produktu działalności, którym mogą być usługi zdrowotne, gdzie układane jest to według potrzeb zgłaszanych przez pacjentów. Zarządzanie finansami obejmuje trzy sfery: zarządzanie kapitałem, zastosowanie środków oraz ich finansowanie, co jest skomplikowane ze względu na konieczność balansowania celów medycznych oraz ekonomicznych. Specyfika działalności publicznych i prywatnych podmiotów leczniczych wpływa także na sposób zarządzania finansami, gdzie publiczne podmioty skupiają się na wykonaniu prawa do ochrony zdrowia, a prywatne na strategii działania określonej przez właścicieli. Rachunkowość w szpitalach jest narzędziem do ewidencji finansowej, ale także zasadniczym elementem zarządzania, planowania oraz kontroli. Wyzwania związane z rachunkiem kosztów i zarządzaniem finansami są złożone i wymagają dokładnego podejścia dostosowanego do specyfiki szpitali.

Literatura

1. Chluska J., *Problemy analizy kosztów w zarządzaniu SPOZ*, Zeszyty Naukowe Uniwersytetu Szczecińskiego, nr 679, Częstochowa 2011.
2. Cygańska M., *Integracja informacji finansowych i klinicznych na potrzeby zarządzania operacyjnego szpitalem*, Wydawnictwo Uniwersytetu Warmińsko-Mazurskiego w Olsztynie, Olsztyn 2018.
3. Dyhdalewicz A., Rutkowska I., *Specyfika ustalania wyniku finansowego podmiotów leczniczych – studium przypadku*, Akademia Zarządzania, nr 3(4) 2019, Białystok 2019.
4. Hass-Symotiuk M., Kludacz-Alessandri M., Cygańska M., *Rachunkowość podmiotów leczniczych I jej potencjał informacyjny. Dobre praktyki w ochronie zdrowia*, Wydawnictwo Wolters Kluwer, Warszawa 2020.
5. Orliński R., Niestrata-Ortiz M., *Wynik na podstawowej działalności operacyjnej szpitala*, Zeszyty Naukowe Uniwersytetu Ekonomicznego w Katowicach, nr 285, UEK, Katowice 2016.
6. Orliński R., *Rachunek kosztów pacjenta na przykładzie szpitala* [w:] *Wyzwania w zarządzaniu kosztami i dokonaniami*, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu, nr 442, Wrocław 2016, s. 358-367.
7. Rabiej E., *Rachunek kosztów leczenia w aspekcie wyceny świadczeń zdrowotnych*, Wydawnictwo Uniwersytetu Rzeszowskiego, Rzeszów 2013.

8. Stańczak-Strumiłło K., Kotapski R., *Zarzadzanie finansami w podmiotach leczniczych*, Wydawnictwo Uniwersytetu Gdańskiego, Gdańsk 2021.
9. Talarska M., *Rachunek kosztów w szpitalu publicznym*, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu, nr 182, Wrocław 2011.
10. Wawrowski R., *Specyfika polityki rachunkowości samodzielnych publicznych zakładów opieki zdrowotnej* [w:] *Polityka rachunkowości a kształtowanie wyniku finansowego*, nr 201, UEK, Katowice 2014, s. 393-403.
11. Wiercińska A., *Przydatność danych pochodzących z rachunkowości w analizach benchmarkingowych szpitala* [w:] *Finanse, Rynki Finansowe Ubezpieczenia*, nr 2/2026 (80), Uniwersytet Gdański, Gdańsk 2016, s. 411-420.

Akty normatywne

1. Ustawa z dnia 15 kwietnia 2011 r. o działalności leczniczej (Dz.U. z 2024 r., poz. 799 z późn.zm.).
2. Ustawa z dnia 27 sierpnia 2009 r. o finansach publicznych (Dz.U. z 2023 r., poz. 1270 z późn.zm.).
3. Ustawa z dnia 29 września 1994 r. o rachunkowości (Dz.U. z 2023 r., poz. 120 z późn.zm.).

Aleksandra Dołżycka
Studenckie Koło Naukowe Rachunkowości "ASSETS"

dr hab. inż. Grzegorz Lew, prof. PRz
Opiekun Koła Naukowego "ASSETS"

Rola blockchain w rachunkowości

Streszczenie

Celem artykułu jest zbadanie roli technologii blockchain w rachunkowości, z uwzględnieniem jej teoretycznych podstaw oraz praktycznych zastosowań. Blockchain, jako rozproszona księga cyfrowa, cechuje się decentralizacją, niezmiennością i transparentnością, co może znacząco usprawnić procesy rejestrowania transakcji, audytu, zarządzania ryzykiem oraz zapewnienia zgodności z regulacjami. Analiza obejmuje zarówno zalety, takie jak zwiększona przejrzystość i bezpieczeństwo, jak i wyzwania, w tym skalowalność i konieczność podnoszenia kwalifikacji pracowników. Blockchain może zrewolucjonizować rachunkowość, eliminując potrzebę pośredników i redukując ryzyko błędów ludzkich. Umożliwia także natychmiastowy dostęp do danych finansowych w czasie rzeczywistym, co może znacząco poprawić podejmowanie decyzji biznesowych.

Słowa kluczowe: blockchain, rachunkowość, technologia, transparentność, decentralizacja

1. Wprowadzenie

Technologia blockchain, pierwotnie stworzona, jako fundament kryptowaluty Bitcoin, zyskała szybko rozgłos, jako potencjalny game-changer w wielu branżach, w tym w rachunkowości[1]. Jej unikalne właściwości, takie jak decentralizacja, niezmienność oraz transparentność, otwierają nowe możliwości dla zarządzania danymi finansowymi, audytu i sprawozdawczości. Rachunkowość, będąca kluczowym elementem funkcjonowania każdej organizacji, polega na precyzyjnym rejestrowaniu, klasyfikacji i raportowaniu transakcji finansowych. Tradycyjne systemy księgowe, mimo iż rozwinięte i powszechnie stosowane, nie są wolne od wad, takich jak podatność na błędy ludzkie, ryzyko oszustw oraz ograniczona przejrzystość i audytowalność. Właśnie w tych obszarach technologia blockchain może przynieść znaczące korzyści. Istotną rolę we wdrażaniu nowoczesnych technologii w rachunkowości odgrywają również specjaliści odpowiedzialni za prowadzenie ksiąg rachunkowych. W ostatnich latach pojawiło się wiele raportów dotyczących zmian, jakie czekają różne zawody w obliczu rosnącej robotyzacji procesów gospodarczych. Wśród

[1] Nakamoto, S. "Bitcoin: A Peer-to-Peer Electronic Cash System." 2008. oraz Swan, M. "Blockchain: Blueprint for a New Economy." O'Reilly Media, Inc., 2015

zawodów uznawanych za "zagrożone" często wymienia się księgowych. Przykładem są centra usług wspólnych w zakresie rachunkowości, gdzie technologia Robot Process Automation (RPA) stała się standardem[2].

Celem artykułu jest zbadanie roli technologii blockchain w rachunkowości, z uwzględnieniem jej teoretycznych podstaw oraz praktycznych zastosowań Technologia blockchain stanowi przełomowy postęp w dziedzinie informatyki, który zrewolucjonizował podejście do przechowywania i przesyłania danych. Łańcuch bloków to rozproszona księga cyfrowa, która w bezpieczny i przejrzysty sposób rejestruje i weryfikuje transakcje. Kluczową cechą tej technologii jest niezmienność każdej transakcji lub bloku danych zapisanego w blockchainie, co oznacza, że nie mogą one zostać zmienione lub usunięte bez konsensusu większości uczestników sieci. Zapewnia to wysoki poziom bezpieczeństwa oraz minimalizuje ryzyko manipulacji danymi lub potencjalnych działań o charakterze oszukańczym.

2. Kluczowe elementy technologii blockchain

Koncepcja blockchain wykracza poza zwykłą innowację, stanowiąc fundamentalne przeobrażenie podejścia do organizacji i zarządzania w gospodarce. W ramach nowo powstającej subdyscypliny ekonomiki blockchain można wyróżnić dwa główne podejścia: jedno skoncentrowane na innowacjach (innovation-centred), a drugie na zarządzaniu (governance-centred)[3]. To drugie podejście, zakorzenione w nowej ekonomii instytucjonalnej oraz teorii wyboru publicznego, wydaje się szczególnie obiecujące z naukowego punktu widzenia. Umożliwia ono płynne przejście od tradycyjnego modelu blockchain do analizy automatycznych, spontanicznie powstających podmiotów nowego typu, takich jak Ethereum, które posiadają niejednoznaczny status prawny (DAO) oraz skomplikowaną kwestię odpowiedzialności[4]. Blockchain, jako technologia nowej generacji redefiniuje funkcjonowanie internetu, wprowadzając zdecentralizowane modele operacyjne, które mogą stanowić zagrożenie dla tradycyjnych, centralnie zarządzanych struktur przedsiębiorstw. Jej potencjał do przekształcania istniejących systemów gospodarczych i organizacyjnych czyni ją kluczowym obszarem badań i wdrożeń w dziedzinie współczesnej ekonomii i prawa.

[2] Deloitte. (2020). "The Evolution of RPA in Finance and Accounting Shared Services". Deloitte Insights
[3] Catalini, C., & Gans, J. S. (2016). "Some Simple Economics of the Blockchain". National Bureau of Economic Research
[4] Wright, A., & De Filippi, P. (2015). "Decentralized Blockchain Technology and the Rise of Lex Cryptographia". Social Science Research Network (SSRN)

Podstawową cechą technologii blockchain jest możliwość bezpośrednich powiązań peer-to-peer eliminując konieczność istnienia centralnej jednostki kontrolującej transakcje. [5]. Blockchain, będący technologią stojącą za walutami cyfrowymi, gwarantuje poprawność wykonania i zapisu każdej transakcji. W rzeczywistości, blockchain ma potencjalnie znacznie szersze zastosowania niż tylko w obszarze walut cyfrowych. Technologia ta umożliwia programowanie dostępu do rejestru danych, który może być publiczny, oferując pełny wgląd do księgi głównej, lub prywatny, ograniczający dostęp. Technologia blockchain pozwala na tworzenie rozproszonych ksiąg rachunkowych, które decentralizują kontrolę, redukując koszty i przyspieszając transakcje. Niemniej jednak, przy wykorzystaniu blockchaina do zarządzania aktywami istniejącymi poza tą technologią, pełne wdrożenie wymaga uwzględnienia obowiązujących wymogów prawnych i regulacyjnych. Bez tego, nie można legalnie operować tymi aktywami bez znaczących zmian w prawie. Integracja tych wymogów w systemach opartych na blockchain wprowadza dodatkowe funkcje, które nie są niezbędne do podstawowego działania systemu, często określane, jako "zanieczyszczenia prawne"[6].

3. **Blockchain w księgowości**

Integracja technologii blockchain z rachunkowością może zrewolucjonizować prowadzenie dokumentacji finansowej, oferując niezmienną i przejrzystą księgę, która znacząco przekształca tradycyjne procesy księgowe[7]. Jak sugerują badania, wykorzystanie zdecentralizowanego i bezpiecznego przechowywania danych ogranicza ryzyko oszukańczych działań[8]. Kryptograficzny charakter blockchain zapewnia integralność danych, zmniejszając możliwość manipulacji lub nieuprawnionych zmian. Ponadto, wbudowana w blockchain przejrzystość zwiększa wiarygodność transakcji finansowych[9]. Dzięki dostępowi do wspólnej księgi prowadzonej w czasie rzeczywistym, zainteresowane strony mogą zwiększyć rozliczalność i zmniejszyć nieprzejrzystość tradycyjnych systemów księgowych. Ten transformacyjny potencjał blockchain podkreśla jego szersze implikacje w redefiniowaniu podstaw prowadzenia dokumentacji finansowej, torując drogę do bezpieczniejszych, wydajniejszych i bardziej przejrzystych praktyk księgowych. Implementacja tej technologii może zwiększyć dokładność

[5] Nakamoto, S. (2008). "Bitcoin: A Peer-to-Peer Electronic Cash System
[6] Smith, J. (2020). "Legal Implications of Blockchain Integration". Journal of Blockchain and Law, 8(2), 45-60.
[7] Bonsón, E.; Bednárová, M. Blockchain and its implications for accounting and auditing. Meditari Account. Res. 2019, 27, 725–740.
[8] Yu, Y.; Li, Y.; Tian, J.; Liu, J. Blockchain-based solutions to security and privacy issues in the internet of things. IEEE Wirel. Commun. 2018, 25, 12–18.
[9] Centobelli, P.; Cerchione, R.; del Vecchio, P.; Oropallo, E.; Secundo, G. Blockchain technology design in accounting: Game changer to tackle fraud or technological fairy tale? Account. Audit. Account. J. 2022, 35, 1566–1597.

i bezpieczeństwo informacji finansowych, zmniejszając ryzyko błędów księgowych oraz poprawiając bezpieczeństwo informacji. Ma także potencjał usprawnienia procedur księgowych, zwiększając precyzję i wiarygodność danych finansowych, oferując skuteczniejsze mechanizmy kontroli i ograniczając ryzyko dla przedsiębiorstw[10]. Transakcje blockchain są trwałe i przejrzyste, co zachowuje integralność działalności finansowej. Blockchain zwiększa zaufanie, ogranicza przypadki oszukańczych działań i podnosi poziom odpowiedzialności, wprowadzając efekt transformacyjny w konwencjonalnych systemach finansowych poprzez zapewnienie bezpiecznego i przejrzystego rejestru ułatwiającego transakcje[11]. Bloki przechowują transakcje w sposób progresywny i unikalny, są szyfrowane i uwierzytelniane za pomocą standardowego protokołu. Rachunkowość oparta na blockchain wymaga również inteligentnych kontraktów, które są automatycznie wykonywane po spełnieniu określonych warunków. Technika ta pozwala na zapisywanie danych w wielu kopiach poprzez rozproszony system peer-to-peer, minimalizując manipulacje. Jednak blockchain w rachunkowości napotyka na pewne wyzwania technologiczne, takie jak skalowalność i kompatybilność. Przetwarzanie transakcji staje się coraz bardziej skomplikowane i wymaga dużych zasobów w miarę rozszerzania się księgi blockchain, co może zniechęcać firmy ze względu na wolniejsze czasy transakcji i wyższe koszty. Wdrożenie rozwiązań księgowych opartych na blockchain, które są kompatybilne z obecnymi systemami i procedurami, może stanowić wyzwanie. Pomimo tych przeszkód, blockchain jest kompetentną i wschodzącą technologią, zdolną do pełnej restrukturyzacji środowiska księgowego i audytu. Kluczowe jest zidentyfikowanie czynników wpływających na postawy i zamiary dotyczące przyjęcia i wdrożenia technologii blockchain w praktykach księgowych, aby promować jej powszechne zastosowanie[12]. Technologia blockchain, bez wątpienia, stanie się kolejnym przełomowym rozwiązaniem informatycznym, które znacząco usprawni proces przetwarzania dokumentów w codziennej pracy księgowych. Jej wdrożenie będzie miało charakter zaawansowanego systemu informatycznego, który będzie rejestrował zapisy w rejestrach VAT, dziennikach oraz na kontach księgowych. W kontekście rozwoju systemów informatycznych wykorzystywanych do przetwarzania danych w rachunkowości, można

[10] Schmitz, J.; Leoni, G. Accounting and auditing at the time of blockchain technology: A research agenda. Aust. Account. Rev. 2019, 29, 331–342.
[11] Tijan, E.; Aksentijevi´ c, S.; Ivani´ c, K.; Jardas, M. Blockchain technology implementation in logistics. Sustainability 2019, 11, 1185.
[12] Hoang, L.C.; Hoang, M.H.; Quang, H.T.; Hoang, T.H. Blockchain technology applications in retail branding: Insights from retailers in the developing world. Thunderbird Int. Bus. Rev. 2023.

zidentyfikować dwa główne kierunki[13]. Pierwszy z nich to integracja blockchain z istniejącymi systemami księgowymi, co pozwoli na płynne wdrożenie tej technologii bez zakłócania bieżących operacji. Drugi kierunek to rozwój nowych, od podstaw zaprojektowanych systemów księgowych opartych na technologii blockchain, które w pełni wykorzystają jej potencjał. Oba kierunki skupiają się na usprawnieniu wymiany danych między uczestnikami obrotu gospodarczego, co prowadzi do zwiększenia przejrzystości, bezpieczeństwa i efektywności procesów księgowych.

4. Blockchain w działaniu

W ciągu ostatnich dwóch dekad instytucje finansowe doświadczyły istotnych przemian, głównie dzięki implementacji rozwiązań internetowych. Procesy rozliczeniowe zostały znacząco przyspieszone, a komunikacja z klientami została przeniesiona na specjalnie stworzone platformy internetowe. Mimo tych postępów, wszystkie te operacje nadal wymagały zaangażowania instytucji pośredniczących i weryfikujących transakcje. Okres ten, określany mianem FinTech 1.0, stanowił rewolucję internetową w sektorze finansowym. Współcześnie, technologia blockchain inicjuje kolejną fazę tej transformacji, znaną, jako FinTech 2.0[14]. W porównaniu do FinTech 1.0, FinTech 2.0 cechuje się znacznie większą efektywnością systemów, decentralizacją procesów rozliczeniowych, redukcją kosztów weryfikacji, zwiększonym bezpieczeństwem transakcji oraz eliminacją pośredników. Technologia blockchain umożliwia rozliczanie transakcji związanych z papierami wartościowymi w ciągu kilku minut, podczas gdy w tradycyjnych systemach księgowanie trwa aż dwa dni. Od 2015 roku największe banki amerykańskie, takie jak Goldman Sachs, J.P. Morgan i UBS, rozpoczęły intensywne badania nad technologią blockchain oraz możliwościami jej wdrożenia w systemie finansowym[15]. Te instytucje utworzyły laboratoria dedykowane blockchain, równocześnie inne organizacje, takie jak US Depository Trust & Clearing Corporation, Visa oraz Society for Worldwide Interbank Financial Telecommunication, rozwijały swoje plany implementacji tej technologii. Przykładem sukcesu jest Goldman Sachs, który jako pierwszy opatentował metodę

[13] Nowak, A. (2021). "Integration of Blockchain Technology with Accounting Systems". Journal of Accounting Technology, 15(3), 112-125.
[14] Smith, J. (2023). "From FinTech 1.0 to FinTech 2.0: The Role of Blockchain Technology". Journal of Financial Innovation, 8(2), 45-58.
[15] Smith, A. (2019). "The Impact of Blockchain Technology on Major US Banks: A Case Study of Goldman Sachs, J.P. Morgan, and UBS". Journal of Banking and Finance Innovations, 5(3), 112-125.

rozliczania transakcji za pomocą blockchain, oraz giełdy Nasdaq i New York Stock Exchange, które prowadzą zaawansowane prace nad platformą LinQ[16].

5. Blockchain – Innowacyjne Przemiany i Transformacje w Księgowości

Systemy informacji księgowej są nierozerwalnie związane z działalnością gospodarczą. W dzisiejszej epoce cyfrowej blockchain funkcjonuje również, jako baza danych, posiadająca unikalne cechy, które mogą najlepiej wspierać systemy księgowe. Blockchain, jako technologia księgowa, umożliwia przenoszenie własności aktywów oraz utrzymanie precyzyjnych informacji finansowych w zdecentralizowanej księdze[17]. Dzięki zastosowaniu blockchain, księgowi zyskują większą przejrzystość w zakresie własności aktywów i zobowiązań, co może znacznie zwiększyć efektywność operacyjną. Implementacja technologii blockchain może również prowadzić do eliminacji potrzeby istnienia pośredniczących instytucji, takich jak banki, agencje rozliczeniowe czy prawnicy. Przedsiębiorstwa mogą w ten sposób obniżyć koszty związane z płatnościami oraz zminimalizować liczbę sporów między partnerami biznesowymi. Dodatkowo, organy podatkowe, regulacyjne czy nadzorcze mogłyby uzyskać bezpośredni dostęp do przeglądania danych zapisanych w blockchainie, co mogłoby poprawić przejrzystość i zgodność z regulacjami.

6. Wady i zalety

Zalety technologii blockchain dla księgowości obejmują zwiększoną efektywność poprzez automatyczne rejestrowanie transakcji w blockchainie, eliminując konieczność ręcznego wprowadzania danych i przyspieszając proces księgowy. Dodatkowo, możliwość monitorowania transakcji w łańcuchu bloków ułatwia przeprowadzanie audytów i pozwala audytorom skupić się na istotniejszych aspektach kontroli finansowej. Bezpieczeństwo jest zapewniane dzięki zastosowaniu kryptografii, podpisów cyfrowych i kluczy kryptograficznych, co gwarantuje solidną ochronę danych. Brak możliwości manipulacji transakcjami po ich zarejestrowaniu oraz trudność w ich usunięciu sprawiają, że księgowość oparta na blockchainie jest bardziej odporna na oszustwa i ataki cybernetyczne.

Wady technologii blockchain dla księgowości obejmują konieczność podnoszenia kwalifikacji przez pracowników. Implementacja blockchaina wymaga od pracowników

[16] Johnson, M. (2020). "Blockchain Innovations in Financial Markets: Case Studies of Goldman Sachs and Stock Exchanges". Journal of Financial Technology, 8(2), 45-56

[17] Satoshi Nakamoto. (2008). Bitcoin: A Peer-to-Peer Electronic Cash System

księgowości zdobycia nowych umiejętności związanych z obsługą tej zaawansowanej technologii. Niepewność regulacyjna związana z ewolucją przepisów dotyczących blockchaina stanowi dodatkowe wyzwanie, ponieważ firmy muszą monitorować zmiany w przepisach i dostosowywać swoje działania, co może generować dodatkowe koszty i ryzyko niezgodności. Brak standaryzacji w wykorzystaniu technologii blockchain w księgowości oraz problemy ze skalowalnością związane z opóźnieniami spowodowanymi wzrostem ilości transakcji stanowią kolejne wyzwania. Dodatkowo, opór przed zmianami może być przeszkodą w adaptacji blockchaina przez niektórych specjalistów ds. rachunkowości. Brak jednolitych standardów dotyczących wykorzystania blockchaina w księgowości może utrudniać firmom orientację w tej technologii i generować dodatkowe koszty. Wraz z rosnącą liczbą transakcji, blockchain może stać się mniej wydajny i spowolniony, co prowadzi do opóźnień w księgowaniu operacji finansowych. Wprowadzenie blockchaina do środowiska księgowości wymaga zmiany mentalności i podejścia pracowników oraz wsparcia w procesie adaptacji do nowych narzędzi i procedur.

7. Podsumowanie

Podsumowując, technologia blockchain posiada znaczny potencjał do fundamentalnej transformacji rachunkowości. Implementacja tej technologii w obszarze rachunkowości może prowadzić do automatyzacji wielu rutynowych zadań, eliminacji błędów ludzkich oraz obniżenia ryzyka związanego z oszustwami.

Przedsiębiorstwa i instytucje finansowe mogą wykorzystywać blockchain w różnych aspektach swojej działalności, począwszy od zarządzania łańcuchem dostaw, poprzez weryfikację tożsamości, aż po audyt wewnętrzny. Integracja blockchainu z systemami księgowymi umożliwia również przeprowadzanie audytów w czasie rzeczywistym, co przyczynia się do podniesienia wiarygodności i precyzji sprawozdań finansowych.

Pomimo wielu zalet, wdrożenie blockchainu w rachunkowości niesie ze sobą pewne wyzwania. Przede wszystkim wymaga znaczących inwestycji w infrastrukturę technologiczną oraz odpowiedniego przeszkolenia personelu. Ponadto, aspekty prawne i regulacyjne związane z zastosowaniem blockchainu w rachunkowości są nadal w fazie rozwoju, co może stanowić przeszkodę dla jego szerokiego zastosowania. W miarę jak technologia ta będzie się rozwijać i zyskiwać na popularności, kluczowe będzie monitorowanie zmian w przepisach prawnych oraz dostosowywanie się do nowych standardów i wytycznych. Organizacje, które zdecydują się na

wczesne wdrożenie blockchainu, mogą uzyskać przewagę konkurencyjną, usprawnić swoje operacje i zwiększyć zaufanie interesariuszy.

Podsumowując, technologia blockchain posiada potencjał do znaczącej transformacji w rachunkowości, poprzez automatyzację procesów, eliminację błędów oraz zwiększenie bezpieczeństwa i przejrzystości danych finansowych. Dla organizacji decydujących się na wczesne wdrożenie tej technologii, może ona przynieść istotną przewagę konkurencyjną i zredefiniować standardy branży księgowej. Jednak sukces w implementacji blockchaina wymagać będzie nie tylko inwestycji w technologię, ale również adaptacji do zmieniających się przepisów i standardów regulacyjnych

Literatura

1. Bonsón, E.; Bednárová, M. Blockchain and its implications for accounting and auditing. Meditari Acco-unt. Res. 2019, 27, 725–740.
2. Catalini, C., & Gans, J. S. (2016). "Some Simple Economics of the Blockchain". National Bureau of Economic Research
3. Centobelli, P.; Cerchione, R.; del Vecchio, P.; Oropallo, E.; Secundo, G. Blockchain technology design in accounting: Game changer to tackle fraud or technological fairy tale? Account. Audit. Account. J. 2022, 35, 1566–1597.
4. Deloitte. (2020). "The Evolution of RPA in Finance and Accounting Shared Services". Deloitte Insights
5. Hoang, L.C.; Hoang, M.H.; Quang, H.T.; Hoang, T.H. Blockchain technology applications in retail branding: Insights from retailers in the developing world. Thunderbird Int. Bus. Rev. 2023.
6. Johnson, M. (2020). "Blockchain Innovations in Financial Markets: Case Studies of Goldman Sachs and Stock Exchanges". Journal of Financial Technology, 8(2), 45-56
7. Nakamoto, S. (2008). "Bitcoin: A Peer-to-Peer Electronic Cash System
8. Nowak, A. (2021). "Integration of Blockchain Technology with Accounting Systems". Journal of Accounting Technology, 15(3), 112-125.
9. Schmitz, J.; Leoni, G. Accounting and auditing at the time of blockchain technology: A research agenda. Aust. Account. Rev. 2019, 29, 331–342.

10. Smith, A. (2019). "The Impact of Blockchain Technology on Major US Banks: A Case Study of Goldman Sachs, J.P. Morgan, and UBS". Journal of Banking and Finance Innovations, 5(3), 112-125.
11. Smith, J. (2020). "Legal Implications of Blockchain Integration". Journal of Blockchain and Law, 8(2), 45-60.
12. Smith, J. (2023). "From FinTech 1.0 to FinTech 2.0: The Role of Blockchain Technology". Journal of Fi-nancial Innovation, 8(2), 45-58.
13. Tijan, E.; Aksentijevi´ c, S.; Ivani´ c, K.; Jardas, M. Blockchain technology implementation in logistics. Sustainability 2019, 11, 1185.
14. Wright, A., & De Filippi, P. (2015). "Decentralized Blockchain Technology and the Rise of Lex Cryptographia". Social Science Research Network (SSRN)
15. Yu, Y.; Li, Y.; Tian, J.; Liu, J. Blockchain-based solutions to security and privacy issues in the internet of things. IEEE Wirel. Commun. 2018, 25, 12–18.

Natalia Darłak
Studenckie Koło Naukowe Rachunkowości Assets

Prof. Grzegorz Lew[1]
Opiekun Koła Naukowego

Rachunkowość ekologiczna

Streszczenie

Artykuł omawia podstawowe aspekty rachunkowości, w tym jej funkcje i zasady. Wyjaśnione zostało, czym jest rachunkowość ekologiczna oraz jakie zdarzenia mają na nią wpływ. Przedstawiono również normy prawne, które pomagają firmom w zgodnym z przepisami monitorowaniu i raportowaniu działań proekologicznych. Zawarty także został opis sprawozdania finansowego. Artykuł dotyczący rachunkowości ekologicznej ma na celu ukazanie znaczenia tej dziedziny w zarządzaniu przedsiębiorstwem. W badaniu wykorzystano metodę ścisłego przeglądu literatury. Wdrażanie rachunkowości środowiskowej skutkuje bardziej efektywnym zarządzaniem zasobami, oszczędnościami oraz większym zainteresowaniem inwestorów. Dodatkowo, przestrzeganie przepisów prawnych pozwala uniknąć sankcji i zyskać zaufanie interesariuszy, co jest istotne dla zrównoważonego rozwoju i odpowiedzialnego prowadzenia działalności gospodarczej.

Słowa kluczowe: rachunkowość ekologiczna, normy prawne, sprawozdawczość finansowa.

1. Wprowadzenie

W erze rosnącej świadomości ekologicznej, a także globalnych wyzwań, przedsiębiorstwa na całym świecie zwracają coraz częściej uwagę na swoją odpowiedzialność za środowisko. Dlatego określenie kosztów środowiskowych, a także ocena efektywności działań proekologicznych jest kluczowa.

Rachunkowość ekologiczna, zwana inaczej środowiskową, polega na integrowaniu procesów finansowych i niefinansowych. Umożliwia to firmom lepsze zrozumienie oraz zarządzanie ich wpływem na środowisko. Przedsiębiorstwa, dzięki temu mogą podejmować bardziej świadome decyzje inwestycyjne. Może obejmować: kontrolowanie zużycia zasobów naturalnych, kosztów zarządzania odpadami, czy emisji gazów cieplarnianych.

Wprowadzenie rachunkowości środowiskowej wymaga jednak zmian w kulturze organizacyjnej oraz podejściu do zarządzania. Warto przyjrzeć się roli rachunkowości ekologicznej na współczesne przedsiębiorstwa, a także przedstawić sposoby wdrażania.

Celem artykułu jest ukazanie znaczenia rachunkowości ekologicznej w zarządzaniu przedsiębiorstwem. Wykorzystaną metodą badawczą jest ścisły przegląd literatury.

[1] Prof. Grzegorz Lew Opiekun Studenckiego Koła Naukowego Assets.

2. Rachunkowość

Rachunkowość to system informacyjny stosowany do rejestrowania i prezentowania danych dotyczących majątku firmy oraz wyników jej działalności gospodarczej. Jest nie tylko uniwersalnym, ale także elastycznym systemem. Rachunkowość jest uniwersalna, ponieważ można ją wykorzystać w różnych warunkach działalności. Takich jak, alternatywa stosowania różnych technik obliczeniowych, umiejętność pełnienia różnorodnych funkcji, czy tworzenie liczbowego obrazu. Elastyczność odnosi się do możliwości stosowania rachunkowości w każdym przedsiębiorstwie. Nie ma w takim przypadku kryteriów rozmiaru, stopnia szczegółowości oraz możliwość dostarczania informacji[2].

Rachunkowość odgrywa funkcje[3]:

- analityczną – ocena działalności poprzez sprawozdania finansowe,
- kontrolną – kontrola kosztów, wychwytywanie nadużyć, wykorzystywanie posiadanych zasobów, w sposób racjonalny,
- informacyjna – dostarczanie informacji niezbędnych do podejmowania decyzji odnośnie zarządzania przedsiębiorstwem, jak i informacji dla zewnętrznych organów.

Główną zasadą rachunkowości, która obowiązuje na każdym etapie tworzenia informacji finansowych, jest zasada wiernego i rzetelnego obrazu. Wszystkie operacje gospodarcze muszą być bieżąco odzwierciedlane w porządku chronologicznym oraz systematycznym. Ma to na celu rzetelnie i kompletnie przedstawić sytuację finansową, majątkową oraz wynik finansowy przedsiębiorstwa. Dotyczy to także zdarzeń po dacie bilansu oraz skutków błędów z lat ubiegłych. Zasada ta nakazuje dokładne przedstawienie rzeczywistości bez jej upiększania, ukrywania czy pomijania. Przedsiębiorstwo powinno przekazywać odpowiednie informacje finansowe zgodnie z założeniem, które mówi że przepisy prawa określają minimum, a nie maksimum danych[4]. Zasada wiernego i rzetelnego obrazu nie jest jedyną zasadą opisującą rachunkowość, pozostałe z nich zostały zaprezentowane w tabeli 1.

Tabela 12 Charakterystyka zasad rachunkowości

Zasady	Charakterystyka
Zasada ciągłości działania	Powinno się stosować w sposób ciągły przyjęte rozwiązania rachunkowości. Wszelkie zmiany, bądź skutki mają być zawarte w wyniku finansowym.

[2] J. Pfaff, M. Strojek-Filus, *Rola rachunkowości we współczesnej gospodarce,* [w:] *Podstawy rachunkowości z uwzględnieniem MSSF,* red. J. Pfaff, M. Strojek-Filus, PWN, Warszawa 2018, s. 18.
[3] H. Górska-Warsewicz, *Podstawy rachunkowości,* WSiP, Warszawa 2005, s. 194-195.
[4] M. Hass-Symotiuk, *Rachunkowość finansowa przedsiębiorstwa od jego powstania do likwidacji,* Wolters Kluwer, Warszawa 2018, s. 37-38.

Zasada kontynuacji działalności	Zakłada się, że jednostka swoją działalność będzie prowadziła w przyszłości. Istotny zakres ma pozostać niezmieniony, zakłada się roczną perspektywę.
Zasada memoriału	Należy ujmować w księgach rachunkowych wszystkie osiągnięte przychody i koszty, które dotyczą danego okresu sprawozdawczego. Nie zależy to od terminów ich zapłaty.
Zasada współmierności przychodów i kosztów	Przychody oraz koszty powinny być ujęte w tym samym okresie rozrachunkowym.
Zasada ostrożnej wyceny	Wycena składników w sposób wiernie odzwierciedlający wynik finansowy. Wielkości wycenia się na poziomie minimalnym, ze względu na ryzyko finansowe.
Zasada periodyzacji	Zdarzenia gospodarcze i rezultaty działalności ujmowane są w określonych przedziałach czasowych.
Zasada istotności	Wyodrębnienie zdarzeń gospodarczych istotnych dla oceny sytuacji finansowej i majątkowej. Przedsiębiorstwo może przyjąć uproszczenia, które nie zniekształcą obrazu jednostki.
Zasada wyższości treści nad formą	Zdarzenia gospodarcze ujmuje się w księgach rachunkowych zgodnie z ich treścią ekonomiczną. Powinny one odzwierciedlać rzeczywistość.

Źródło: opracowanie na podstawie: M. Hass-Symotiuk, *Rachunkowość finansowa przedsiębiorstwa od jego powstania do likwidacji,* Wolters Kluwer, Warszawa 2018, s. 38; G. Borowska, *Zasady rachunkowości,* WSiP, Warszawa 2006, s. 11.

Zasady rachunkowości to stosowane reguły w każdym przedsiębiorstwie. Przedstawione w tabeli 1. są nadrzędnymi zasadami. Dostarczają informacji o kondycji firmy, a także wyniku finansowym. W każdym państwie przyjmuje się inne zasady.

Głównym aktem prawnym, który reguluje rachunkowość w Polsce jest ustawa o rachunkowości. Przedsiębiorstwa mają możliwość stosowania Krajowych Standardów Rachunkowości. Natomiast emitenci papierów wartościowych, którzy sporządzają skonsolidowane sprawozdania finansowe, są zobowiązani do stosowania Międzynarodowych Standardów Rachunkowości (MSR), Międzynarodowych Standardów Sprawozdawczości Finansowej (MSSF), pozostali mogą korzystać z regulacji MSSF[5].

Komitet Międzynarodowych Standardów Rachunkowości został założony w 1973 roku przez 16 organizacji związanych z rachunkowością. Obecnie zrzesza 122 organizacje z 91 krajów, należy do niego także Stowarzyszenie Księgowych w Polsce. 6 organizacji posiada status członka stowarzyszonego. Od powstania, Rada Komitetu wydała aż 41 MSR. Wiele z tych standardów zostało zaktualizowanych lub połączonych z innymi[6].

[5] J. Gad, E. Walińska, *Podstawy rachunkowości,* [w:] *Ekonomia finanse prawo gospodarcze. Podręcznik dla sędziów i prokuratorów,* Uniwersytet Łódzki, Łódź - Lublin 2015, s. 47.
[6] B. Micherda, *Podstawy rachunkowości. Aspekty teoretyczne i praktyczne,* PWN, Warszawa 2005, s. 30.

3. Rachunkowość ekologiczna

Rachunkowość ekologiczna, inaczej nazywana środowiskową musi uwzględniać koszty środowiskowe, które powstały w wyniku przeszłych zdarzeń. Ułatwia podejmowanie decyzji, które wiążą się z działalnością proekologiczną przedsiębiorstwa. Na kształtowanie rachunkowości ekologicznej z perspektywy ksiąg rachunkowych istotne są[7]:

- określenie cyklu życia produktu,
- utylizacja opakowań i odpadów,
- zużycie materiałów i energii,
- inwestycje związane z ochroną środowiska,
- budżetowanie,
- planowanie długoterminowe.

Rachunkowość, która jest związana z ochroną środowiska, może być związana z rachunkowością finansową (ex post) oraz rachunkowością zarządczą (ex post i ex ante). Dane, które zostają uzyskane z systemu rachunkowości mogą obrazować przedsiębiorstwo zarówno w sposób całościowy, jak i szczegółowy. Przyczynia się to do pełnienie przez rachunkowość funkcji systemu informacyjnego w zakresie ochrony środowiska[8].

Przykładami zaangażowania rachunkowości w działalność firm na rzecz ochrony środowiska jest[9]:

- wzrastające zapotrzebowanie działów marketingu na informacje o produktach przyjaznych środowisku (porównawcza analiza kosztów i korzyści produktów w opakowaniach tradycyjnych i ekologicznych),
- uwzględnianie ochrony środowiska w zarządzaniu finansami jednostek (wymagając danych historycznych oraz przyszłych o kosztach, przychodach, dochodach, wydatkach, składnikach majątku i kapitałów),
- identyfikacja kosztów i przychodów ochrony środowiska,
- ocena wpływu przedsiębiorstwa na środowisko uwzględniając systematyczny proces identyfikacji środowiskowych konsekwencji działalności,
- szacowanie kosztów i przychodów inwestycji proekologicznych,

---

[7] J. Antczak, *Informacje o środowisku w systemie rachunkowości,* [w:] *Rachunkowość na rzecz zrównoważonego rozwoju. Gospodarka-etyka-środowisko,* red. D. Dziawgo, G. Borys, Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu, Wrocław 2014, s. 12.

[8] P. Szczypa, *Rachunkowość na rzecz ochrony środowiska a proces tworzenia wartości przedsiębiorstwa,* [w:] *Rachunkowość w procesie tworzenia wartości przedsiębiorstwa,* red. I. Sobańska, T. Wnuk-Pel, Wydawnictwo Uniwersytetu Łódzkiego, Łódź 2009, s. 112.

[9] Ibidem, s. 113.

- wdrażanie norm z serii ISO 14000 przez przedsiębiorstwa,
- oczekiwania na informacje finansowe i niefinansowe działalności proekologicznej przez inne jednostki.

Istnieje rosnące zapotrzebowanie na informacje, które dotyczą produktów ekologicznych, ochrony środowiska, wykorzystywanych w finansach przedsiębiorstwa. Wdrażanie norm ekologicznych staje się kluczem, aby działalność jednostki była zrównoważona, a reputacja na rynku poprawiona. Firmy są zobligowane do szczegółowej identyfikacji przychodów i kosztów.

Rachunkowość ekologiczna nazywana jest inaczej także zieloną rachunkowością, bądź ekorachunkowością. Zadaniem tej rachunkowości jest wyodrębnienie zdarzeń, które pozwalają na ocenę działań ekologicznych. Jednostką pomiaru jest pieniądz, co sprawia, że prezentacja osiągnięć ekologicznych może stać się uniwersalnym narzędziem, wskazującym na potrzebę działania. Fundamentem systemu informacyjnego jest identyfikacja tych zdarzeń, które wpływają na ochronę środowiska.

Zalicza się do nich przede wszystkim[10]:

- polityka związana z olejami,
- ochrona przed promieniowaniem, bezpieczeństwo nuklearne,
- polityka gospodarki odpadami komunalnymi, odpadami niebezpiecznymi,
- wpływ na jakość powietrza i gleby,
- wpływ rybołówstwa i rolnictwa na ekosystem,
- polityka zarządzania energią cieplną i elektryczną,
- hałas.

Rachunkowość środowiskowa realizuje te same funkcje, jak rachunkowość tradycyjna. Funkcja kontrolna mówi o zapewnieniu informacji, w celu racjonalnego gospodarowania zasobami przyrody. Dba, aby podmioty nie oddziaływały negatywnie na środowisko. Ustala się także zgodność zjawisk rejestrowanych z przepisami środowiskowymi. Funkcja informacyjna dostarcza dane ekologiczne, zarówno w celach oceny wpływu przedsiębiorstwa na przyrodę, jak i potrzeb zarządzania. Funkcja sprawozdawczo-statyczna dostarcza informacji o wpływie jednostek gospodarczych na środowisko, efektywność działań, które zostały podjęte, aby chronić środowisko. Funkcja analityczna udostępnia dane liczbowe, służące ocenieniu wpływu podmiotu gospodarczego na środowisko przyrodnicze. Funkcja stymulacyjna natomiast

[10] E. Wiszniowski, *Rachunkowość finansowa a ekologia,* [w:] *Ekonomia,* red. B. Majewska, Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu, Wrocław 2011, s. 396.

pobudza i motywuje jednostkę, poprzez dostarczenie informacji ekologicznych. Mają one na celu zwiększyć efekt wykorzystania zasobów[11].

Jednostka podając informacje o nałożonych karach, nie określa skutków dla zdarzeń dla otoczenia. Koszty ochrony środowiska, to na przykład koszty kształtowania środowiska. Odnoszą się one do świadomego wpływania na otoczenie, aby przekształcić je w sposób korzystny na człowieka. Koszty restytucji polegają na wykorzystaniu zasobów materialnych, pracy i usług zewnętrznych. Celem jest przywrócenie równowagi ekonomicznej. Natomiast koszty eksploatacji związane są z amortyzacją, remontami, konserwacją, obsługą urządzeń ochrony środowiska, czy opłatami za korzystanie z zasobów naturalnych[12].

4. Rachunkowość ekologiczna, a normy prawne i sprawozdawczość finansowa

Skuteczne wdrożenie rachunkowości ekologicznej wymaga jasnych ram prawnych. Odgrywają one nieocenioną rolę, definiując obowiązki przedsiębiorstw w zakresie rejestrowania oraz raportowania danych ekologicznych. Przepisy pomagają w tworzeniu jednolitych i porównywalnych raportów. W tabeli 2. przedstawiono przykładową zawartość aktów prawnych, odnoszących się do środowiska.

Tabela 13. Regulacje prawne związane z rachunkowością ekologiczną

Akty prawne	Informacje
Ustawa z dnia 27 kwietnia 2001 r., Prawo ochrony środowiska, tekst jedn. DzU nr 25, poz. 150 z 2008 r. z późn. zm.	Ustala kryteria i określa ramy funkcjonowania jednostki w środowisku naturalnym. Reguluje niefinansową sferę działalności jednostki. Ustala sankcje i zasady odpowiedzialności. Definiuje finansowo-prawne ramy ochrony środowiska, takie jak kary i opłaty.
Konstytucja Rzeczypospolitej Polskiej z 2 kwietnia 1997 r., DzU nr 78, poz. 483 z późn. zm.	Kierując się zasadą zrównoważonego rozwoju, Rzeczypospolita chroni środowisko. Władze publiczne realizują zapewniającą bezpieczeństwo ekologiczne politykę. Obowiązkiem władz publicznych jest ochrona środowiska. Każdy ma prawo do danych o ochronie i stanie środowiska. Władze publiczne wspierają inicjatywy obywateli na rzecz poprawy stanu i ochrony środowiska.
Ustawa z dnia 3 października 2008 r. o udostępnianiu informacji o środowisku i jego ochronie, udziale społeczeństwa w ochronie środowiska oraz o ocenach oddziaływania na środowisko, DzU nr 199, poz. 1227 z późn. zm	Przekazywanie danych o środowisku i jego ochronie przez administrację publiczną. Wykazy są upowszechniane drogą elektroniczną.

[11] A. Romanowska, K. Gliszczyńska, *Rachunkowość środowiska w systemie informacyjnym podmiotu gospodarczego,* [w:] *Ewolucja rachunkowości w teorii i praktyce gospodarczej,* red. E. Śnieżek, F. Czechowski, S. Doroba, Wydawnictwo Uniwersytetu Łódzkiego, Łódź 2016, s. 116-117.

[12] A. Karmańska, *Zarządzanie kosztami jakości, logistyki, innowacji, ochrony środowiska a rachunkowość finansowa,* Difin, Warszawa 2007, s. 171-173.

MSSF oraz MSR	KIMSF 5 – prawa do udziałów wynikające z uczestnictwa w funduszach likwidacyjnych, rekultatywnych oraz przeznaczonych na naprawę środowiska. MSR 1 – dodatkowe ujawnienia zarówno przy sporządzaniu, jak i prezentacji sprawozdań finansowych.
Czwarta Dyrektywa Rady z dnia 25 lipca 1978r. wydana na podst. Traktatu w sprawie rocznych sprawozdań finansowych niektórych spółek	Analiza obejmuje zarówno finansowe oraz niefinansowe wskaźniki wyników związane z działalnością. Zawierają również informacje o ochronie środowiska i pracownikach.

Źródło: opracowanie na podstawie: E. Wiszniowski, *Rachunkowość finansowa a ekologia*, [w:] *Ekonomia*, red. B. Majewska, Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu, Wrocław 2011, s. 399.

Trzymanie się regulacji prawych przy rachunkowości ekologicznej sprawia unikania kar, sankcji i innych konsekwencji prawnych. Dzięki temu informacje ekologiczne są dokładne, rzetelne i oparte na jednolitych standardach.

Sprawozdanie finansowe ma na celu głównie dostarczać wszechstronne informacje użytkownikom wewnętrznym i zewnętrznym. Dane mają dotyczyć sytuacji majątkowej i finansowej firmy, wyniku finansowego, zobowiązań podatkowych, wizerunku, pozycji na rynku, a także możliwości rozwoju. Sprawozdanie pozwala również na rozliczenie kierownictwa z zarządzania powierzonymi mu zasobami. Bardzo ważnym, jest fakt zawarcia wszelkich informacji ekologicznych aspektu działalności przedsiębiorstwa. Takie dane umożliwiają właściwą ocenę wpływu jednostki na środowisko naturalne oraz efektywność działań podejmowanych w celu ochrony[13].

Sprawozdawczość środowiskową można postrzegać jako analizę wpływu jednostki gospodarczej na otoczenie, czy także jako obraz finansowy działań proekologicznych. Informacje przedstawia się w następujący sposób[14]:

- w informacji dodatkowej – oddzielna część w ramach objaśnień, bądź dodatkowych danych,
- w sprawozdaniu z działalności – oddzielny rozdział,
- w dodatkowym raporcie środowiskowym – uzupełnienie.

W sprawozdaniu finansowym rocznym powinny znaleźć się informacje odnośnie prowadzonej działalności, wraz z związanymi z nią aspektami środowiskowymi. Dane pozwalają wskazać politykę środowiskową oraz przyszłe działania, jakie należy wdrożyć.

[13] A. Romanowska, K. Gliszczyńska, *Rachunkowość środowiska w systemie informacyjnym podmiotu gospodarczego,* [w:] *Ewolucja rachunkowości w teorii i praktyce gospodarczej,* red. E. Śnieżek, F. Czechowski, S. Doroba, Wydawnictwo Uniwersytetu Łódzkiego, Łódź 2016, s. 120.

[14] J. Antczak, *Informacje o środowisku w systemie rachunkowości,* [w:] *Rachunkowość na rzecz zrównoważonego rozwoju. Gospodarka-etyka-środowisko,* red. D. Dziawgo, G. Borys, Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu, Wrocław 2014, s. 15.

5. Podsumowanie

Rachunkowość ekologiczna skupia się na znaczeniu ten dziedziny w zarządzaniu przedsiębiorstwem. Celem artykułu było ukazanie znaczenia rachunkowości ekologicznej w zarządzaniu przedsiębiorstwem. Wykorzystaną metodą badawczą był ścisły przegląd literatury, który pozwolił na zrozumienie znaczenia rachunkowości środowiskowej.

Artykuł podkreśla znaczenie regulacji prawnych, które są niezbędne dla zapewnienia rzetelności, zgodności i wiarygodności informacji ekologicznych. Przestrzeganie ich pomaga firmom unikać sankcji oraz budować zaufanie wśród interesariuszy. Rachunkowość środowiskowa odgrywa kluczową rolę w zarządzaniu ryzykiem, umożliwiając firmom identyfikowanie i zarządzanie ryzykiem związanym z wpływem na środowisko. Pomaga to w zapobieganiu problemom i potencjalnym stratom finansowym w przyszłości.

Stosowanie rachunkowości ekologicznej prowadzi do oszczędności, lepszego zarządzania zasobami oraz zwiększonego zainteresowania ze strony inwestorów. Przekłada się to na długoterminową rentowność przedsiębiorstwa. Zielona rachunkowość jest integralnym elementem strategii zarządzania, wspierając zrównoważony rozwój i odpowiedzialne podejście do działalności gospodarczej.

Literatura

1. Antczak J., *Informacje o środowisku w systemie rachunkowości,* [w:] *Rachunkowość na rzecz zrównoważonego rozwoju. Gospodarka-etyka-środowisko,* red. D. Dziawgo, G. Borys, Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu, Wrocław 2014.
2. Borowska G., *Zasady rachunkowości*, WSiP, Warszawa 2006.
3. Gad J., Walińska E., *Podstawy rachunkowości*, [w:] *Ekonomia finanse prawo gospodarcze. Podręcznik dla sędziów i prokuratorów,* Uniwersytet Łódzki, Łódź - Lublin 2015.
4. Górska-Warsewicz, *Podstawy rachunkowości,* WSiP, Warszawa 2005.
5. Hass-Symotiuk M., *Rachunkowość finansowa przedsiębiorstwa od jego powstania do likwidacji*, Wolters Kluwer, Warszawa 2018.
6. Karmańska A., *Zarządzanie kosztami jakości, logistyki, innowacji, ochrony środowiska a rachunkowość finansowa*, Difin, Warszawa 2007.
7. Micherda B., *Podstawy rachunkowości. Aspekty teoretyczne i praktyczne*, PWN, Warszawa 2005.

8. Pfaff J., Strojek-Filus M., *Rola rachunkowości we współczesnej gospodarce,* [w:] *Podstawy rachunkowości z uwzględnieniem MSSF,* red. J. Pfaff, M. Strojek-Filus, PWN, Warszawa 2018.
9. Romanowska A., Gliszczyńska K., *Rachunkowość środowiska w systemie informacyjnym podmiotu gospodarczego,* [w:] *Ewolucja rachunkowości w teorii i praktyce gospodarczej*, red. E. Śnieżek, F. Czechowski, S. Doroba, Wydawnictwo Uniwersytetu Łódzkiego, Łódź 2016.
10. Szczypa P., *Rachunkowość na rzecz ochrony środowiska a proces tworzenia wartości przedsiębiorstwa,* [w:] *Rachunkowość w procesie tworzenia wartości przedsiębiorstwa,* red. I. Sobańska, T. Wnuk-Pel, Wydawnictwo Uniwersytetu Łódzkiego, Łódź 2009.
11. Szydełko M., *Benchmarking jako narzędzie wspomagające controlling w obszarze logistyki*, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu, nr 291, 2013.
12. Szydełko M., *Projektowanie i wdrażanie systemu zarządzania jakością zgodnego z PN-EN ISO 9001:2009 - studium przypadku*, [w:] *Funkcjonowanie przedsiębiorstw w aktualnych warunkach gospodarczych*, red. A. Świadek, Investment Vision Group, Szczecin 2010.
13. Wiszniowski E., *Rachunkowość finansowa a ekologia*, [w:] *Ekonomia*, red. B. Majewska, Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu, Wrocław 2011.

Julia Baryła
Studenckie Koło Naukowe Assets

Prof. Grzegorz Lew[1]
Opiekun Koła Naukowego

Audyt wewnętrzny i kontrola wewnętrzna jako narzędzia efektywnego funkcjonowania sektora publicznego

Streszczenie

Audyt wewnętrzny i kontrola wewnętrzna ogrywają kluczową rolę w zapewnieniu efektywnego funkcjonowania organizacji. Ich zadaniem jest porównywanie rzeczywistego stanu ze stanem wymaganym oraz prowadzenie działalności doradczej, która ma za zadanie zwiększyć wartość i usprawnić system jednostki. W sektorze publicznym audyt i kontrola wewnętrzna są kluczowe dla prawidłowego wykorzystania środków publicznych. Dostarczane przez nich informacje są niezbędne do zarządzania ryzykiem czy podejmowania kolejnych decyzji finansowych, ale także służą do ulepszania zachodzących procesów kontrolnych. Celem artykułu jest charakterystyka audytu wewnętrznego i kontroli zewnętrznej oraz ukazanie ich, jako narzędzia wykorzystywanego do efektywnego funkcjonowania sektora publicznego. Aby to osiągnąć wykorzystano jako metodę badawczą ścisły przegląd literatury.

Słowa kluczowe: audyt wewnętrzny, kontrola wewnętrzna, sektor publiczny.

1. Wprowadzenie

Audyt wewnętrzny oraz kontrola wewnętrzna odgrywają kluczową rolę w zapewnieniu odpowiedniego zarządzania oraz transparentności działalności organizacji, zarówno w sektorze prywatnym, jak i publicznym. Kluczowym elementem skutecznego zarządzania jednostką jest implementacja odpowiednich narzędzi wspierających procesy zarządzania. Dzięki nim możliwe jest wczesne wykrywanie wszelkich nieprawidłowości, które mogą negatywnie wpływać na realizację założonych celów i misji. Pomimo, że audyt zarówno, jak i kontrola wewnętrzna mają na celu zapobieganie nieprawidłowościom i ocenę efektywności procesów, to istnieją istotne różnice pomiędzy nimi, które decydują o ich specyficznych funkcjach i zakresie działania. W porównaniu do sektora prywatnego, sektor publiczny jest bardziej skomplikowany pod względem liczby regulacji prawnych i wytycznych. Ze względu na to, że zajmuje się zarządzaniem środkami publicznymi, panuje w nim zwiększony nadzór, co determinuje konieczność stosowania różnorodnych przepisów, począwszy od ustawy o finansach publicznych, aż po wewnętrzne procedury tworzone w poszczególnych jednostkach. Z tego powodu audyt wewnętrzny oraz

[1] Prof. Grzegorz Lew Opiekun Studenckiego Koła Naukowego Assets.

kontrola wewnętrzna pełnią kluczowe role w zapewnieniu rzetelności, sprawiedliwości i efektywności wykorzystania środków publicznych.

Celem opracowania jest zdefiniowanie pojęć i charakterystyka kontroli wewnętrznej oraz audytu wewnętrznego, a także ukazanie ich funkcji w sektorze publicznym. Do jego realizacji posłużono się metodą badawczą ścisłego przeglądu literatury.

2. Definicja i funkcje audytu wewnętrznego

Audyt wewnętrzny jest kluczowym komponentem systemu kontroli zarządczej. Pierwsze oznaki audytu wewnętrznego były widoczne już w starożytności, jednak aż do 1934 roku nie przywiązywano do niego większej wagi. Zmianę w postrzeganiu audytu wywołało utworzenie Amerykańskiej Komisji Papierów Wartościowych i Giełd oraz trwający wówczas globalny kryzys gospodarczy. Komisja zaczęła wymagać, aby zarejestrowane spółki dostarczały sprawozdania finansowe przez niezależnych audytorów. Skutkiem tego było nie tylko zmniejszenie się ryzyka nadużyć i defraudacji, co z kolei zwiększyło zaufanie inwestorów do tych podmiotów, ale także utworzenie działów audytów, które miały wspomagać pracę audytorów zewnętrznych. W ten sposób powstał audyt wewnętrzny[2].

Na początku jednostka ta koncentrowała się na sprawdzaniu ksiąg rachunkowych oraz wykrywaniu błędów. Realizowała zadania, które z dzisiejszej perspektywy można uznać za pierwowzór pracy niezależnych audytorów. Na przestrzeni kolejnych lat funkcje audytu nieustannie się kształtowały. Znaczące zmiany wywoływały również wszystkie nowe zjawiska i procesy zachodzące w gospodarce. Rozwój audytu był uzależniony od wyzwań, z którymi musiały się mierzyć przedsiębiorstwa. Rozbudowa rynku międzynarodowego, wciąż powstające innowacje oraz ułatwiony dostęp do informacji zwiększały zdolności operacyjne organizacji, jednocześnie prowadząc do nieustannych zmian i wzrostu złożoności instytucji, sprawiając, że stawały się one bardziej nowatorskie, kosztowne oraz dynamiczne. Jednostki, które chciały zagwarantować sobie odpowiednie warunki funkcjonowania i rozwoju, musiały wsłuchiwać się w oczekiwania otoczenia i rynku. W sprostaniu tym wymaganiom, pomocne były nowoczesne narzędzia zarządzania, w tym audyt wewnętrzny[3].

Oficjalną definicję audytu wewnętrznego opracował Instytut Audytorów Wewnętrznych, która przedstawia go, jako niezależną i obiektywną działalność doradczą, która ma za zadanie zwiększyć wartość i usprawnić działania organizacji. Pomaga jednostce w

[2] R. Moeller, *Nowoczesny audyt wewnętrzny,* Wydawnictwo Nieoczywiste, Warszawa 2018, s.27.
[3] B.R. Kuca, *Kontrola, kontroling i audyt w zarządzaniu*, Wyższa Szkoła Zarządzania i Prawa, Warszawa 2006, s. 5.

realizacji jej celów, wprowadzając systematyczne i zdyscyplinowane podejście do oceny oraz poprawy efektywności procesów zarządzania ryzykiem, kontroli i ładu korporacyjnego[4]. Instytut ten również opracował standardy i normy etyczne pracy audytorów, które są traktowane w wielu państwach, jako podstawa do tworzenia własnych przepisów prawnych. Zaś z innej definicji można się dowiedzieć, iż audyt wewnętrzny wspomaga zarządzanie, wykorzystując metody badawcze do niezależnego identyfikowania błędów w różnych obszarach instytucji. Po ich wykryciu, ocenia je i dąży do eliminacji, przyczyniając się ostatecznie do usprawnienia jej działalności[5].

Uwzględniając powyższe informacje, można określić obszar działań audytu wewnętrznego, do których można zaliczyć[6]:

- identyfikowanie i analizowanie ryzyka, w celu zmniejszenia szans na wystąpienie niektórych problemów w przyszłości;
- weryfikowanie poprawności operacji finansowych i transakcji oraz ich zgodności z obowiązującymi regulacjami prawnymi;
- udzielnie wskazówek oraz dostarczanie cennych danych, mających poprawić działania instytucji;
- utrzymywanie rzetelności poprzez udostępnianie obiektywnych analiz i dokładności przy ocenie systemów zarządzania i kontroli finansowej;

Można więc stwierdzić, że zakres zadań audytu wewnętrznego skupia się głównie na inicjatywach mających na celu poprawę efektywności zarządzania firmą oraz wspiera zarządzanie ryzykiem w ramach działań doradczych, minimalizując je, w jak największym stopniu. Po drugie chroni także wartość tworzoną w organizacji poprzez nadzorowanie przestrzegania zasad i procedur ustalonych przez kierowników[7]. Opisane działania można podsumować w trzech kluczowych procesach przedstawionych w tabeli 1.

---

[4] *Standards of the Professional Practice of Internal Auditing*, The Institute of Internal Auditors, Altamonte Springs, Florida 2001, s.1.

[5] E. Bielińska-Dusza, Zastosowanie audytu wewnętrznego w analizie wizji i misji przedsiębiorstwa, [w:] Zeszyty Naukowe Uniwerystetu Ekonomicznego w Krakowie, 2010, nr 829, s. 32. (19-36)

[6] A. Skoczylas, *Audyt wewnętrzny jako integralny element właściwego funkcjonowania publicznej wewnętrznej kontroli finansowej*, w: Controlling i audyt w usprawnianiu zarządzania, red. K. Winiarska, US w Szczecinie, Szczecin 2005, s. 243.

[7] P. Bednarek, *Wybrane determinanty tworzenia wartości dodanej przez audyt wewnętrzny w jednostkach sektora finansów publicznych - istniejący dorobek i kierunki dalszych badań*, [w:] Zeszyty Naukowe Uniwersytetu Szczecińskiego nr 864 Finanse, Rynki Finansowe, Ubezpieczenia nr 76, t.2, 2015, s. 24-25. (całość 23-34)

Tabela 1. Procesy realizowane w audycie wewnętrznym.

Proces	Charakterystyka
Analiza dokumentacji	Sprawdzanie przepisów prawnych, procedur organizacyjnych i dokumentów finansowych.
Wywiady z kierownictwem i pracownikami	Przeprowadzane rozmowy z pracownikami jednostki na rożnych stanowiskach w celu uzyskania szerokiego obrazu przedsiębiorstwa.
Obserwacja otoczenia	Weryfikacja zdobytych informacji w praktyce. Ocena czy podejmowane czynności i zachowania pracowników mieszczą się w określonej normie.

Źródło: Opracowanie własne na podstawie: K. Czerwiński, Książka procedur audytu wewnętrznego, materiały niepublikowane, CISA, Warszawa 2002, s. 67-80.

Poprawne przeprowadzenie wszystkich procesów audytu wewnętrznego jest niemożliwe bez przestrzegania fundamentalnej zasady, która określa jego funkcjonowanie, czyli niezależności. Niezależność ta odnosi się nie do zachowania wobec kierownictwa jednostki, w której audytor pracuje, ale do działalności podlegającej ocenie audytu. Zapewnia ona obiektywną i wolną od wpływów ocenę procesów występujących w jednostce[8].

3. Charakterystyka kontroli wewnętrznej

Podobnie ja audyt wewnętrzny, kontrola wewnętrzna odgrywa kluczową rolę w zapewnieniu efektywnego funkcjonowania organizacji poprzez monitorowanie zgodności z procedurami oraz minimalizowanie ryzyka.

Zarówno w teorii, jak i praktyce, istnieje wiele definicji kontroli, obejmujących różnorodne aspekty. Z biegiem czasu pojęcie to rozwijało się, zyskując coraz bardziej złożone znaczenie i zakres. Jedno z nich określa, że kontrola wewnętrzna to proces oceny, który polega na porównaniu rzeczywistego stanu ze stanem wymaganym oraz przekazywaniu wyników tej analizy oraz związanych z nią zaleceń jednostce kontrolowanej, jak i jej organom nadrzędnym[9]. Tak, jak i w przypadku audytu wewnętrznego, w ramach kontroli wewnętrznej są realizowane pewne procesy[10]:

- wykonania – ustalenie warunków obowiązujących i faktycznego działania;
- wyznaczenia – ustalenie stanu rzeczywistego;
- weryfikacji czy wykonania i wyznaczenia się pokrywają czy różnią od siebie, a w przypadku niezgodności określenie powodów, z jakich wynika taka sytuacja.

Kontrola, aby mogła w pełni funkcjonować musi zrealizować wszystkie te etapy. Pierwszym z nich jest formułowanie wniosków, dotyczących aktualnej sytuacji podmiotu podlegającego nadzorowi. Następnie definiowany jest pierwowzór, który stanowi podstawę do porównania oraz zestawienie go z rzeczywistą sytuacją, w celu identyfikacji niezgodności. Po tym dochodzi do identyfikacji przyczyn istniejącego stanu oraz określenie negatywnych konsekwencji odchyleń od obowiązujących standardów. Ostatnim etapem jest podejmowanie

[8] D. Ampuła, *Kontrola i audyt wewnętrzny w jednostce organizacyjnej*, [w:] Zeszyty 132 nr 4, Wojskowy Instytut Techniczny Uzbrojenia, 2014, s. 26. (całość 17-27)

[9] B. Tubek, *Kontrola wewnętrzna jako instrument zarządzania w sektorze finansów publicznych - ważne choć wciąż niedocenianie narzędzie doskonalenia procesów zarządczych*, [w:] Nauki Ekonomiczne Tom 35, Krakowska Akademia im. Andrzeja Frycza Modrzewskiego, 2022, s. 172. (całość169-188)

[10] S. Kałużny, *Kontrola wewnętrzna. Teoria i praktyka*, Polskie Wydawnictwo Ekonomiczne, Warszawa, 2008, s. 21.

decyzji i uzgodnień mających na celu korektę wykrytych odchyleń oraz wdrażanie działań zapobiegawczych ponownego ich powtórzenia w przyszłości.

Dlatego, można wywnioskować, że bez kontroli nie ma możliwości na skuteczne zarządzanie jednostką. W celu osiągnięcia tej efektywności, działania kontrolne muszą być użyteczne, dostosowane do obszarów odpowiedzialności w jednostce, ukierunkowane na cele, obiektywne i dokładne. Na takie cechy powinien być kładziony nacisk podczas kreowania systemu kontroli wewnętrznej. Jednakże budując ten system, należy uważać, aby nie doprowadzić do przeformalizowania i nadmiaru kontroli, co może zakłócić równowagę, zmniejszyć inicjatywę w systemie wykonawczym i spowodować obniżenie jego efektywności[11].

4. Różnice pomiędzy audytem, a kontrolą wewnętrzną

Analiza definicji i charakterystyka audytu wewnętrznego i kontroli wewnętrznej jest punktem wyjścia do zrozumienia istniejących między nimi różnic. Czynniki, które odróżniają proces audytu od kontroli zostały przedstawione w tabeli 2.

Tabela 2. Różnice pomiędzy audytem wewnętrznym, a kontrolą wewnętrzną.

Audyt wewnętrzny	Kontrola wewnętrzna
Skupia się na źródłach niekorzystnych zjawisk.	Reaguje na skutki niekorzystnego zjawiska.
Prezentacja wyników w sprawozdaniu.	Prezentacja wyników w protokole pokontrolnym.
Wiele odgórnych uregulowań prawnych.	Własne regulacje.
Wdrożony w przypadku ryzyka związanego z danym obszarem badań.	Wdrożona na każdym szczeblu struktury organizacyjnej.
Niezależny.	Ograniczona zakresem upoważnienia.

Źródło: K. Winiarska, *Audyt wewnętrzny*, Difin, Warszawa, 2008, s. 77-78; H. Szymańska, *Ogólne zasady audytu wewnętrznego w jednostkach sektora finansów publicznych*, [w:] *Audyt wewnętrzny w jednostkach sektora finansów publicznych*, T. Kiziukiewicz (red.), Difin, Warszawa, 2009, s.34.

Pierwsza kategoria odnosi się do realizowanych celów. Procedura audytu, mając je na uwadze, monitoruje działanie całej jednostki oraz poszczególnych obszarów, skupiając się na ocenie ich efektywności i przyczynach na podstawie, których powstały. Dodatkowo audytorzy, chcąc uzyskać szerszy i bardziej dokładny obraz jednostki, uwzględniają czynnik ludzki przez przeprowadzanie licznych rozmów z pracownikami oraz kierownictwem. Dzięki temu audyt pełni też funkcję doradczą. Z kolei kontrola wewnętrzna i jej działania głównie koncentrują się na zarządzaniu i nadzorze. Ma określać poziom nieprawidłowości oraz porównywać stan rzeczywisty z obowiązującymi przepisami prawnymi[12].

Kolejną różnicą jest sposób prezentowania wyników przeprowadzanych badań. W przypadku kontroli, końcowe wyniki zawierają ocenę całości działań pracowników jednostki oraz uwzględniają różne czynności zarządcze podejmowane przez kierownictwo. Wszystkie wnioski są dokumentowane w protokole pokontrolnym. Natomiast w audycie zalecenia i wnioski przedstawiane są w sprawozdaniu. Uwzględnia ono wszystkie mechanizmy i procesy zachodzące w jednostce, z naciskiem na ich kompleksową ocenę.

Uwzględniając w rozbieżnościach uregulowania prawne, audyt wewnętrzny w jednostkach sektora publicznego podlega licznym regulacjom dotyczącym jego

[11] B. Nadolna, *System kontroli wewnętrznej w przedsiębiorstwie*, [w:] Studia i Prace Wydziału Nauk Ekonomicznych i Zarządzania 16, 2009, s. 281. (całość 271-282)
[12] A. Heigl, *Controlling, intern revision*, Gustav Fischer Verlag, Stuttgart, New York, 1998, s.6.

przeprowadzenia. Należy tu wspomnieć przede wszystkim o Ustawie o finansach publicznych, Karcie Audytu Wewnętrznego oraz Kodeksie Etyki Audytora Wewnętrznego. Z kolei w przypadku kontroli wewnętrznej, jednostki mają prawo do tworzenia własnych wewnętrznych regulacji, które będą określać sposób jej funkcjonowania[13].

Analizując różnice pomiędzy audytem, a kontrolą zewnętrzną ważne jest zwrócenie uwagi na kryterium częstotliwości. Kontrola jest prowadzona ciągle i powinna być przeprowadzana na każdym szczeblu struktury organizacyjnej jednostki. Za to audyt, mimo, że też odbywa się w sposób ciągły to jednak jego częstotliwość zależy od poziomu ryzyka związanego z obszarem danego badania[14].

Elementem odróżniającym jest także niezależność. Dla audytu niezależność jest kluczowa dla zachowania wiarygodności przeprowadzonego procesu. Kładzie on nacisk na autonomię swoich działań i ocen, podczas gdy kontrola wewnętrzna działa w ramach określonych kompetencji i procedur, określonych przez politykę organizacji.

Oprócz wymienionych różnic, należy wspomnieć także o podobieństwach między pojęciami kontroli wewnętrznej i audytu wewnętrznego. Jednym z ich wspólnych cech jest, to że wykonujący tę pracę są pracownikami jednostki, którą sprawdzają. Obra procesy mają na celu zapewnienie, że organizacja działa zgodnie z określonymi zasadami, przepisami i standardami. Chronią także aktywa organizacji przed stratami wynikającymi z błędów, oszustw lub niewłaściwego zarządzania. Zarówno audyt wewnętrzny, jak i kontrola wewnętrzna oceniają efektywność i skuteczność operacji, procesów oraz systemów zarządzania ryzykiem[15]. Dążą do ciągłej poprawy procesów organizacyjnych, wskazując obszary wymagające usprawnień. Angażują się również w edukację i szkolenie pracowników na temat znaczenia kontroli i audytu, promując świadomość ryzyka i zgodności w organizacji.

5. Audyt wewnętrzny i kontrola wewnętrzna w sektorze publicznym

W sektorze publicznym niezwykle ważne jest prawidłowe funkcjonowanie systemu audytu i kontroli wewnętrznej. Zapewniają one wykorzystanie publicznych środków w sposób odpowiedzialny i zgodny z prawem, a działalność instytucji publicznych jest prowadzona zgodnie z najwyższymi standardami zarządzania. Kierownicy jednostki mają obowiązek tworzenia pisemnych procedur kontroli finansowej dotyczących procesów związanych z gromadzeniem i rozdysponowaniem środków publicznych oraz gospodarowanym mieniem. To właśnie ten obowiązek stał się jedną z przyczyn wprowadzenia wymogu przeprowadzania audytu wewnętrznego w dużej części polskiej administracji publicznej. Obiektywny, profesjonalny i niezależny audyt wewnętrzny stanowi kluczowe źródło informacji dla kierownika jednostki sektora finansów publicznych na temat funkcjonowania jego jednostki. Dane dostarczane przez audytora wewnętrznego umożliwiają kierownikowi ocenę, czy wdrożony w jednostce system kontroli wewnętrznej realizuje swoje zadania. Oprócz ustalonych zasad dotyczących przeprowadzania audytu, istotne jest także skuteczne zorganizowanie pracy

---

[13] H. Szymańska, *Ogólne zasady audytu wewnętrznego w jednostkach sektora finansów publicznych*, [w:] *Audyt wewnętrzny w jednostkach sektora finansów publicznych*, T. Kiziukiewicz (red.), Difin, Warszawa, 2009, s.34.

[14] M. Chmielewska, *Zestawienie różnic pomiędzy audytem wewnętrznym, a kontrolą wewnętrzną w sektorze publicznym*, [w:] *Finanse, rachunkowość, kontrola i audyt w sektorze publicznym i prywatnym. Studium przypadków*, T. Gabrusewicz, K. Marchewka-Bartkowiak, M. Wiśniewski (red.), CeDeWu, Warszawa 2015, s. 99. (całość 87-101)

[15] O. Martyniuk, *Audyt wewnętrzny, a kontrola wewnętrzna*, [w:] *Prace i Materiały Wydziału Zarządzania Uniwersytetu Gdańskiego*, 2005, nr 2, s. 137-139. (s.131-142)

komórki audytu wewnętrznego. Dzięki temu audytor może dostarczać kierownikowi jednostki rzetelne informacje, które są niezbędne do zarządzania ryzykiem działalności, zgodnie z obowiązującymi przepisami prawa oraz Standardami Audytu Wewnętrznego[16].

Audyt wewnętrzny w sektorze publicznym jest oparty na kilku fundamentalnych zasadach. Po pierwsze audyt wewnętrzny jest obowiązkowy dla jednostek sektora finansów publicznych. Po drugie, zgodnie z zasadami, audyt jest prowadzony przez audytorów wewnętrznych zatrudnionych w tych jednostkach, z pewnymi wyjątkami, ale one dotyczą jednostek podległych i nadzorowanych. Osoby te powinny posiadać odpowiednie uprawnienia. Ważne jest również, aby system audytu podlegał koordynacji, a zadania centralnej jednostki harmonizującej są realizowane przez Ministra Finansów oraz Generalnego Inspektora Kontroli Skarbowej w przypadku środków pochodzących z Unii Europejskiej. Te zasady zapewniają odpowiednie ramy działania audytu[17].

W finansach publicznych kontrola wewnętrzna została określona mianem zarządczej, aby nie kojarzyć jej jedynie z weryfikacja zgodności z procedurami, lecz uwzględnić aspekty systemowe i podkreślać odpowiedzialność kierowników jednostek sektora za jej skuteczność. Przejęcie odpowiedzialności za projektowanie i wdrażanie systemów kontroli, ustanowiło obowiązek ich przeprowadzania. Kierownicy są również odpowiedzialni za systematyczne monitorowanie systemów, aby zapewnić ich realizację w zakresie niezbędnym do osiągania celów jednostek. Wprowadzanie ich może początkowo wiązać się ze zwiększonym nakładem pracy kosztów, ale dzięki uporządkowaniu procedur kontrolnych, przewiduje się, że w sektorze publicznym zostaną zidentyfikowane i usunięte powtarzające się czynności oraz zbędne punkty decyzyjne[18].

6. Podsumowanie

Podsumowując, audyt wewnętrzny i kontrola wewnętrzna odgrywają niezwykle istotną rolę w sektorze publicznym, zapewniając skuteczne zarządzanie i gospodarowanie środkami publicznymi. Działalność sektora finansów publicznych wymaga przestrzegania wielu procedur i szczegółowych wytycznych, co jest niezwykle istotne ze względu na zarządzanie środkami publicznymi. Największy nacisk kładziony jest na aspekt nadzoru. Kluczowym elementem skutecznego zarządzania jest implementacja odpowiednich narzędzi wspierających system zarządzania. Wdrożenie klarownych procedur audytowych i kontrolnych nie tylko minimalizuje ryzyko nadużyć i błędów, ale również umożliwia lepsze wykorzystanie dostępnych zasobów oraz realizację strategicznych celów organizacji. Ich idealnym odpowiednikiem są audyt wewnętrzny i kontrola wewnętrzna.

---

[16] A. Mazurek-Różynek, *Koordynacja audytu wewnętrznego z kontrolą wewnętrzną i audytem zewnętrznym w jednostkach sektora finansów publicznych*, [w:] *Studia i prace wydziału nauk ekonomicznych i zarządzania* nr 16, 2009, s.11-12. (11-19)

[17] M. Wiśniewska, *Audyt wewnętrzny w sektorze publicznym w Polsce-geneza, funkcjonowanie, koncepcje rozwoju*, „Samorząd terytorialny w Polsce i w Europie. Aktualne problemy i wyzwania", K. czarnecki, A. Lutrzykowski, R. Musiałkiewicz, Włocławek 2017, s. 403.

[18] E.W. Babuśka, *Koncepcja kontroli zarządczej w sektorze finansów publicznych*, [w:] *Zeszyty Naukowe Uniwersytetu Szczecińskiego. Finanse, rynki finansowe, ubezpieczenia nr 42*, Uniwersytet Ekonomiczny w Krakowie, 2011, s. 19-20. (11-22)

Audyt i kontrola często są mylone i podawane, jako synonimy co może prowadzić do nieporozumień. Ważne jest podkreślenie, że audyt i kontrola są zupełnie różnymi procesami, które są złożone i wieloaspektowe, składające się z licznych elementów. Należy je rozróżniać i analizować osobno, zamiast traktować jako jedno.

W kontekście dynamicznie zmieniającego się otoczenia regulacyjnego, ważne jest nieustanne doskonalenie i adaptacja tych instrumentów do nowych wyzwań i potrzeb społeczeństwa. Ostatecznie, audyt wewnętrzny i kontrola wewnętrzna stają się fundamentem stabilności i zaufania w sektorze publicznym, wspierając cele służące dobru wspólnemu oraz efektywne zarządzanie zasobami publicznymi, stają się do tego najlepszymi narzędziami.

Literatura

1. Ampuła D., Kontrola i audyt wewnętrzny w jednostce organizacyjnej, [w:] Zeszyty 132 nr 4, Wojskowy Instytut Techniczny Uzbrojenia, 2014, s. 17-27.
2. Babuśka E.W., Koncepcja kontroli zarządczej w sektorze finansów publicznych, [w:] Zeszyty Naukowe Uniwersytetu Szczecińskiego. Finanse, rynki finansowe, ubezpieczenia nr 42, Uniwersytet Ekonomiczny w Krakowie, 2011, s. 11-22.
3. Bednarek P., Wybrane determinanty tworzenia wartości dodanej przez audyt wewnętrzny w jednostkach sektora finansów publicznych - istniejący dorobek i kierunki dalszych badań, [w:] Zeszyty Naukowe Uniwersytetu Szczecińskiego nr 864 Finanse, Rynki Finansowe, Ubezpieczenia nr 76, t.2, 2015, s. 23-34.
4. Bielińska-Dusza E., Zastosowanie audytu wewnętrznego w analizie wizji i misji przedsiębiorstwa, [w:] Zeszyty Naukowe Uniwersytetu Ekonomicznego w Krakowie, 2010, nr 829, s. 19-36.
5. Chmielewska M., Zestawienie różnic pomiędzy audytem wewnętrznym, a kontrolą wewnętrzną w sekto-rze publicznym, [w:] Finanse, rachunkowość, kontrola i audyt w sektorze publicznym i prywatnym. Studium przypadków, T. Gabrusewicz, K. Marchewka-Bartkowiak, M. Wiśniewski (red.), CeDeWu, Warszawa 2015, s. 87-101.
6. Heigl A., Controlling, intern revision, Gustav Fischer Verlag, Stuttgart, New York, 1998.
7. Kałużny S., Kontrola wewnętrzna. Teoria i praktyka, Polskie Wydawnictwo Ekonomiczne, Warszawa, 2008.
8. Kuca B.R., *Kontrola, kontroling i audyt w zarządzaniu*, Wyższa Szkoła Zarządzania i Prawa, Warszawa 2006.
9. Martyniuk O., Audyt wewnętrzny, a kontrola wewnętrzna, [w:] Prace i Materiały Wydziału Zarządzania Uniwersytetu Gdańskiego, 2005, nr 2, s. 131-142.
10. Mazurek-Różynek A., Koordynacja audytu wewnętrznego z kontrolą wewnętrzną i audytem zewnętrznym w jednostkach sektora finansów publicznych, [w:] Studia i prace wydziału nauk ekonomicznych i zarządzania nr 16, 2009, s.11-19.
11. Moeller R., Nowoczesny audyt wewnętrzny, Wydawnictwo Nieoczywiste, Warszawa 2018.
12. Nadolna B., System kontroli wewnętrznej w przedsiębiorstwie, [w:] Studia i Prace Wydziału Nauk Ekonomicznych i Zarządzania 16, 2009, s. 271-282.
13. Skoczylas A., Audyt wewnętrzny jako integralny element właściwego funkcjonowania publicznej wewnętrznej kontroli finansowej, w: Controlling i audyt w usprawnianiu zarządzania, red. K. Winiarska, US w Szczecinie, Szczecin 2005.

14. Standards of the Professional Practice of Internal Auditing, The Institute of Internal Auditors, Altamonte Springs, Florida 2001.
15. Szymańska H., Ogólne zasady audytu wewnętrznego w jednostkach sektora finansów publicznych, [w:] Audyt wewnętrzny w jednostkach sektora finansów publicznych, T. Kiziukiewicz (red.), Difin, Warszawa, 2009.
16. Tubek B., Kontrola wewnętrzna jako instrument zarządzania w sektorze finansów publicznych - ważne choć wciąż niedocenianie narzędzie doskonalenia procesów zarządczych, [w:] Nauki Ekonomiczne Tom 35, Krakowska Akademia im. Andrzeja Frycza Modrzewskiego, 2022, s. 169-188.
17. Wiśniewska M., Audyt wewnętrzny w sektorze publicznym w Polsce-geneza, funkcjonowanie, koncepcje rozwoju, „Samorząd terytorialny w Polsce i w Europie. Aktualne problemy i wyzwania", K. Czarnecki, A. Lutrzykowski, R. Musiałkiewicz, Włocławek 2017.

Kinga Dębska
Studenckie Koło Naukowe Rachunkowości „ASSETS”

dr Agnieszka Lew
Opiekun Koła Naukowego

Amortyzacja jako narzędzie modelowania kosztami w przedsiębiorstwie

Streszczenie

Artykuł prezentuje różne rozwiązania ewidencyjne w zakresie amortyzacji, które przedsiębiorstwo może wybrać zarówno na poziomie bilansowym, jak i podatkowym. Celem artykułu jest podkreślenie w jaki sposób przyjęta metoda amortyzacji ma wpływ na stronę kosztową przedsiębiorstwa, oddziałując jednocześnie na wynik finansowy, zarówno w krótkiej, jak i długiej perspektywie. Analizie poddano metodę liniową, degresywną, naturalną oraz odpis jednorazowy, jako metodę badawczą wykorzystując ścisły przegląd literatury. Wnioskuje się, że jeśli przedsiębiorstwo chce aktywnie korzystać z możliwości modelowania kosztów, jakie daje amortyzacja, powinno oddzielić amortyzację bilansową od podatkowej.

Słowa kluczowe: koszty, metody amortyzacji, amortyzacja bilansowa i podatkowa.

1. Wprowadzenie

Środki trwałe są ważnym elementem aktywów większości przedsiębiorstw działających na dużą skalę. Są to obiekty będące własnością jednostki gospodarczej – po zwykle jednorazowym wydaniu środków pieniężnych podczas zakupu nie ponosi się dodatkowych cyklicznych opłat (jak ma to miejsce choćby w przypadku leasingu). Choć środki trwałe z reguły nie wymagają ponoszenia dodatkowych wydatków, nie są wolne od kosztów. Ze względu na okres użyteczności środków trwałych, podlegają one zużywaniu się, co odzwierciedla amortyzacja.

Ustawa o rachunkowości w wielu jej obszarach oferuje użytkownikom szeroki wachlarz możliwości wyboru stosowanych rozwiązań. Dobra znajomość jednostki, której rachunkowość się prowadzi, w połączeniu z szeroką wiedzą na temat wariantów oferowanych przez UoR oraz inne ustawy (przede wszystkim podatkowe), pozwala zarządzać zapisami w księgach rachunkowych w granicach prawa, maksymalizując ich użyteczność dla celów spółki. Zasada ta ma również zastosowanie w przypadku amortyzacji, która może być rejestrowana nie tylko przy użyciu różnych stawek amortyzacji, jak również różnych metod. Kompilacja tych decyzji wpływa na wartość rejestrowanych kosztów zarówno w krótkim, jak i w długim okresie. Dodatkowym aspektem, który wpływa na sposób ujmowania amortyzacji, jest jej wymiar podatkowy, w zakresie którego obowiązują dodatkowe reguły, nieco odmienne od tych

bilansowych. Analiza dostępnych rozwiązań oraz zrozumienie interesów przedsiębiorstwa pozwala modelować częściowo wymiar rejestrowanych kosztów, za pomocą amortyzacji.

Celem artykułu jest analiza ujmowania kosztów amortyzacji w rachunkowości, pod kątem dostępnych metod i rozwiązań proponowanych przez przepisy bilansowe oraz podatkowe. Rozważaniom poddano również problem, jakie konsekwencje przynosi wybór konkretnej metody w kontekście wymiaru rejestrowanych w ten sposób kosztów oraz manipulacji wynikiem finansowym. Wykorzystaną metodą badawczą jest ścisły przegląd literatury.

2. Istota amortyzacji środków trwałych

Środki trwałe można określić jako znaczący element majątku przedsiębiorstwa, ponieważ ze względu na mnogość ich rodzajów, mogą służyć różnym celom w jednostce. Ich wspólną cechą, determinującą klasyfikację jako środek trwały, jest zakładany okres użytkowania, wynoszący powyżej jednego roku. Dodatkowymi założeniami, które musi spełnić dany składnik aktywów aby zostać uznanym za środek trwały, jest jego kompletność, zdatność do użytku oraz przeznaczenie do użytkowania na potrzeby jednostki[1]. Użytkowanie środków trwałych wiąże się ze zużywaniem[2], które często, ale nie zawsze prowadzi do niezdatności do użytkowania danego środka.

W rachunkowości odzwierciedleniem tego zużywania się jest amortyzacja. Rejestrowanie jej jest zasadne, ponieważ środki trwałe wraz z upływem czasu zmieniają swoje cechy, obniżają efektywność i jakość pracy, co wiąże się z redukcją ich wartości użytkowej. Wymiar zużywania się może być zróżnicowany, w zależności od stopnia eksploatacji danego środka trwałego, czy też jego indywidualnego stanu (poziomu technicznego, jakości wykonania)[3].

W celu ustalenia wartości odpisów amortyzacyjnych stosuje się roczne stawki amortyzacji – które wyrażane są poprzez okres ekonomicznej użyteczności środka trwałego. Czasem trudne może być jego określenie, ponieważ zależy od wielu czynników: liczby zmian, na których pracuje dany środek trwały, tempa postępu techniczno-ekonomicznego czy też wydajności środka trwałego[4]. Okres żywotności danego środka trwałego powinien możliwie najbardziej wiarygodnie odzwierciedlać rzeczywistość. Należy przy tym pamiętać, że proces amortyzacji (ujmowanej w korespondencji z kontem umorzeniowym) po upływie okresu ekonomicznej

[1] G. Borowska, I. Frymark, *Księgowość i kalkulacja. Część 1*, Wydawnictwo WSiP, Warszawa 2013, s.367.
[2] M. Bąk, *Środki trwałe i ich zużycie w cyklu życia produktu*, „Prace naukowe Uniwersytetu Ekonomicznego we Wrocławiu", 2012, nr. 255, s. 13.
[3] Ibidem, s. 14.
[4] M. Bąkowski, A. Wasilewska, *Zasady amortyzacji środków trwałych w księgach rachunkowych osób prawnych (część 2)*, „KNUV", 2014, nr. 3(41), s. 92.

użyteczności sprawi, że w księgach rachunkowych dany środek trwały będzie wykazywał wartość zerową. Nie oznacza to jednak, że przedsiębiorstwo ma się go pozbywać – choć rachunkowo nie będzie mieć wartości, nadal może służyć jednostce, jeśli w praktyce jest do tego zdatny. Nawet jeśli jego wydajność znacząco spadła, wciąż może wspomagać przynoszenie jednostce korzyści ekonomicznych, przy jednoczesnym braku ponoszeniu kosztów, co z punktu widzenia minimalizacji kosztów jest sytuacją niezwykle pożądaną.

Wysokość naliczanej miesięcznie amortyzacji zależy nie tylko od zaimplementowanej stawki oraz wartości początkowej środka trwałego, ale też od zastosowanej metody. W prawie bilansowym i podatkowym najczęściej można spotkać się z metodami: liniową, degresywną i naturalną[5].

3. Amortyzacja liniowa

Metoda ta jest najprostszą w swojej istocie i bywa nazywana również metodą amortyzacji równomiernej[6]. Jej podstawowym założeniem jest dokonywanie odpisów w równej wysokości, przez cały okres użyteczności ekonomicznej środka trwałego. Oznacza to, że wartość początkową danego obiektu mnoży się przez stawkę amortyzacji, uzyskując wartość rocznego odpisu – takiego samego w każdym roku.

Dobór stawki amortyzacyjnej jest zatem w tym przypadku jedynym czynnikiem mającym wpływ na wartość odpisu w czasie. W każdym miesiącu wymiar kosztów amortyzacji w kontekście danego środka trwałego będzie taki sam – może być jedynie wyższy, doprowadzając do umorzenia w krótszym okresie, lub niższy, amortyzując obiekt w dłuższej perspektywie czasu. Należy przy tym pamiętać, że dobór stawki nie jest zupełnie dowolny – musi mieścić się w zasadnym przedziale zgodnym z Wykazem Stawek Amortyzacyjnych oraz ujęty w polityce rachunkowości przedsiębiorstwa.

Metoda liniowa nie daje wielu możliwości w zakresie rozkładania wartości odpisów w czasie – jedynym czynnikiem mogącym na to wpływać jest przyjęta stawka amortyzacyjna. Po jej ustaleniu, odpisy pozostają niezmienne w czasie – jeśli nie nastąpi ulepszenie środka trwałego zwiększające jego wartość początkową. W takim przypadku, musi nastąpić rekalkulacja miesięcznego odpisu.

---

[5] J. Godlewska, T. Fołta, *Zaawansowana rachunkowość finansowa z uwzględnieniem sprawozdawczości finansowej i prawa podatkowego*, Stowarzyszenie Księgowych w Polsce, Warszawa 2022, s. 192.

[6] M. Stypa, *Amortyzacja środków trwałych w ujęciu podatkowym i bilansowym*, [w:] *Ekonomia człowieka. Wymiary i aspekty*, red. J. Zimny, Katolicki Uniwersytet Jana Pawła II w Lublinie, Stalowa Wola 2017, s. 97.

4. Amortyzacja degresywna

Nazwa metody degresywnej już poprzez swoje nazewnictwo wskazuje na istotę swojego założenia – a więc odpisy nie tylko zmieniające swoją wartość w czasie, ale również ulegające zmniejszeniom. Nazywana jest również metodą amortyzacji przyspieszonej, ponieważ całkowity okres umorzenia danego środka trwałego przy zastosowaniu metody degresywnej jest krótszy niż w przypadku metody liniowej[7]. Oba warianty wiążą się z zastosowaniem corocznie tej samej stawki amortyzacji – ponieważ zależy ona od natury środka trwałego, a nie przyjętej metody. Różnice w wysokości odpisów wynikają zatem z odmiennej podstawy obliczeniowej. W przypadku metody liniowej wartość odpisu odnosi się zawsze do wartości początkowej, zaś metoda degresywna wartość odpisu ustala na podstawie bieżącej wartości środka trwałego[8], a więc wartości początkowej pomniejszonej o dotychczasowe umorzenie. Stąd wynika degresywny charakter tej metody – wraz z upływem czasu skumulowane umorzenie wzrasta, a wartość netto będąca podstawą obliczeniową zmniejsza się.

Amortyzacja metodą degresywną składa się z dwóch etapów[9]:

- w etapie pierwszym naliczanie amortyzacji odbywa się poprzez zastosowanie stawki wskazanej w Wykazie, podwyższonej o współczynnik nie większy niż 2,0, od wartości bieżącej środka trwałego,
- etap drugi rozpoczyna się w momencie, w którym odpis dokonywany degresywnie byłby mniejszy niż odpis ustalony przy pomocy metody liniowej – w tej sytuacji następuje przejście na metodę liniową przy stosowaniu tej samej stawki, jednak bez współczynnika podwyższającego. Etap drugi trwa do momentu całkowitego umorzenia środka trwałego.

Zastosowanie metody degresywnej amortyzacji powoduje, że w początkowych okresach eksploatacji środków trwałych rejestruje się zwiększone odpisy amortyzacyjne, niż w końcowych. W związku z tym, koszty amortyzacji w większej skali obciążają wynik finansowy na początku używania danego składnika aktywów – na końcu zaś odnotowuje się zwiększenie wyniku[10]. Wariant ten stosuje się najczęściej do takich środków trwałych, które na początku swojego cyklu życia ulegają szybszemu zużywaniu się – zwiększone odpisy mają w ten sposób

[7] J. Iwin-Garzyńska, *Kapitał amortyzacyjny w zarządzaniu finansami*, PWE, Warszawa 2005, s. 99.
[8] J. Duraj, *Podstawy ekonomiki przedsiębiorstwa*, PWE, Warszawa 2000, s. 482.
[9] D. Bem, *Amortyzacja podatkowa środków trwałych oraz wartości niematerialnych i prawnych jako źródło finansowania wewnętrznego, metody amortyzacji, problemy*, „Studia i prace kolegium zarządzania i finansów", 2007, nr. 82, s. 52.
[10] M. Grabowska, *Rola amortyzacji w kształtowaniu majątkowej strategii gospodarowania kapitałem zasobowym przedsiębiorstwa*, „Acta Universitatis Lodziensis. Folia Oeconomica", 2010, nr. 236, s. 126.

poniekąd zbilansować późniejsze potencjalne koszty remontów i napraw. Opisany mechanizm stosowania metody degresywnej ma zastosowanie w przypadku amortyzacji podatkowej, ponieważ ustawa o rachunkowości nie przewiduje zmiany metody amortyzacji w ciągu roku[11]. Zmienny charakter odpisów oraz jego korelacja z wynikiem jednostki może skłaniać przedsiębiorstwa do wykorzystywania metody degresywnej w zakresie manipulacji wynikiem finansowym. Zwiększony wymiar kosztów w początkowej jej fazie zmniejsza wynik finansowy, zmniejszając tym samym zobowiązanie podatkowe – co z punktu widzenia jednostki może być pożądane w latach bardziej zyskownych.

Warto podkreślić, że opisana procedura jest jedną z odmian metody degresywnej, adresowaną przez przepisy podatkowe (nazywana również metodą degresywno-liniową). W literaturze znaleźć można również wariant polegający na metodzie malejącego salda, czy też metodzie sumy cyfr rocznych[12].

5. Amortyzacja naturalna

Naturalna metoda amortyzacji polega na ustalaniu wysokości odpisów amortyzacyjnych zgodnie z rzeczywistym zużyciem danego środka trwałego, poprzez określenie zaistniałego czasu eksploatacji i odpowiednie rozdzielenie wartości początkowej[13]. Może też mieć postać ustalania odpisów na podstawie wielkości produkcji[14].

Aby zastosowanie metody naturalnej było możliwe, musi istnieć sposób ustalenia łącznego potencjału użytkowego środka trwałego oraz jednostki pomiarowej. Jednocześnie, musi być możliwy wiarygodny pomiar tego potencjału w wybranej jednostce czasu za pomocą ustalonej jednostki miary. Zaletą tej metody jest jej dokładność – w najbardziej rzetelny sposób odzwierciedla rzeczywiste zużywanie się środków trwałych. Co więcej, w przypadku czasowego wyłączenia z użycia danego składnika aktywów (np. w przypadku sezonowości produkcji), zostaje to ujęte bezpośrednio[15], a amortyzacja nie ulega przeszacowaniu.

Metoda naturalna może podlegać największym wahaniom w wysokości odpisów spośród omówionych w artykule. Wahania te zwykle nie są jednak zwykle kontrolowane przez człowieka, w związku z czym wymiar kosztów amortyzacji w danym okresie jest najmniej przewidywalny.

---

[11] J. Godlewska, T. Fołta, *Zaawansowana rachunkowość…*, op. cit. s. 192.
[12] Ibidem.
[13] P. Jasiorska, A. Krawczyk, P. Owczarek, *Specyfika ustalania amortyzacji na przykładzie branży wydobywczej*, [w:] *Rachunkowość – ludzie, pasja,* historie, red. E. Śnieżek, Wydawnictwo SIZ, Łódź 2020, s. 64.
[14] M. Grabowska, *Rola amortyzacji…*, op. cit. s. 126.
[15] J. Godlewska, T. Fołta, *Zaawansowana rachunkowość…*, op. cit. s. 192-193.

6. Odpis jednorazowy

Niektóre środki trwałe mają stosunkowo niską wartość początkową – z praktycznego punktu widzenia amortyzacja takich aktywów przez okres ich ekonomicznej użyteczności wydaje się mało zasadna. Przepisy podatkowe przewidują zatem możliwość zastosowania jednorazowego odpisu amortyzacji.

Istotą metody jednorazowej jest jednorazowe uznanie za koszty uzyskania przychodu całkowitej wartości początkowej środka trwałego. W świetle przepisów podatkowych należy spełnić dwa warunki, aby odpis jednorazowy był możliwy: wartość środka trwałego nie może przekraczać 3500 zł (od 2018 r. próg ten podwyższono do 10000 zł[16]) a wydatek poniesiony w związku z jego nabyciem musi zostać poniesiony w miesiącu oddania go do użytkowania[17].

Z punktu widzenia kosztów w przedsiębiorstwie, metoda odpisu jednorazowego daje jednostkom relatywnie największe pole do manipulacji wypracowanym zyskiem. Odpis jednorazowy uznawany jest podatkowo, zatem kiedy prognozuje się wypracowanie wysokiej podstawy opodatkowania, zakup środka trwałego o niskiej wartości i jednorazowe odpisanie go w koszty daje możliwość zmniejszenia zobowiązania podatkowego, bez dodatkowych konsekwencji kosztowych w przyszłości. Jest to rozwiązanie najbardziej doraźne, ponieważ nie wiąże się z dodatkowymi odpisami w przyszłych okresach – przedsiębiorstwo nie musi liczyć się z występowaniem kosztów stałych w momencie, w którym nie wie, jak będzie kształtował się jego wynik (w odróżnieniu od metody liniowej czy degresywnej).

7. Podsumowanie

Amortyzacja jest kosztem, który daje możliwość szerokiego ujęcia, na wiele różnych sposobów. Metody te są regulowane zarówno od strony bilansowej, poprzez ustawę o rachunkowości, jak również przez przepisy podatkowe. Oba akty prawne charakteryzuje występowanie pewnych różnic i rozbieżności – jednym z nich jest możliwość dokonywania odpisów już od miesiąca przyjęcia środka trwałego do użytkowania albo miesiąca następnego (ujęcie bilansowe), lub od miesiąca następnego po przyjęciu do użytkowania (ujęcie podatkowe). Podobnie stawki amortyzacji (a więc i okres umarzania) może się różnić w przepisach bilansowych i podatkowych. W przypadku zdecydowania się na odrębne ujmowanie amortyzacji bilansowej i podatkowej, należy odpowiednio ująć to w księgach rachunkowych –

[16] J. Iwin-Garzyńska, *Podatkowy kapitał amortyzacyjny – zarys problemu*, „Studia Ekonomiczne. Zeszyty Naukowe Uniwersytetu Ekonomicznego w Katowicach", 2018, nr. 358, s. 96.
[17] J. Iwin-Garzyńska, *Amortyzacja podatkowa jako instrument wspierania inwestycji małych przedsiębiorstw w Polsce*, „Zeszyty Naukowe Uniwersytetu Szczecińskiego. Ekonomiczne problemy usług nr. 116", 2015, nr. 848, s. 626.

najczęściej poprzez prowadzenie odpowiednio rozbudowanej analityki do kont syntetycznych. W praktyce jednak, celem uniknięcia komplikacji i uproszczenia rozliczania amortyzacji dąży się do znalezienia rozwiązania wspólnego dla ustawy o rachunkowości i przepisów podatkowych.

Przedsiębiorstwo decydując się na wykorzystanie korzyści metod podatkowych, które dostrzega, godzi się na odrębne naliczanie amortyzacji bilansowej i podatkowej. Może być to jednak zabieg pożądany i opłacalny, jeśli jednostce zależy na zwiększenie kosztów w początkowej fazie eksploatacji środka trwałego, może zdecydować się na amortyzację degresywną, zamiast liniowej. Korzystnym dla przedsiębiorstw rozwiązaniem podatkowym jest też możliwość jednorazowego odpisania w koszty uzyskania przychodu środków trwałych o niskiej wartości początkowej. Metoda ta pozwala w najbardziej doraźny sposób zmniejszyć wymiar zobowiązania podatkowego odprowadzanego do urzędu. Główną sprzecznością między obszarem bilansowym i podatkowym jest interes – z punktu widzenia księgowego jednostki chcą maksymalizować zysk netto. Z punktu widzenia podatkowego, większy zysk wiąże się z większym zobowiązaniem podatkowym. Dlatego też warto rozważyć odrębne ujmowanie amortyzacji bilansowej i podatkowej – rozwiązania proponowane przez oba akty prawne umożliwiają modelowanie wysokości tego rodzaju kosztów w czasie celem zaspokojenia obu interesów. Należy jednak pamiętać, że nadrzędnym celem wyboru metody i stawki amortyzacji danego środka trwałego, powinno być oddanie jego rzeczywistego zużycia, a nie manipulacja wynikiem jednostki.

Literatura

1. Bąk M., *Środki trwałe i ich zużycie w cyklu życia produktu*, „Prace naukowe Uniwersytetu Ekonomicznego we Wrocławiu", 2012, nr. 255.
2. Bąkowski M., Wasilewska A., *Zasady amortyzacji środków trwałych w księgach rachunkowych osób prawnych (część 2)*, „KNUV", 2014, nr. 3(41).
3. Bem D., *Amortyzacja podatkowa środków trwałych oraz wartości niematerialnych i prawnych jako źródło finansowania wewnętrznego, metody amortyzacji, problemy*, „Studia i prace kolegium zarządzania i finansów", 2007, nr. 82.
4. Borowska G., Frymark I., *Księgowość i kalkulacja. Część 1*, Wydawnictwo WSiP, Warszawa 2013.
5. Duraj J., *Podstawy ekonomiki przedsiębiorstwa*, PWE, Warszawa 2000.

6. Godlewska J., Fołta T., *Zaawansowana rachunkowość finansowa z uwzględnieniem sprawozdawczości finansowej i prawa podatkowego*, Stowarzyszenie Księgowych w Polsce, Warszawa 2022.
7. Grabowska M., *Rola amortyzacji w kształtowaniu majątkowej strategii gospodarowania kapitałem zasobowym przedsiębiorstwa,* „Acta Universitatis Lodziensis. Folia Oeconomica", 2010, nr. 236.
8. Iwin-Garzyńska J., *Amortyzacja podatkowa jako instrument wspierania inwestycji małych przedsiębiorstw w Polsce*, „Zeszyty Naukowe Uniwersytetu Szczecińskiego. Ekonomiczne problemy usług nr. 116", 2015, nr. 848, s. 626.
9. Iwin-Garzyńska J., *Kapitał amortyzacyjny w zarządzaniu finansami,* PWE, Warszawa 2005.
10. Iwin-Garzyńska J., *Podatkowy kapitał amortyzacyjny – zarys problemu*, „Studia Ekonomiczne. Zeszyty Naukowe Uniwersytetu Ekonomicznego w Katowicach", 2018, nr. 358.
11. Jasiorska P., Krawczyk A., Owczarek P., *Specyfika ustalania amortyzacji na przykładzie branży wydobywczej*, [w:] Rachunkowość – ludzie, pasja, historie, red. E. Śnieżek, Wydawnictwo SIZ, Łódź 2020.
12. Stypa M., *Amortyzacja środków trwałych w ujęciu podatkowym i bilansowym*, [w:] *Ekonomia człowieka. Wymiary i aspekty*, red. J. Zimny, Katolicki Uniwersytet Jana Pawła II w Lublinie, Stalowa Wola 2017.

KOŁO

NAUKOWE

INTERAKCJI

CZŁOWIEK-KOMPUTER

„GEST”

Wiktor Kuczek
Koło Naukowe Interakcji Człowiek – Komputer „GEST”

mgr. inż. Dawid Kalandyk
Opiekun naukowy

Analiza działania sieci CNN oraz propozycja modelu rozpoznającego kolekcjonerskie karty do gry

Streszczenie

Artykuł opisuje zasady działania konwolucyjnej sieci neuronowej CNN, oraz jej poszczególnych warstw, a także konkretne operacje i obliczenia przez nie realizowane. Autor w szczególności skupia się na opisaniu procesu przygotowywania odpowiedniej bazy danych, która jest przystępna do wykorzystania w procesie uczenia rozpoznawania kart. W pracy przedstawiono również budowę proponowanego modelu sieci w języku Python. Artykuł analizuje również szereg aspektów procesu nauki proponowanej sieci, jak i działania poszczególnych warstw. Na końcu autor przedstawia wyniki analizy w postaci wykresów oraz tabel w celu zobrazowaniu procesu i wyników nauki sieci neuronowej. Opisany model jest częścią większego projektu koła naukowego GEST o nazwie Rzeszow University of Technology – Artifical Intelligence Applications (RUT-AI Applications).

Słowa kluczowe: konwolucja, CNN, sieć neuronowa, baza danych, rozpoznawanie kart.

1. Wprowadzenie

Zbieranie kart kolekcjonerskich to popularne i rozwijające hobby, które również jest przedmiotem handlu na całym globie. Ich klasyfikacja, ocena, przechowywanie danych o nich oraz aktualizowanie stanu swojej kolekcji jest kluczowe dla sprzedawców, aukcjonerów oraz zwykłych entuzjastów. W standardowych warunkach zidentyfikowanie karty, za czym to idzie jej wartości, wymaga eksperckiej wiedzy oraz poświęcenia dodatkowego czasu. Dzięki postępom w dziedzinie sztucznej inteligencji istnieje możliwość zautomatyzowania tego procesu przy użyciu konwolucyjnych sieci neuronowych (ang. Convolutional Neural Network - CNN).

Celem jest opracowanie systemu opartego na konwolucyjnych sieciach neuronowych CNN, który jest w stanie automatycznie klasyfikować karty kolekcjonerskie na podstawie zdjęć zrobionych przez użytkownika lub obrazów z internetu. System ma na celu zwiększenie prędkości i efektywności w identyfikowaniu kart w porównaniu do ludzi oraz aby uniknąć błędów przez nich popełnianych.

Do osiągnięcia celu należy przygotować bazę danych zawierających odpowiednią liczbę zdjęć kart. Taki zestaw musi być odpowiednio podzielony na klasy, które w późniejszym etapie będą podstawą do procesu rozpoznawania kart. Do procesu uczenia na tej bazie wykorzystany został język Python wraz z jego biblioteką TensorFlow.

2. Przygotowanie bazy z danymi

2.1 Opis zbioru

Zbiór danych stworzony został przez autora z własnoręcznie zrobionych zdjęć kart kolekcjonerskich z gry karcianej **One Piece**. Składa się on z katalogu ***"train"*** przeznaczonego na obrazy, które w późniejszych etapach zostaną użyte do uczenia modelu oraz katalogu ***"test"*** z danymi do testu, czyli sprawdzania w jakim stopniu model nauczył się rozpoznawać określone karty. W zbiorze testowym nie mogą, a przynajmniej dla rzetelnej oceny jakości nauki sieci nie powinny znajdować się obrazy, które również zawarte są w katalogu *"train"*. Ostateczny zbiór użyty do uczenia sieci neuronowej składa się z **21** katalogów, każdy z nich reprezentuje jedną klasę, czyli jedną unikatową kartę, której rozpoznawania model będzie musiał się nauczyć. We wszystkich katalogach znajduje się łącznie **1260** obrazów, które są podzielone według proporcji **80:20**. Oznacza to, że w katalogu *"train"* łącznie znajduje się **1008** obrazów z czego każda klasa zawiera ich **48**, natomiast w katalogu "test" łączna ilość zdjęć wynosi **252**, czyli po **12** opisujących każdą klasę. Zastosowanie tych proporcji pozwala na wydajniejsze i efektywniejsze uczenie się modelu, w konsekwencji czego użytkownik uzyskuje zadowalające oraz użyteczne rezultaty.

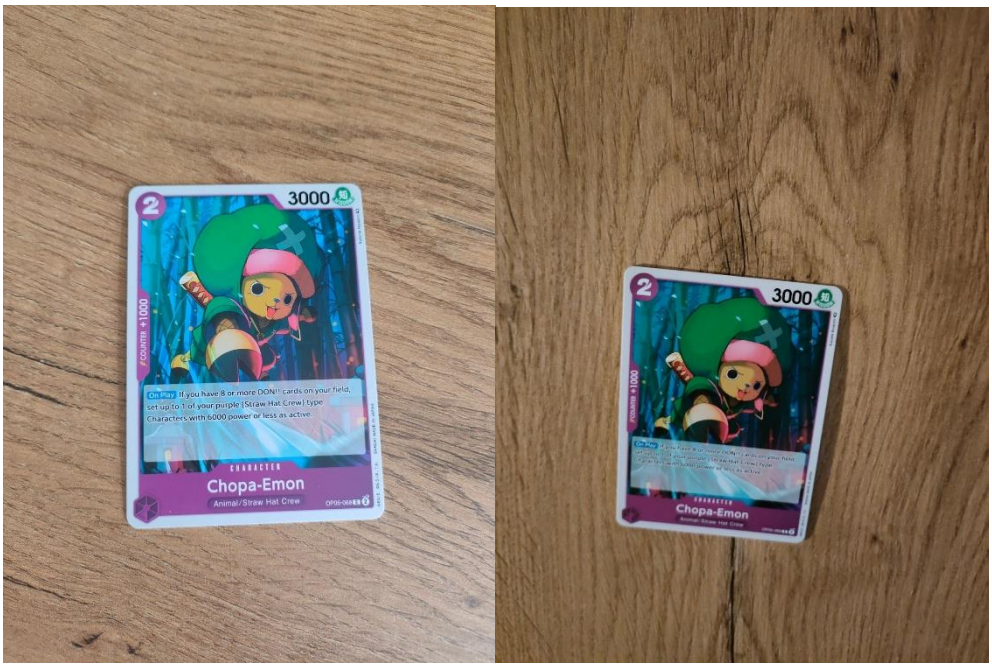
Rysunek 1. Wszystkie karty
Źródło: opracowanie własne.

2.2 Dobór danych

Każda z kart [Rysunek 1] reprezentuje jedną z klas. Dla utrudnienia nauki modelu zadbano o wybranie odpowiednich kart. Część z nich posiada wspólne cechy. na przykład białe ramki lub czerwony prostokąt w miejscu nazwy karty. Niektóre z nich wykonane są w stylu bardziej “komiksowym” inne zaś bardziej “trójwymiarowo”. Innymi różnice, które można zaobserwować to kolory ramek (oprócz w większości białych są rownież żółte lub niebieskie), dodatkowy półprzezroczysty blok z opisem umiejętności. Jedne karty dzielą liczby na nich zapisane, inne się różnią. Podobnie z dwoma ostatnimi kartami, obie mają identyczne wzory i napisy, różnią się tylko kolorami. Zastosowanie takiego charakteru bazy danych nie pozwala modelowi kierować się tylko na przykład samymi kolorami albo kształtami. Dzięki temu sieć w trakcie procesu nauki poszukuje unikalnych cech każdej karty i potem je z nią kojarzy. Inaczej sytuacja by wyglądało w przypadku rozpoznawaniu konkretnych gier karcianych, wtedy model musiałby się skupić na schematach, czyli wszystkich cechach wspólnych w ramach konkretnej gry.

2.3 Preprocessing obrazów

Na początek wykonano 5 zdjęć każdej z 21 kart pod różnym nachyleniem kamery, dzięki czemu zdjęcia miały zróżnicowane warunki oświetlenia.



Rysunek 2. Zdjęcia nie edytowane w rozdzielczości 3468x4624
Źródło: opracowanie własne.

W przypadku zdjęć na [Rysunek 2] tło zajmuje znaczną część obrazu, co może negatywnie wpływać na jakość i długość nauki. Dane zbierane przez program mogą zostać zaburzone, na przykład przez ilość danych do przetworzenia, kolor tła czy ilość kształtów widocznych w tle. W tym celu przycięto ręcznie zdjęcia tak, aby większość tła została odrzucona, aby model skupił się na samej karcie. W celu zapewnienia pewnego poziomu ujednolicenia położenia kart w obrębie zdjęcia, autor zdecydował się na ręczny proces kadrowania.

Rysunek 3. Przycięte zdjęcie o wymiarach 1921x2522
Źródło: opracowanie własne.

Zdjęcia po przycięciu nie były jeszcze gotowe do nauki, ponieważ były za duże, w konsekwencji czego nauka modelu trwałaby zbyt długo. Na dodatek po ręcznym przycinaniu rozmiary pomimo bycia w mniej więcej podobnym rozmiarze, to jednak różnice wielkości 300px negatywnie wpłynęłyby na efektywność procesu nauki. W wyniku tego przy pomocy języka Python oraz bibliotek ***cv2*** i ***os*** zdjęcia został zmniejszone do rozmiarów bardziej przystosowanych do uczenia sieci.

```
def resize_image(image_path, target_width, target_height):
    image = cv2.imread(image_path)resized_image = cv2.resize(image, (target_width, target_height))
    filename = os.path.basename(image_path)
    output_path = os.path.join('/CardRecognition/resized', 'resized_256x324_' + filename)
    cv2.imwrite(output_path, resized_image)
    print(f"Resized image saved to: {output_path}")
input_directory = '/CardRecognition/cropped'
output_directory = '/CardRecognition/resized'
os.makedirs(output_directory, exist_ok=True)
target_width = 256
target_height = 324
```

```
for filename in os.listdir(input_directory):
    if filename.endswith('.jpg') or filename.endswith('.png'):
        image_path = os.path.join(input_directory, filename)
        resize_image(image_path, target_width, target_height)
```

Listing 1. Program do zmieniania wielkości obrazu
Źródło: opracowanie własne.

Program przedstawiony w [Listing 1] kolejno wczytuje plik z katalogu ze zdjęciami, które zostały wcześniej przycięte, aby następnie zmienić ich rozmiar na **256x324**. Taki rozmiar został wybrany ze względu na to, że w miarę zachowuje proporcje karty, a wymiarowość danych jest przystępna dla modelu do uczenia się. Następnie edytowany plik zapisywany jest w katalogu ***"resized"*** dla ułatwienia dalszego procesu z nazwą "***resized_256x324_ + nazwa pliku wejsciowego"***. Funkcja wykonująca to zadanie znajduje się w pętli, która przechodzi po każdym pliku w katalogu ***"cropped"*** do momentu aż wszystkie pliki w nim zostaną poddan edycji.

Rysunek 4. Zdjęcie przeskalowane do 256x324
Źródło: opracowanie własne.

Jak można zauważyć zdjęcie z [Rysunek 4] w porównaniu do zdjęcia z [Rysunek 3] jest zdecydowanie słabszej jakości, a tekst na nim jest zdecydowanie mniej czytelny, co jednak powinno ułatwić sieci proces nauki poprzez znacząca redukcję wymiarowości danych.

Ostatnim krokiem przed ukończeniem bazy zdjęć jest uzyskanie obróconych zdjęć, aby program nie nauczył się rozpoznawać tylko tych kart, które są w ustawieniu pionowym. Ponownie wykorzystano do tego zadania język Python oraz te same biblioteki co przy przeskalowywaniu.

```
def rotate_image(image_path, rotation_degrees):
    image = cv2.imread(image_path)
    height, width = image.shape[:2]
    for i in range(11):
        rotation_matrix = cv2.getRotationMatrix2D((width/2, height/2), rotation_degrees * (i+1), 1)
        rotated_image = cv2.warpAffine(src=image, M=rotation_matrix, dsize=(width, height))
        filename, file_extension = os.path.splitext(os.path.basename(image_path))
        output_path = os.path.join('CardRecognition/rotated', f'rotated_{filename}_{i+1}{file_extension}')
        cv2.imwrite(output_path, rotated_image)
        print(f"Rotated image saved to: {output_path}")
input_directory = 'CardRecognition/resized'
output_directory = 'CardRecognition/rotated'
os.makedirs(output_directory, exist_ok=True)
target_width = 256
target_height = 324
for filename in os.listdir(input_directory):
    if filename.endswith('.jpg') or filename.endswith('.png'):
        image_path = os.path.join(input_directory, filename)
        rotate_image(image_path, 30)
```

Listing 2. Program do obracania zdjęć
Źródło: opracowanie własne.

Program opisany w [Listing 2] kolejno wczytuje plik z katalogu ze zdjęciami, które zostały wcześniej przeskalowane, aby następnie obrócić je o **30** stopni jedenaście razy. Powodem takiej parametryzacji procesu obróbki danych była chęć uzyskania parzystej liczby zdjęć w każdej klasie (łącznie z pionowym zdjęciem). Z katalogu "resized" zostają kolejno wczytywane wszystkie zdjęcia, a następnie obracane według ustalonych danych. Kolejno edytowany plik zapisywany jest w katalogu ***"rotated"*** z nazwą ***"rotated_nazwaPkiku_numerObrotu"***. Funkcja wykonująca to zadanie znajduje się w pętli, która przechodzi po każdym pliku w katalogu ***"resized"*** do momentu, aż wszystkie pliki w nim zostaną zedytowane. Przykładowe obrócone zdjęcia zostały przedstawione na [Rysunek 5].

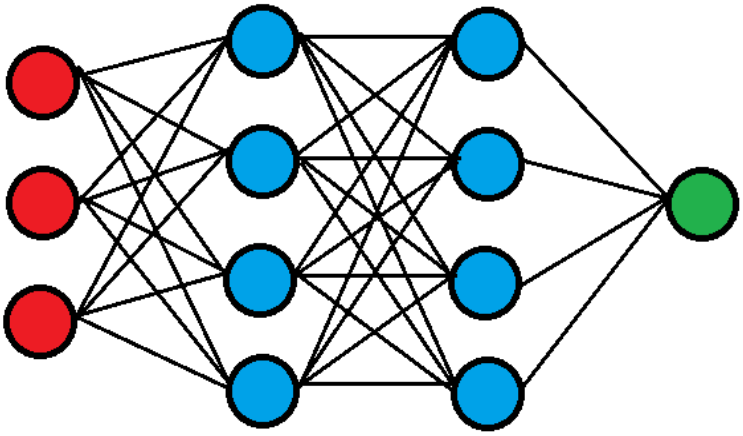
Rysunek 5. Zdjęcia z 4 przykładowymi obrotami
Źródło: opracowanie własne.

Jak można zauważyć zdjęcia w trakcie obracania nie zmieniają swojej wielkości w konsekwencji czego ich kawałek zostaje odcięty, a miejsce, w którym domyślnie był obraz zostaje zastąpione kolorem czarnym. Można by stwierdzić, że utrata części zdjęcia to coś złego. W tym przypadku to nic bardziej mylnego. Dzięki temu model uczy się w bardziej "rygorystycznych" warunkach. Oznacza to, że w przyszłości, gdy użytkownik poda do rozpoznania ucięte zdjęcie karty lub nawet porwaną kartę, istnieje większe prawdopodobieństwo prawidłowego zidentyfikowania obrazu.

3. Konwolucyjna sieć neuronowa

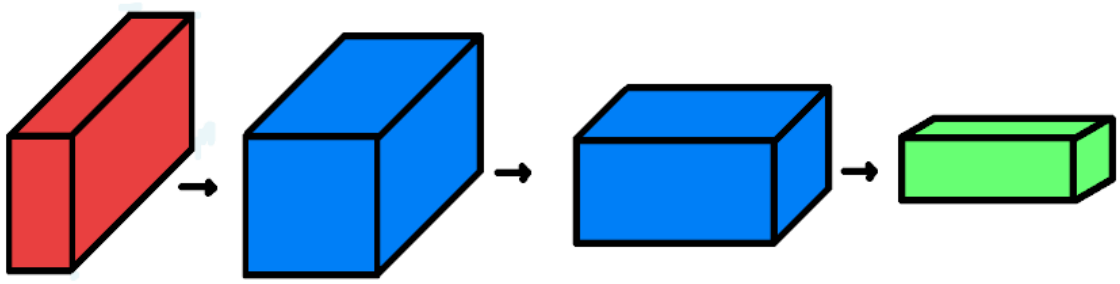
Konwolucyjna Sieć Neuronowa, z angielskiego CNN - Convolution Neural Network, inaczej Splotowa Sieć Neuronowa to jedna z metod Głębokiego Uczenia sieci neuronowych do analizy w głównej mierze obrazu, jednak jest też możliwe wykorzystania jej możliwości do rozpoznawania wzorów w dźwiękach, sygnałach czy też przedziałach czasowych. Jednak to nie koniec możliwych zastosowań. Innymi możliwościami związanymi z obrazami są, na przykład rozmazywanie/blurrowanie lub zmienianie stylistyki obrazu.

Typowa trójwarstwowa sieć neuronowa składa się z warstwy wejściowej, warstwy ukrytej oraz warstwy wyjściowej i wygląda to następująco [Rysunek 6]:



Rysunek 6. Wygląd klasycznej sieci neuronowej trójwarstwowe
Źródło: opracowanie własne.

Natomiast sieć konwolucyjna przygotowana dla obróbki obrazów składa się z neuronów ułożonych w trójwymiarze. W każdym z takich przypadków trójwymiarowa warstwa wejściowa zmieniana przez warstwy ukryte w trójwymiarową warstwę wyjściową. Dzieje się to przez fakt, że obraz jest zapisany w formacie **RGB**, w wyniku czego zamiast tylko wysokości i szerokości macierzy dochodzi też jej głębokość, ponieważ kolor czerwony jest reprezentowany przez macierz numer 1, zielony macierz numer 2, a niebieski numer 3, które to są na siebie nałożone, w konsekwencji czego powstaje macierz trójwymiarowa. Taką przykładową sieć można przedstawić w następujący sposób:



Rysunek 7. Przykładowy wygląd sieci konwolucyjnej
Źródło: opracowanie własne.

3.1 Warstwa wejściowa

Warstwa wejściowa w przypadku operowaniu na obrazach to warstwa, do której podawane jest zdjęcie do nauki. Rozmiar tej warstwy dyktowany jest przez rozmiar obrazu, czyli w przypadku tego projektu jest to 256x324 i na dodatek x3 ze względu na trzy kanały RGB. Skutkiem tego rozmiar warstwy wynosi 248 832 neurony.

3.2 Warstwa ukryta

Inaczej warstwa konwolucyjna, odpowiadająca za przetwarzenie otrzymanego z warstwy wejściowej obrazu oraz ta, w której dzieje się najwięcej obliczeń. Jest podstawowym elementem każdej sieci CNN. Może się w niej znajdować wiele warstw ukrytych w zależności od upodobania lub potrzeb. W skład każdej z nich wchodzi dowolna liczba neuronów, na dodatek nie musi być ona równa. Część tą można przedstawić jeszcze dokładniej, ponieważ składa się ona z takich procesów jak konwolucja, aktywacja oraz pooling. Od każdego z tych procesów zależy jak dokładnie bądź szybko model osiągnie wymagany cel.

3.3 Warstwa wyjściowa

W przypadku sieci CNN najczęściej nazywana jako warstwa w pełni połączona (fully connected layer) działa podobnie do tradycyjnej warstwy neuronowej, gdzie każdy neuron jest połączony z każdym neuronem w poprzedniej warstwie. Warstwy te są zazwyczaj używane na końcu sieci, aby skompresować wyekstrahowane cechy do określonej liczby klas w klasyfikacji.

4. Omówienie procesów

4.1 Konowlucja

Konwolucja jest kluczową operacją matematyczną w dziedzinie analizy sygnałów, przetwarzania obrazów oraz w kontekście sieci neuronowych, szczególnie w splotowych sieciach neuronowych. Zrozumienie, czym jest konwolucja, jest najważniejszą rzeczą, aby w pełni móc zrozumieć, w jaki sposób działają CNN i dlaczego są one potrzebne oraz tak skuteczne w analizie obrazów i innych form danych strukturalnych.

W skład warstwy konwolucyjnej wchodzą filtry, kernele, liczba przeskoków, padding, a na samym końcu aktywacja. **Filtry** odpowiadają za to jak zmieni się głębokość warstwy, naprzykład jeśli jej wymiary to 32x32x3 to po przejściu przez 64 filtry rozmiar będzie wynosił 32x32x64. **Kernele** to inny rodzaj filtrów wpływające na dwie pierwsze wartości macierzy, czyli wysokość i szerokość. Są to małe macierze, najczęściej w rozmiarach 2x2, 3x3 lub 5x5, w których zapisane są wartości pomagające programowi rozpoznać, na przykład krawędzie. Przykładowym kernelem jest pionowy ***filtr Sobela***, a jego wartości to:

-1	0	1
-2	0	2
-1	0	1

Rysunek 8. Filtr Sobela dla krawędzi pionowy
Źródło: opracowanie własne.

Kolejnym elementem jest padding. Nazywana jest tak technika dodawania dodatkowych pikseli w około obrazu wejściowego co przekłada się na dodanie zer wokół macierzy wejściowej, czyli w przypadku, gdy domyślnie rozmiar obrazu wynosi 8x8 po dodaniu paddingu wyniesie 9x9. Jednak, dlaczego taka operacja jest konieczna? Można wyróżnić dwa najważniejsze przypadki. Jednym z nich jest zmniejszenie się rozmiaru. Wynik operacji można uzyskać ze wzoru:

$$(n - f + 1)x(n - f + 1),$$

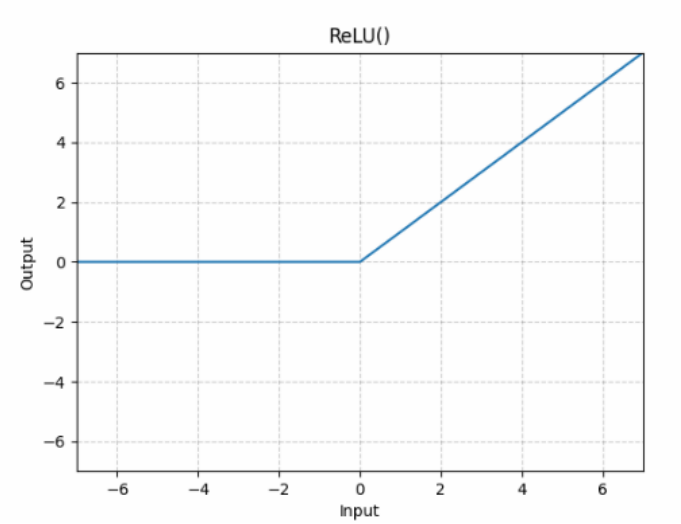
gdzie n oznacza rozmiar obrazu (n x n), a f rozmiar filtra (f x f), więc zdjęcie 8x8 po przejściu przez filtr 3x3 będzie miało rozmiar 6x6, zatem z każdym powtórzeniem operacji konwolucji obraz się zmniejsza co nakłada ograniczenie, ile razy można wykonać to działanie. Innym problemem pojawiającym się, gdy padding nie jest zaimplementowany jest fakt, że

niektóre wartości prawie wcale nie są sczytywane, dla przykładu wartości na rogach będą wzięte pod uwagę tylko jeden raz, na brzegach nawet tylko trzy razy w przypadku rozmiaru 6x6, podczas gdy wartości, znajdujące się bliżej centrum macierzy zostaną wzięte do obliczeń wiele razy.

Po wykonaniu tych czynności nadchodzi pora, aby otrzymana macierz przeszła przez funkcję aktywacji. Takich funkcji jest wiele jednak dwoma najczęściej używanymi są ***ReLu*** oraz ***softmax***. Wzór na ReLu wygląda następująco:

$$relu(x) = \max(0, x),$$

gdzie **x** oznacza wartość z macierzy, która przechodzi przez funkcje. W wyniku tego pozbywane są wszystkie liczby ujemne, dodatnie natomiast pozostają niezmienne.



Rysunek 9. Wykres funkcji ReLU
Źródło: opracowanie własne.

Funkcja **softmax** jest bardziej złożona obliczeniowo więc z reguły jest używana w ostatniej warstwie w momencie klasyfikacji obrazu. Jej głównym zadaniem jest przekształcanie wektora wartości rzeczywistych w wektor prawdopodobieństw, które sumują się do 1. Oznacza to, że każdy element wektora wejściowego jest eksponentowany, czyli podnoszony staje się potęgą liczby Eulera ***e***. Krok ten również jest zabezpieczeniem, aby na wyjściu nie było liczb ujemnych. Następnie eksponentowane wartości są następnie dzielone przez sumę wszystkich eksponentowanych wartości. Ten krok normalizuje wyniki tak, aby suma tych wynosiła 1, co jest wymagane dla prawdopodobieństw. Wzór funkcji wygląda następująco:

$$\text{softmax}(c_i) = \frac{e^{c_i}}{\sum_{j=1}^{n} e^{c_j}}.$$

4.2 Pooling

Obraz po przejściu przez filtry w warstwie konwolucyjnej, a właściwie jego rozmiar, jest w stanie drastycznie się zwiększyć, na przykład z 32x32x3 na 32x32x64. Powoduje to, że ilość obliczeń się zwiększa, czego konsekwencją jest wydłużenie się procesu wykonania kolejnych obliczeń. Aby zapobiec temu problemowi, zastosowuje się warstwę poolingu, która jest pewnym rodzajem agregacji danych, który oprócz wspomnianego skrócenia czasu działania sieci, za czym idzie uproszczenie modelu, to dodatkowo w konkretnych przypadkach pozwala zapobiec overfittingowi, czyli przeuczeniu modelu sieci. To zjawisko jest szczególnie niekorzystne w momencie, gdy sieć uczy się na ograniczonej ilości danych, o których nie ma pewności, czy są poprawne i wymagane do późniejszego użytku w przyszłości. Jednym z przykładów może być model operujący na giełdzie, który się nauczył na danych z tylko kilku miesięcy. Jego działanie długoterminowe może nie przynieść oczekiwanych skutków, czyli zysków, a wręcz przeciwnie może przysłużyć się do strat.

Jak duży wpływ na rozmiar macierzy ma warstwa poolingu można przedstawić samymi wzorami. Jeśli rozmiar tego “filtra” zostanie przyjęty na 2x2, a wielkość przeskoku wyniesie 2, do przewidzenia wyniku tej operacji jest możliwe wykorzystanie poniższych wzorów:

$$W2 = \frac{(W1-F)}{S} + 1,$$

$$H2 = \frac{(H1-F)}{S} + 1,$$

$$D2 = D2\ ,$$

gdzie W1, H1 i D1 to nic innego jak szerokość (Width) x wysokość (Height) x głębokość (Depth) obrazu, który chodzi do tej warstwy, a F i S, to kolejno rozmiar filtra oraz wielkość skoku. Dla powyższych danych oraz dla macierzy wejściowej 6x6x3, po przejściu przez pooling otrzymany obraz posiada rozmiar 3x3x3.

Operację tą można podzielić na dwa różne rodzaje - MaxPooling oraz AvgPooling. Dzięki zachowywaniu najważniejszych informacji, MaxPool jest w stanie bezproblemowo pozbyć się zbędnych cech na obrazie takich jak szum. Pozwala to na bardziej szczegółowe rozpoznawanie krawędzi, więc najlepiej tą funkcje stosować w modelach, których celem jest detekcja obiektów, która posiadają specyficzne i wyraziste cechy. Przykładem tego jest rozpoznawanie twarzy, każda wyróżniająca rysa będzie zarejestrowana co pomoże uefektywnić naukę. Podobnie jest z rozpoznawaniem kształtów. W tym przypadku tło zostanie bardziej zignorowane, a sieć skupi się na konturach

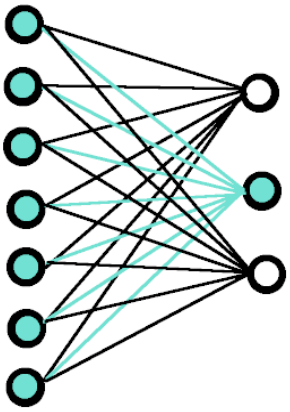
Pomimo wszystkich zalet MaxPoolingu, AvgPooling również ma swoje zastosowania. Uśrednianie wartości pozwala pozbycia się pewnego rodzaju “anomalii” na zdjęciu, to jest pojedynczych ekstremalnych wartości. Funkcja skupia się głównie na kontekście obrazu, co jest potrzebne przy segmentacji obrazów lub analizy tekstur. Jednak najbardziej wyróżniającym wykorzystaniem uśredniania jest analiza medyczna. Przy badaniu skanów rentgenowskich każda niestandardowa zmiana w ciele zostanie zarejestrowana. Funkcja ta znalazła zastosowanie przy badaniu zmian w strukturze płuc, na przykład po covidzie. Na samym początku stosowana funkcja MaxPool do pozbycia szumów, następnie do samego końca AvgPool. Proces ten został dokładniej udokumentowany w artykule [3].

4.3 Odrzucanie i spłaszczanie

Warstwa **Dropout** ma przekazaną wartość ułamkową z jaką częstotliwością ma odrzucać dane, natomiast spłaszczanie, z angielskiego **Flattening**, to proces, który stosowany jest przed warstwą wyjściową, jednak jest możliwość wliczać do “w pełni połączonych warstw”. Jego celem jest zmienić kształt macierzy, aby w prosty sposób było możliwe przekazania danych do standardowej warstwy neuronów.

4.4 FC

Na sam koniec procesu dane wchodzą do warstwy wyjściowej a w tym przypadku dokładniej, “w pełni połączonych warstw”, z angielskiego Fully Connected Layers, w skrócie FC. Odpowiadają one za pobranie jako wejście danych z poprzednich operacji, a następnie wykonują obliczenia dla ostatecznego sklasyfikowania obrazu lub innego badanego obiektu. Dla przykładu dane mogą przejść najpierw przez 512 neuronów, które swoje wyjście przekazują do kolejnych 256, w obydwu tych warstwach funkcją aktywacyjną jest ReLu dla efektywnego procesu obliczeń. Jednak liczba neuronów dla ostatniej warstwy jest z góry określona, a dokładnie w momencie utworzenia zbioru danych do uczenia. Otóż liczba wyjść, a zatem neuronów na wyjściu, musi być równa ilości unikalnych klas obrazów. W tym etapie najczęściej używaną funkcją aktywacyjną jest softmax, ze względu na większą szczegółowość.



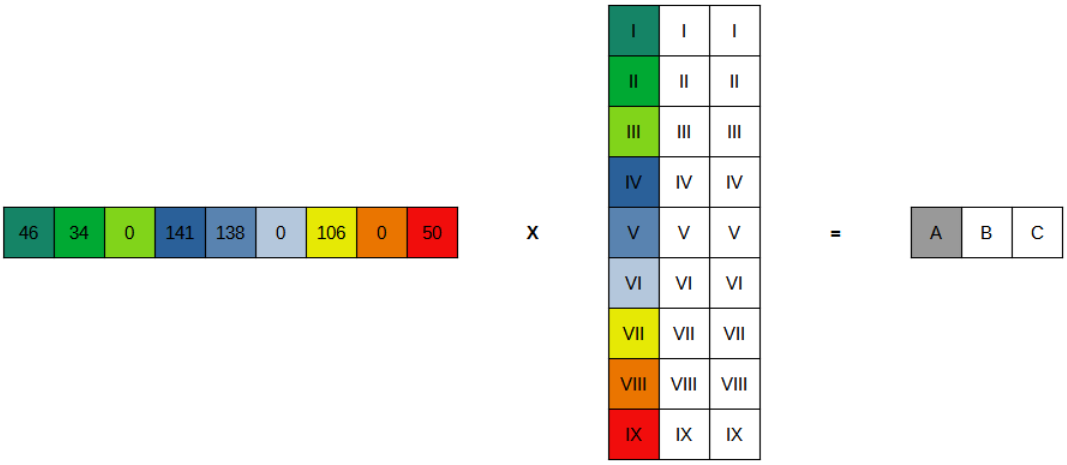
Rysunek 10. Przedstawienie sposobu w jakim łączone są neurony w FC
Źródło: opracowanie własne.

Należy również zwrócić uwagę na to, że w warstwie w pełni połączonej każdy neuron jest połączony z każdym neuronem znajdującym się w poprzedniej warstwie, tak jak to widać na [Rysunek 10]. Oznacza to, że każda warstwa FC przyjmuje wszystkie wyjścia z poprzedniej warstwy jako wejścia, a każde wyjście z warstwy FC jest kombinacją liniową tych wejść. Warstwę w pełni połączoną, w przypadku, gdy podany jest wektor wejściowy cech $\mathbf{x=[x1,x2,\ldots,xn]}$, można ją opisać matematycznie następującym wzorem:

$$y_i = \sigma \sum_{j=1}^{n} w_{ij}\, x_i + b_j,$$

gdzie:

- w_{ij} to waga łącząca i-te wejście z j-tym neuronem,
- b_j to przesunięcie (bias) dla j-tego neuronu,
- σ to funkcja aktywacji, czyli w przypadku tego projekty ReLu lub softmax.



Rysunek 11. Proces obliczania wartości wyjściowej
Źródło: opracowanie własne.

Jak przedstawiono na [Rysunek 11], wcześniej spłaszczona macierz zostaje przekazana do modelu podobnego jak na [Rysunek 10], tylko w tym przypadku złożonym z 9 neuronów, a następnie dane z wektora wejściowego mnożone są z macierzą wag w wyniku czego powstaje wektor wyjściowy.

5. Stworzenie architektury i badania

Wykorzystując wcześniej przedstawione informacje, autor sporządził optymalną architekturę przy pomocy biblioteki TensorFlow. Wspomniana struktura znajduje się poniżej:

```
cnn_model = Sequential()
cnn_model.add(Conv2D(filters=64, kernel_size=5, strides=(2, 2), padding='same', activation='relu',
    input_shape=[256, 324, 3]))
cnn_model.add(AvgPool2D(pool_size=2, strides=2))
cnn_model.add(Conv2D(filters=64, kernel_size=5, strides=(2, 2), padding='same', activation='relu'))
cnn_model.add(AvgPool2D(pool_size=2, strides=2))
cnn_model.add(Conv2D(filters=64, kernel_size=3, padding='same', activation='relu'))
cnn_model.add(AvgPool2D(pool_size=2, strides=2))

cnn_model.add(tf.keras.layers.Dropout(0.4))

cnn_model.add(Flatten())
cnn_model.add(Dense(units=512, activation='relu'))
cnn_model.add(Dense(units=256, activation='relu'))
cnn_model.add(Dense(units=21, activation='softmax'))
```

Listing 3. Budowa architektury
Źródło: opracowanie własne.

Proces uczenia rozpoczyna się dzięki wywołaniu funkcji:

```
    model.compile(optimizer=tf.keras.optimizers.Adam(learning_rate=learning),
loss='categorical_crossentropy', metrics=['accuracy'])
```

Listing 4. Konfigurowanie modelu uczenia
Źródło: opracowanie własne.

Wybranym na poczet tego projektu optymalizatorem jest **adam**. To zaawansowany optymalizator, który łączy zalety dwóch innych : RMSProp i Momentum. Adam adaptacyjnie dostosowuje współczynniki uczenia dla poszczególnych wag i wykorzystuje momenty gradientów. Aktualizacje wag w Adam są obliczane według wzoru:

$$m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t,$$

$$v_t = \beta_2 v_{t-1} + (1-\beta_2) g_t^2,$$

$$\overline{m_t} = \frac{m_t}{1-\beta_1^t},$$

$$\overline{v_t} = \frac{v_t}{1-\beta_2^t},$$

$$\theta_t = \theta_{t-1} - \eta \frac{\overline{m_t}}{\sqrt{\overline{v_t}}+\epsilon},$$

gdzie:

- θ to waga modelu,
- m_t oraz v_t to pierwsze i drugie momenty gradientów,
- β_1 i β_2 to współczynniki dla momentów,
- η to współczynnik uczenia,
- ϵ to mała wartość zapobiegająca dzieleniu przez zero.

"loss = categorical_crossentropy", czyli połączoenie funkcji aktywacyjnej softmax z krzyżową entropią, za pomocą których ostatecznie będą klasyfikowane obrazy. Funkcje straty kategorycznej entropii krzyżowej najlepiej przedstawić następująco. Dla danego przykładu i, gdzie yi jest rzeczywistą etykietą, a yi jest wektorem przewidywanych prawdopodobieństw dla każdej klasy, funkcja straty kategorycznej entropii krzyżowej jest definiowana jako:

$$Loss(y_i, \overline{y_i}) = -\sum_{j=1}^{K} y_{ij} \log(y_i),$$

gdzie, K to liczba klas, y_{ij} to element j wektora rzeczywistych etykiet y_i, który jest równy 1, jeśli rzeczywista klasa to j, a w przeciwnym razie 0. $\overline{y_i}$ to element j wektora przewidywanych prawdopodobieństw $\overline{y_i}$ dla klasy j.

Summary() to wizualne przedstawienie w konsoli wszystkich warstw oraz rozmiaru danych, które zostały otrzymane po wszystkich obliczeniach. Przykładowe zestawienie przedstawiono na [Rysunek 12].

Layer (type)	Output Shape	Param #
conv2d_3 (Conv2D)	(None, 256, 256, 32)	896
max_pooling2d_3 (MaxPooling2D)	(None, 128, 128, 32)	0
conv2d_4 (Conv2D)	(None, 126, 126, 32)	9,248
max_pooling2d_4 (MaxPooling2D)	(None, 63, 63, 32)	0
conv2d_5 (Conv2D)	(None, 61, 61, 32)	9,248
max_pooling2d_5 (MaxPooling2D)	(None, 30, 30, 32)	0
dropout_1 (Dropout)	(None, 30, 30, 32)	0
flatten_1 (Flatten)	(None, 28800)	0
dense_2 (Dense)	(None, 1024)	29,492,224
dense_3 (Dense)	(None, 23)	23,575

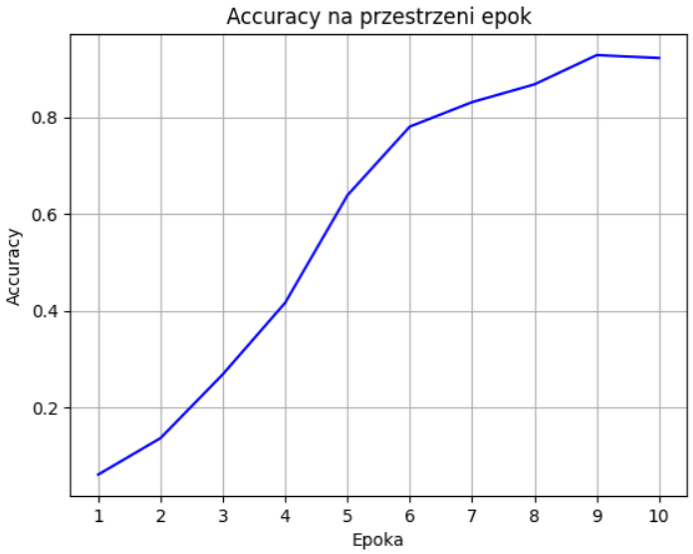
Total params: 29,535,191 (112.67 MB)

Rysunek 12. "Streszczenie" operacji na każdej z warstw
Źródło: opracowanie własne.

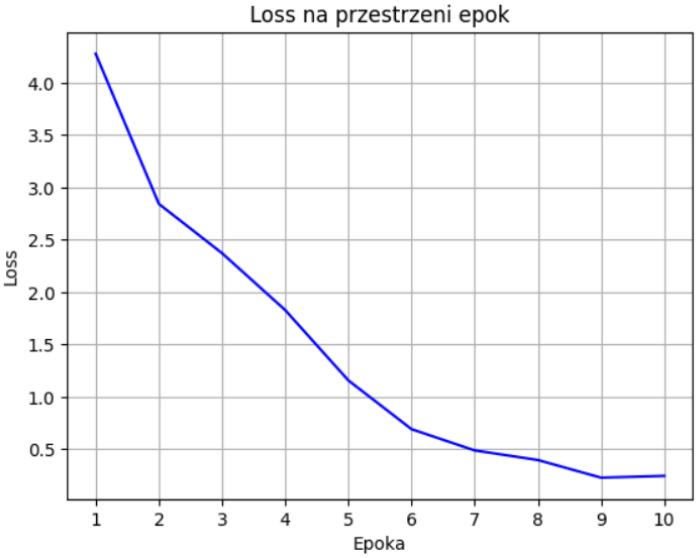
```
train_loss, tarin_acc = cnn_model.evaluate(training_image_set)
ep= [i for i in range(0,10)]
plot.plot(ep, history_of_training.history['accuracy'],label="Training acc")
```

Listing 5. Ocena modelu i tworzenie wykresu
Źródło: opracowanie własne.

Metoda **evaluate()** pozwala na ocenę modelu na podstawie danych treningowych i pozyskać ostateczną wartość celności i straty. Następnie za pomocą takich danych jest możliwość wygenerować wykres, który przedstawia proces nauki. Oczekiwanym wynikiem dla straty (loss) jest wartość zbliżona 0 jednak nie równa, natomiast dla celności (accuracy) zbliżona do 1 ale nie równa.



Rysunek 13. Przykładowy wykres celności na przestrzeni 10 epok
Źródło: opracowanie własne.



Rysunek 14. Przykładowy wykres straty na przestrzeni 10 epok
Źródło: opracowanie własne.

```
y_pred = cnn_model.predict(test_image_set)
predicted_categories = tf.argmax(y_pred, axis=1)
true_categories =  tf.concat([y for x,y in test_image_set], axis=0)
Y_true = tf.argmax(true_categories, axis=1)
```

Listing 6. Przedstawienie przewidywanych i faktycznych kategorii
Źródło: opracowanie własne.

Dzięki funkcji **Predict()** model jest wstanie wytypować, do jakiej klasy przynależą obrazy znajdujące się w katalogu testowym. Dane z predykcji są zapisane w tablicy jednak są one trudne do odczytania przez człowieka oraz są niekorzystne do dalszych obliczeń, dlatego należy je przekształcić.

```
(array([[1.58302680e-01, 9.47204530e-02, 8.86514187e-02, ...,
         1.77595690e-02, 3.52344066e-02, 3.96764092e-02],
        [1.00719996e-01, 2.53435135e-01, 1.23483203e-02, ...,
         1.55811995e-01, 2.80559268e-02, 1.97136998e-02],
        [1.61103830e-01, 4.13037352e-02, 8.89345165e-03, ...,
         2.56890543e-02, 1.20725408e-02, 1.12476572e-02],
        ...,
        [1.10686822e-02, 4.85996105e-04, 1.92247182e-02, ...,
         6.66744600e-04, 1.98434796e-02, 5.60685337e-01],
        [5.39286109e-03, 2.86174635e-03, 2.08529755e-02, ...,
         1.19206589e-03, 1.78178884e-02, 2.20210567e-01],
        [4.16238531e-02, 3.11301183e-02, 9.76318046e-02, ...,
         7.13212462e-03, 3.05144489e-02, 5.14020249e-02]], dtype=float32),
```

Rysunek 15. Przedstawienie tabeli predykcji klas
Źródło: opracowanie własne.

```
array([ 0,  1,  0,  1, 15,  0,  0, 17,  9,  0,  0,  0,  1,  1,  5,  1, 15,
        0,  0,  1,  2, 19,  5, 19,  2,  2,  2, 13,  2,  2,  2, 17,  2,  2,
        2,  2,  3, 14,  7, 14,  3, 14,  3, 17,  3, 14,  3,  3,  1,  8,  5,
        4, 11,  4,  5, 16, 10, 11,  4,  1,  7,  5,  5,  5,  5,  4,  5, 16,
        7,  5,  5, 18, 15, 10, 13, 19, 10,  4,  4,  4,  0,  2,  6,  7,  2,
        7,  7, 19,  7,  0,  7, 10,  0,  8,  6,  0, 14,  2,  7, 19,  2,  8,
        7,  4,  8,  8, 17, 10,  9,  9,  9,  9,  9,  9,  9,  9,  9,  9,  9,
        9, 10, 10,  8,  1,  8,  6,  8,  8, 10, 10, 10, 10, 10,  4, 15, 15,
       11, 11, 18,  0, 15,  3, 11, 11, 12, 12, 12, 12, 12,  1,  0, 17, 15,
       12,  7, 12, 10,  4, 13, 10, 19,  2,  2, 19, 10,  2, 17,  2, 14,  3,
        3, 14, 16,  3, 14, 14, 14, 14, 14,  3, 15, 15, 14, 15, 15,  1,  3,
        3, 15, 15, 15, 14, 15, 15,  7, 19, 19,  7,  5,  4, 16, 11,  4, 15,
       14,  0, 20, 20, 15,  7,  0,  1,  2, 12, 16,  7,  0, 18, 18, 18, 18,
        1, 18, 18,  0, 18, 18, 18, 15,  7, 20, 19, 19,  4, 12,  1, 12, 12,
        6,  4,  2,  2, 19,  2, 19, 20,  3, 16, 20, 20, 20,  5],
```

Rysunek 16. Przekształcona tabela predykcji klas
Źródło: opracowanie własne.

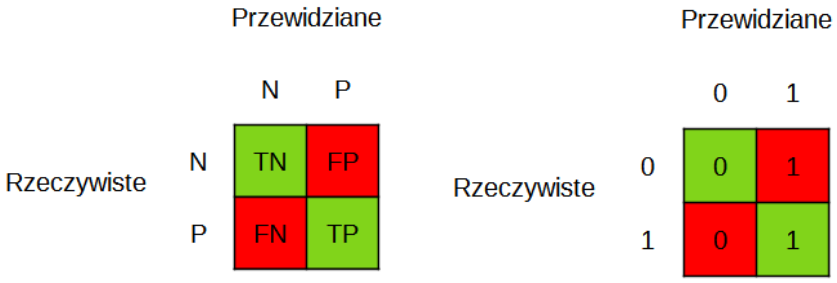
Każda liczba z tablicy na [Rysunek 16] reprezentuje konkretną klasę, którą przypisał model dla danego obrazu, który jest reprezentowany numerem indeksu tabeli. Analogicznie tabela ma taką długość, jaka jest ilość obrazów w katalogu testowym. Następnie wykonano raport z klasyfikacji.

	precision	recall	f1-score	support
Bartholomew_Kuma	0.37	0.58	0.45	12
Basil_Hawkins	0.29	0.33	0.31	12
Bunny_Joe	0.43	0.83	0.57	12
Chopa_Emon	0.43	0.50	0.46	12
Dragon_claw	0.21	0.25	0.23	12
Enel	0.54	0.58	0.56	12
Gamma_Knife	0.25	0.08	0.12	12
Hack	0.25	0.33	0.29	12
Hino_Bird_Zap	0.33	0.25	0.29	12
Lider_Back	0.92	1.00	0.96	12
Lieuthenant_Spacey	0.40	0.50	0.44	12
Monkey_D_Garp	0.57	0.33	0.42	12
Monkey_D_Luffy	0.64	0.58	0.61	12
Nefeltari_Kobra	0.33	0.08	0.13	12
Nola	0.47	0.58	0.52	12
Regular_Back	0.37	0.58	0.45	12
Revolutionary_Army_HQ	0.17	0.08	0.11	12
Sabo	0.00	0.00	0.00	12
Sakazuki	0.82	0.75	0.78	12
Sterry	0.15	0.17	0.16	12
Tho_Toh	0.57	0.33	0.42	12
accuracy			0.42	252
macro avg	0.41	0.42	0.39	252
weighted avg	0.41	0.42	0.39	252

Rysunek 17. Przekształcona tabela predykcji klas
Źródło: opracowanie własne.

Do obliczenia wartości znajdujących się na [Rysunek 17] potrzebne są tylko 4 zmienne:

- TN (True Negative) - to wynik wskazujący, że warunki nie zostały spełnione i w rzeczywistości tak jest,
- TP (True Positive) - to wynik wskazujący, że warunki zostały spełnionei w rzeczywistości tak jest,
- FN (False Negative) - to wynik wskazujący, że warunki nie zostały spełnione, a w rzeczywistości jest na odwrót,
- FP (False Positive) - to wynik wskazujący, że warunki zostały spełnione, a w rzeczywistości jest na odwrót.



Rysunek 18. Test diagnostyczny przedstawiony słownie i binarnie
Źródło: opracowanie własne.

Następnie liczbę wystąpień każdego stanu należy wykorzystać do wzorów:

$$\text{Precision} = \frac{TP}{TP+FP},$$

$$\text{Recall} = \frac{TP}{TP+FN},$$

$$F1 = (2 \cdot \text{Precision} \cdot \text{Recall}) / (\text{Precision} + \text{Recall}),$$

$$\text{Accuracy} = \frac{TP+TN}{TP+FN+TN+FP},$$

gdzie:

- Precison to miara, która sprawdza, ile obrazów zostało sklasyfikowanych jako obraz A, w porównaniu do wszystkich, które zostały uznane za obraz A, nawet jeśli to nie prawda. Jeśli precyzja wynosi 0.4, oznacza to, że 40% obrazów sklasyfikowanych jako obraz A, faktycznie nim było.
- Recall to miara, która sprawdza, ile obrazów zostało poprawnie sklasyfikowanych jako obraz A, w porównaniu do wszystkich, które rzeczywiście były obrazem A. Jeśli precyzja wynosi 0.5, oznacza to, że model był wstanie rozpoznać 50% obrazów z kategorii A.
- F1 to średnia harmoniczna z Precision i Recall, która daje ogólną ocenę, jak model dobrze klasyfikuje obrazy jako A, w porównaniu do faktycznej ich ilości.
- Accuracy to miara, która informuje jaki procent wszystkich obrazów sklasyfikował poprawnie.

Jak można zauważyć, model na [Rysunek 17] miał celność tylko na poziomie 42%, co jest zdecydowanie niesatysfakcjonujące. Jednak analiza nie kończy się w tym miejscu. Dane otrzymane z klasyfikacji można przekształcić w confusion matrix (Macierz Pomyłek), która przedstawia, ile razy dana klasa była rozpoznana jako inna z klas.

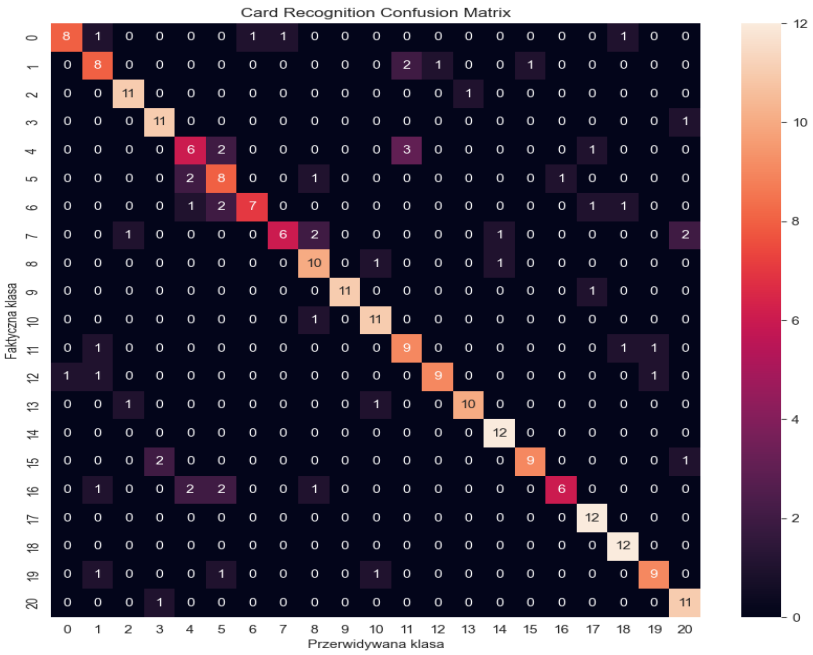
```
confusion = confusion_matrix(Y_true, predicted_categories)
plot.figure(figsize=(10, 10))
heatmap = sns.heatmap(confusion, annot=True)
```

Listing 7. Tworzenie i wizualizacja macierzy pomyłek
Źródło: opracowanie własne.

Funkcja **confusion_matrix** po przekazaniu do niej danych z oszacowanych klas oraz prawdziwych klas zwraca następującą macierz [Rysunek 19], która następnie przedstawiana jest w formie heatmapy (mapy ciepła), którą przestawiono na [Rysunek 4.20].

```
array([[ 8,  1,  0,  0,  0,  0,  1,  1,  0,  0,  0,  0,  0,  0,  0,  0,
         0,  0,  1,  0,  0],
       [ 0,  8,  0,  0,  0,  0,  0,  0,  0,  0,  0,  2,  1,  0,  0,  1,
         0,  0,  0,  0,  0],
       [ 0,  0, 11,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  1,  0,  0,
         0,  0,  0,  0,  0],
       [ 0,  0,  0, 11,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,
         0,  0,  0,  0,  1],
       [ 0,  0,  0,  0,  6,  2,  0,  0,  0,  0,  0,  3,  0,  0,  0,  0,
         0,  1,  0,  0,  0],
       [ 0,  0,  0,  0,  2,  8,  0,  0,  1,  0,  0,  0,  0,  0,  0,  0,
         1,  0,  0,  0,  0],
       [ 0,  0,  0,  0,  1,  2,  7,  0,  0,  0,  0,  0,  0,  0,  0,  0,
         0,  1,  1,  0,  0],
       [ 0,  0,  1,  0,  0,  0,  0,  6,  2,  0,  0,  0,  0,  0,  1,  0,
         0,  0,  0,  0,  2],
       [ 0,  0,  0,  0,  0,  0,  0,  0, 10,  0,  1,  0,  0,  0,  1,  0,
         0,  0,  0,  0,  0],
       [ 0,  0,  0,  0,  0,  0,  0,  0,  0, 11,  0,  0,  0,  0,  0,  0,
         0,  1,  0,  0,  0],
       [ 0,  0,  0,  0,  0,  0,  0,  0,  1,  0, 11,  0,  0,  0,  0,  0,
         0,  0,  0,  0,  0],
       [ 0,  1,  0,  0,  0,  0,  0,  0,  0,  0,  0,  9,  0,  0,  0,  0,
         0,  0,  1,  1,  0],
       [ 1,  1,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  9,  0,  0,  0,
         0,  0,  0,  1,  0],
       [ 0,  0,  1,  0,  0,  0,  0,  0,  0,  0,  1,  0,  0, 10,  0,  0,
         0,  0,  0,  0,  0],
       [ 0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0, 12,  0,
         0,  0,  0,  0,  0],
       [ 0,  0,  0,  2,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  9,
         0,  0,  0,  0,  1],
       [ 0,  1,  0,  0,  2,  2,  0,  0,  1,  0,  0,  0,  0,  0,  0,  0,
         6,  0,  0,  0,  0],
       [ 0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,
         0, 12,  0,  0,  0],
       [ 0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,
         0,  0, 12,  0,  0],
       [ 0,  1,  0,  0,  0,  1,  0,  0,  0,  0,  1,  0,  0,  0,  0,  0,
         0,  0,  0,  9,  0],
       [ 0,  0,  0,  1,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,
         0,  0,  0,  0, 11]], dtype=int64)
```

Rysunek 19. Macierz pomyłek dla modelu z precyzja 78%
Źródło: opracowanie własne.



Rysunek 20. Heatmaoa dla modelu z precyzja 78%
Źródło: opracowanie własne.

6. Podsumowanie

Konwolucyjne sieci neuronowe są rzeczą wręcz konieczną przy uczeniu modelu rozpoznawania jakichkolwiek zdjęć. Umożliwia to uzyskać oczekiwane efekty w relatywnie szybkim czasie, jednak na to ma również wpływ używany sprzęt. Obliczenia nawet na dobrym CPU są znacznie wolniejsze niż na średniej karcie graficznej, z tego powodu w dalszym etapie rozwoju tego projektu koniecznością jest zoptymalizowanie biblioteki TensorFlow tak, aby operowała na GPU. Dzięki temu proces uczenia będzie trwać aż dwa razy krócej.

Model nauczył się w sposób oczekiwany, gdyż bezbłędnie rozpoznaje obrazy, na których nie był trenowany. W tej sytuacji warto będzie wykorzystać wybraną architekturę w przyszłości do rozszerzonej bazy. Warto ją rozwinąć o więcej kart w podobnym stylu, ale równiej wprowadzić większą różnorodność poprzez, na przykład karty holograficzne lub karty z innych gier. Dzięki temu model będzie w stanie rozpoznawać postacie z kart, które nawet w tym momencie się nie ukazały.

Inną opcją rozwoju jest stworzenie aplikacji, w której użytkownik może wybrać zdjęcie ze swojego urządzenia a następnie nauczony model je sklasyfikuje. Następnie istniałaby możliwość dodania karty do kolekcji. Taka aplikacja w przyszłości zostanie zaimplementowana do większego projektu koła naukowego RUT-AI Applications realizowanego przez koł naukowe Interakcji Człowiek-Komputer GEST.

Podsumowując, CNN jest bardzo przydatne do obróbki i klasyfikowania obrazów, co nie tylko będzie w dużej mierze wykorzystywane w przyszłości, a w zasadzie jest używane już na oczach codziennych użytkowników internetu.

Źródła internetowe

1. https://www.geeksforgeeks.org/introduction-convolution-neural-network/ (dostęp: 26.05.2024)
2. https://www.geeksforgeeks.org/cnn-introduction-to-padding/ (dostęp: 26.05.2024)
3. https://www.sciencedirect.com/science/article/pii/S1568494622007050 (dostęp: 27.05.2024)
4. https://cs231n.github.io/convolutional-networks/#overview (dostęp: 26.05.2024)
5. https://en.wikipedia.org/wiki/Convolution (dostęp: 26.05.2024)
6. https://keras.io/api/ (dostęp: 26.05.2024)
7. https://stanford.edu/~shervine/teaching/cs-230/cheatsheet-convolutional-neural-networks (dostęp: 26.05.2024)
8. https://builtin.com/machine-learning/fully-connected-layer (dostęp: 27.05.2024)
9. https://www.geeksforgeeks.org/adam-optimizer/ dostęp: (28.05.2024)

Michnik Łukasz
Koło Naukowe Interakcji Człowiek – Komputer „GEST”

mgr. inż. Dawid Kalandyk
Opiekun Naukowy

Analiza działania sieci CNN oraz propozycja modelu rozpoznającego cyfry pisane odręcznie

Artykuł opisuje zasady działania konwolucyjnej sieci neuronowej CNN, oraz jej poszczególnych warstw, a także konkretne operacje i obliczenia przez nie realizowane. Autor przedstawia również budowę proponowanego modelu sieci w języku Python, oraz opisuje proces jej szkolenia na bazie MNIST, która składa się z rękopisów cyfr. Artykuł analizuje również szereg aspektów procesu nauki proponowanej sieci, który po wyuczeniu zostaje przekonwertowany na język JavaScript i zaimplementowany na stronie internetowej w celu umożliwienia każdemu użytkownikowi własnoręcznego jego przetestowania. Jest to część większego projektu koła naukowego GEST o nazwie Rzeszow University of Technology – Artifical Intelligence Applications (RUT-AI Applications).

Słowa kluczowe: CNN, DCNN, uczenie maszynowe, rozpoznawanie cyfr, sztuczna inteligencja.

1. Wprowadzenie

Sieć splotowa (ang. Convolutional Neural Network – CNN) jest jedną z najpopularniejszych sieci neuronowych służących do klasyfikacji obrazów i filmów. Sieć przyjmuje zdjęcie, przetwarza je analizując poszczególne piksele, a następnie klasyfikuje na podstawie określonych kategorii, rozpoznając rzeczy i obiekty na zdjęciu. Problematyka artykułu opiera się na analizie ww. sieci, opisu działania algorytmów oraz przedstawieniu szkolenia przykładowego modelu przy użyciu języka Python z biblioteką Keras. Na przykładzie procesu szkolenia sieci zostanie pokazany proces doboru konkretnych warstw na podstawie wyników otrzymywanych podczas kolejnych testów. Wykorzystana baza danych służąca do nauki to popularna baza MNIST. Zawiera ona obrazy w rozmiarze 28x28 pikseli przedstawiające pisane odręcznie cyfry od 0 do 9, gdzie każdy piksel opisany jest wartością z przedziału [0; – 255]. Wartość 0 opisuje kolor biały, natomiast 255 jest kolorem czarnym, co jest równoznaczne z pozostawieniem przez użytkownika śladu atramentu w obrębie tego piksela. Następnie utworzony model zostanie przekonwertowany do języka JavaScript i zaimplementowany na stronie internetowej w celu pokazania działania sieci na przykładach stworzonych przez samych użytkowników. Użytkownik będzie w stanie samodzielnie napisać swoją cyfrę, a model po analizie otrzymanego obrazu określi wynik i wypisze go na ekranie. Wykonanie tej części, będzie opierało się na wykorzystaniu biblioteki React, wraz z frameworkiem – Next.js. Model po przekonwertowaniu na język JavaScript zostanie zaimplementowany do aplikacji przy użyciu dodatkowej biblioteki Tensorflow.js, która

umożliwia łatwe testowanie. Projekt, którego dotyczy niniejszy artykuł jest częścią większego projektu koła naukowego GEST o nazwie Rzeszow University of Technology – Artifical Intelligence Applications (RUT-AI Applications). Po ukończeniu prac utworzona aplikacja wraz z opisem teoretycznym będzie dostępna na stronie koła.

2. Konwolucyjna sieć neuronowa

Sieć CNN (ang. Convolutional Neural Network) jest specjalnym rodzajem sieci neuronowej, która została zaprojektowana do analizy danych przestrzennych, takich jak obrazy czy filmy. Przetwarza poszczególne pixele otrzymanego zdjecia, aby później móc wyciągnąć wnioski i je sklasyfikować. Architektura sieci oparta jest na różnych rodzajach warstw neuronów. Każda sieć neuronowa wyróżnia podstawowy podział na warstwy neuronów:

- warstwa wejściowa,
- warstwy ukryte,
- warstwa wyjściowa.

Sieć konwolucyjna również może zostać sklasyfikowana w ten sposób, jedynie z tą różnicą, że jest ona znacznie bardziej skomplikowana z powodu ogromnej różnorodności warstw ukrytych. Każda z nich wykonuje swoją ściśle określoną operację. Ich działanie głównie polega na operacjach na macierzach, dla przykładu warstwa konwolucyjna (Convolutional layer) iteruje po kolejnych składnikach macierzy pixeli za pomocą filtra (kernel) o rozmiarze najczęściej 3x3 lub 5x5, obliczając tzw. mapę cech pomiędzy pixelami obrazu, a wagami w filtrze. Wyróżniamy wiele rodzajów warstw kryjących się pod pojęciam ukrytych, natomiast zostaną one omówione w dalszej części artykułu. Na początku, obraz trafia na warstwę wejściową, przyjmuje ona surowe dane wejściowe, najczęściej obrazy w postaci tensorów o wymiarach (wysokość, szerokość, liczba kanałów). Każdy piksel obrazu jest traktowany jako wejściowy neuron tej warstwy. W późniejszym etapie dane pochodzące z pierwszej wejściowej warstwy trafiają na kolejne, dane wyjściowe poprzedniej warstwy, są danymi wejściowymi kolejnej. Poddanie obrazu pod działanie tych warstw skutkuje w ostatnim etapie otrzymaniem neuronów w liczbie odpowiadającej liczbie klas, w przypadku bazy danych MNIST wymagane jest 10 neuronów, określają one poszczególne cyfry jakie możemy napisać. Najbardziej aktywny neuron jest wynikiem operacji i to właśnie jego numer jest zwracany jako cyfra napisana na obrazie. Poniżej znajduje się dokładny opis czterech najważniejszych warstw w sieci konwolucyjnej, będą one wykorzystane w budowie modelu rozpoznającego cyfry na obrazie.

2.1. Warstwa konwolucyjna (ang. Convolutional Layer)

Serce algorytmu sieci CNN, jest najważniejszą warstwą odpowiadającą za obliczanie tzw. iloczynu skalarnego. Wykonuje operacje macierzowe, przesuwając filtr (kernel) o rozmiarze najczęściej 3x3 lub 5x5 po macierzy pixeli obrazu, mnożąc z każdą iteracją wartości pixeli z wartościami wag w filtrze. Powstała w ten sposób macierz, w zależności od konfiguracji sieci, może zmieniać rozmiar danych wyjściowych, np. w przypadku ustawienia braku paddingu, podana na wejście macierz 28x28 po przejsciu przez warstwę konwolucyjną będzie miała rozmiar 26x26, natomiast z ustawieniem z paddingiem, rozmiar pozostaje niezmieniony. Utworzona macierz jest nazywana mapą cech. Ten proces uwydatnia bardziej specyficzne cechy na obrazie tj. zakrzywienia, kształty i zapisuje je we wspomnianej wcześniej mapie. Ustawienie paddingu oraz wartości stride w przypadku warstw konwolucyjnych jest niezwykle istotne, ponieważ wpływa ono bezpośrednio na dane wyjściowe każdej warstwy konwolucyjnej. Działanie paddingu polega na dodawaniu dodatkowych pixeli wokół krawędzi obrazu wejściowego, zwykle wartości 0. Stosowanie paddingu ma na celu zachowanie rozmiaru wyjściowego, po operacji konwolucji, takiego samego jak rozmiar wejściowy. Typ paddingu w bibliotece Keras, możemy określić poprzez zastosowanie ustawienia *valid* (bez paddingu) oraz *same* (z paddingiem). Stride natomiast określa, o ile pikseli filtr przesuwa się po obrazie podczas operacji konwolucji. Domyślna wartość stride wynosi 1, co oznacza, że filtr przesuwa się o jeden piksel na raz. Większe wartości stride powodują większe przeskoki. Dla przykładu ustawienie stride równego 2 powoduje przeskok filtra o 2 piksele, co ma ogromny wpływ na rozmiar wyjściowej mapy cech, ponieważ w takim przypadku zmniejszy się ona dwukrotnie.

Wzór na rozmiar wyjsciowej mapy cech po operacji konwolucji:

$$R = \frac{W - F + 2P}{S} + 1$$

gdzie:

R – rozmiar wyjściowy,

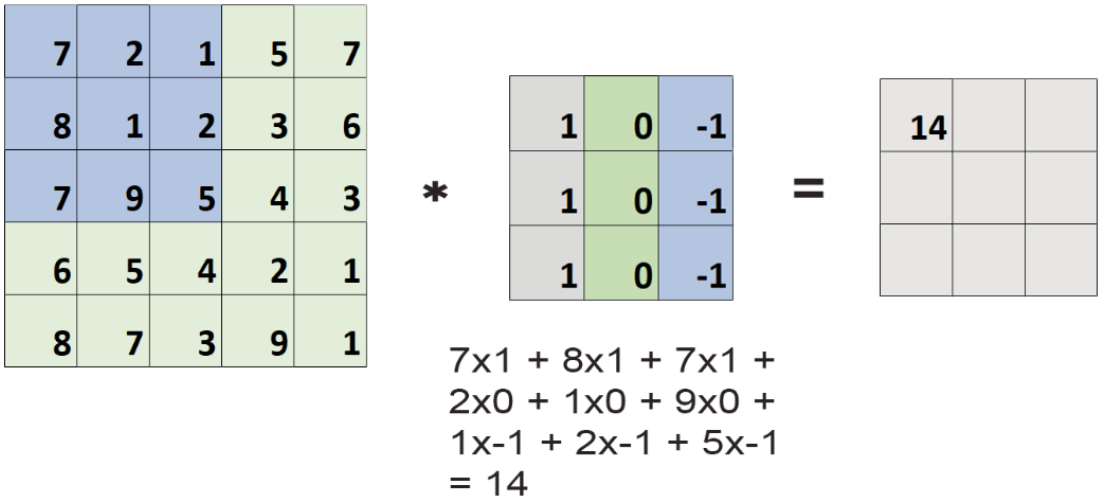
W – rozmiar wejściowy (szerokość lub wysokość),

F – rozmiar filtra,

P – padding,

S – stride

Na rysunku 1 przedstawione zostało powstawanie mapy cech poprzez wykonanie mnożenia macierzy obrazu z kernelem:



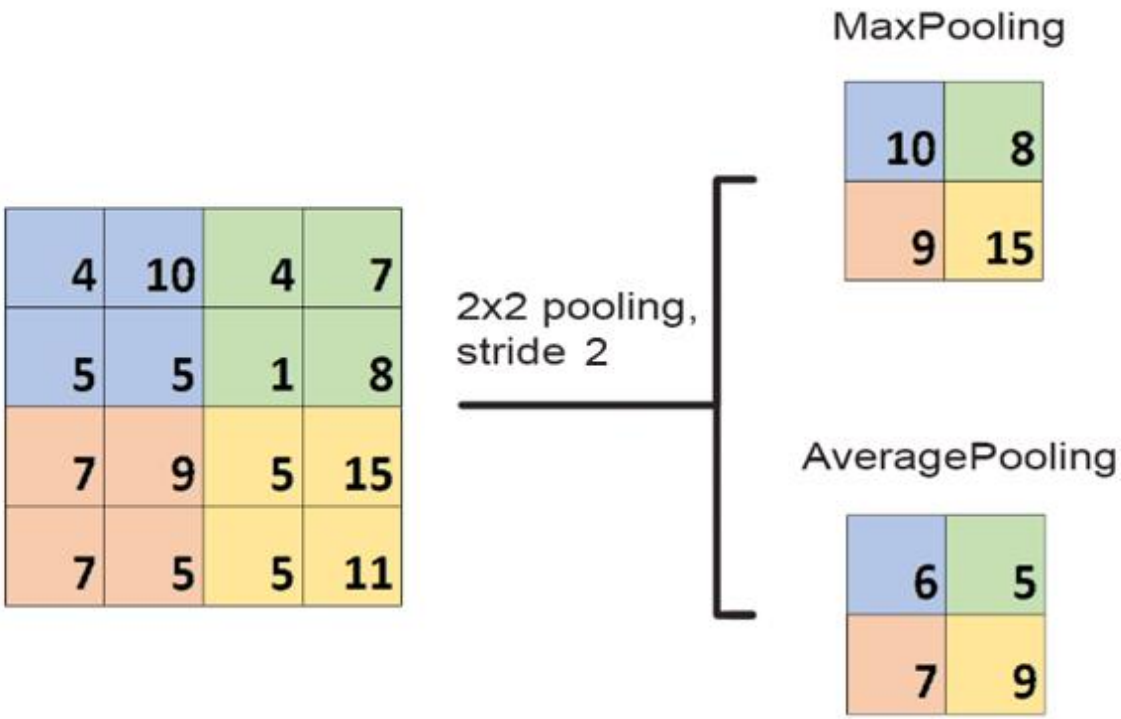
Rysunek 1. Mnożenie macierzy pixeli oraz kernela

Źródło: opracowanie własne

Podczas operacji konwolucji każdy element filtru jest mnożony przez odpowiadający mu element obrazu wejściowego, a następnie wyniki są sumowane, tworząc pojedynczą wartość w wyjściowej mapie cech. Warto zaznaczyć, że podczas tej operacji wagi w filtrze pozostają stałe, gdy przesuwa się on po obrazie. Początkowe wartości wag są losowane, a następnie dostosowywane podczas procesu uczenia sieci ze względu na propagację wsteczną oraz zejście gradientowe. Po każdej operacji konwolucji konieczne jest użycie funkcji aktywacyjnej: *Rectified Linear Unit (ReLU)* , stosowana jest ona w celu przekształcenia mapy cech oraz wprowadzenia nieliniowości. Przygotowana w ten sposób mapa cech jest wyjściem warstwy konwolucyjnej i zostaje podana jako dane wejściowe kolejnej warstwy.

2.2.Warstwa łącząca (ang. Pooling Layer)

Warstwa łącząca jest odpowiedzialna za zmniejszenie rozmiaru mapy cech. Zmniejszają one ilość informacji przy jednoczesnym zachowaniu najwazniejszych cech. Sprawia to, że szkolenie jest mniej kosztowne obliczeniowo i pomaga zapobiegać zbyt nadmiernemu dopasowaniu, w którym model staje się zbyt skoncentrowany na danych szkoleniowych. Wyróżniamy dwa rodzaje warstw łączących, różnią się one działaniem filtra. W przypadku warstwy *Max Pooling* filtr wybiera najwyższą wartość macierzy cech z każdego przesunięcia się filtru po obrazie, natomiast warstwa *Average Pooling*, jak sama nazwa wskazuje, oblicza średnią wartość z wartości znajdujących się w filtrze. Dokładne działanie tej warstwy przedstawione jest na rysunku 2. Widzimy na nim proces obliczania wyjściowej macierzy na podstawie filtru o rozmiarze 2x2 i ustawieniu stride na 2 w celu lepszej wizualizacji działania:



Rysunek 2. Ilustracja przedstawiająca podział oraz działanie warstw łączących

Źródło: opracowanie własne

Powstała w ten sposób macierz jest teraz dwukrotnie mniejsza od mapy wejściowej i jest przygotowana do podania tych danych na kolejną warstwę.

2.3.Warstwa spłaszczająca (ang. Flattening Layer)

Zadaniem tej warstwy jest przekształcenie otrzymanej na wejściu dwuwymiarowej mapy cech, która powstała po operacjach wcześniejszych warstw, na jednowymiarowy wektor. Wykonanie tej operacji jest konieczne do poprawnego działania późniejszych warstw. Po przejściu danych przez tą warstwę, obliczenia macierzowe kończą się, od tego momentu dane będą przekazywane pomiędzy jednowymiarowymi warstami neuronów. W modelu sieci CNN występuje tylko jedna warstwa tego rodzaju

2.4.Warstwa w pełni połączona (ang. Fully connected Layer)

Warstwy w pełni połączone są tradycyjnyni warstwami sieci neuronowych, składają się z neuronów połączonych ze wszystkimi neuronami z poprzedniej warstwy. Jeden model może zawierać wiele takich warstw w zależności od potrzeb badanych danych. Wszystkie z nich mogą posiadać funkcję aktywacyjną ***ReLU***, ważne jest jednak, aby ostatnia zawierała funkcję ***Softmax*** w celu uzyskania rozkładu prawdopodobieństwa na ostateczny zbiór całkowitej liczby klas. Funkcja przekształca surowe wartości wyników z sieci neuronowej na

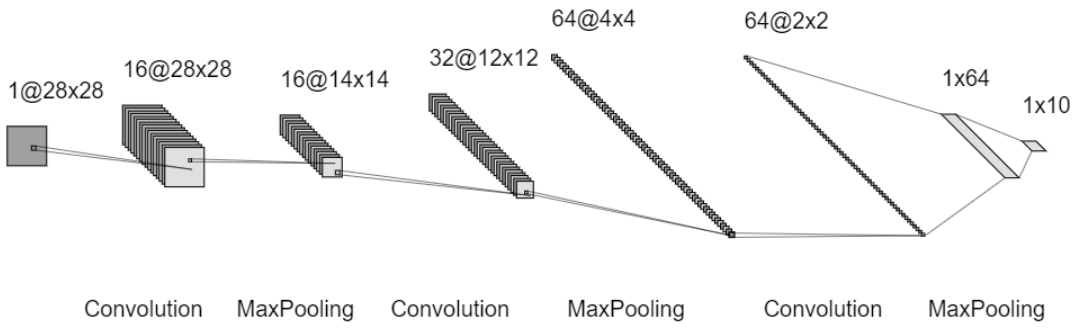
prawdopodobieństwa, które sumują się do 1. Liczba neuronów ostatniej warstwy w pełni połączonej oznacza liczbę klas jakie chcemy uzyskać. W przypadku bazy danych MNIST występują cyfry od 0 do 9, więc potrzebne jest 10 klas / neuronów ostatniej warstwy. Podczas działania modelu, najbardziej aktywny neuron tej ostatniej warstwy będzie oznaczał wynik, czyli liczbę napisaną na obrazie.

3. Przygotowanie bazy danych

Baza danych MNIST zawiera rękopisy przedstawiające cyfry od 0 do 9 w rozmiarze 28x28 pikseli, przekonwertowane do pliku .csv, gdzie każdy wiersz składa się z 785 kolumn opisujących konkretny pixel wartościami od 0 do 255, z wyjątkiem pierwszej kolumny w której jest zapisana ukazana na rysunku cyfra. Baza danych składa się z dwóch plików *mnist_train.csv* zawierający 60000 wierszy z przykładami do trenowania sieci, oraz *mnist_test.csv*, gdzie znajduje się 10000 przykładów służących do testowania sieci. Przed rozpoczęciem nauki modelu należy najpierw przygotować dane, w tym celu konieczne jest dokonanie operacji normalizacji, gdzie wynikiem będą wartości znormalizowane z przedziału od 0 do 1. Zadanie to zostało wykonane przy użyciu dodatkowego skryptu w języku Python, gdzie wykonano normalizacji za pomocą narzędzia *MinMaxScaler* z biblioteki *sklearn,* następnie znormalizowane dane zapisano do plików *mnist_test_normalized.csv* oraz *mnist_train_normalized.csv.* Przygotowane w ten sposób dane są teraz gotowe do dalszych badań.

4. Budowa modelu sieci CNN

Budowa modelu sieci CNN opiera się na języku Python wraz z biblioteką Keras, która zawiera już gotowe funkcję wszystkich poszczególnych warstw. Znormalizowane wcześniej dane zostały wczytane do programu oraz odpowiednio sformatowane i są teraz gotowe do nauki. Budowanie modelu polega na deklaracji modelu sekwencyjnego, a następnie dodawaniu do niego kolejnych warstw. Założono, że budowany model powinien zawierać przynajmniej 3 warstwy konwolucyjne, 3 warstwy MaxPooling oraz 2 warstwy w pełni połączone.



Rysunek 3. Schemat warstw w tworzonym modelu sieci.

Źródło: Opracowanie własne

Rysunek 3 ukazuje schemat poszczególnych warstw w budowanym modelu sieci, widać na nim zmiany rozmiaru tensora po przechodzeniu przez kolejne warstwy. Poniżej opisane zostały kolejne warstwy dodawane do modelu sekwencyjnego:

1. Pierwsza warstwa konwolucyjna - rozmiar macierzy wejściowej wynosi 28x28, w przypadku tego rodzaju warstw należy ustawiać wartości liczby filtrów od najmniejszych, ponieważ obliczenia na tym etapie są wymagające z powodu dużej ilości danych. W związku z tym przyjęto liczbę filtrów równą 16 z rozmiarem 3x3. Przyjęto również wcześniej opisaną funkcję aktywacyjną *ReLU,* dla tej i dla wszystkich kolejnych warstw konwolucyjnych. Wartość *Stride* pozostaje domyślna, czyli wynosi 1. Padding natomiast został ustawiony na *same,* czyli na wyjściu tej warstwy rozmiar mapy cech pozostaje niezmieniony. Wynikiem będzie tensor o rozmiarze 28x28x16.
2. Kolejną warstwą będzie warstwa MaxPooling, została ona wybrana zamiast AveragePooling z powodu większej skuteczności podczas wychwytywania kontrastów obrazu. W przypadku bazy MNIST, gdzie występują jedynie odcienie szarości wybór ten jest bardziej optymalny. Rozmiar filtru tej oraz kolejnych warstw łączących został wyznaczony na 2x2. Po tej warstwie rozmiar macierzy wyjściowej zmniejszy się dwukrotnie, czyli będzie wynosił teraz 14x14x16.
3. Następna będzie warstwa konwolucyjna z liczbą filtrów równą 32, oraz rozmiarem 3x3. Padding zostaje ustawiony na *valid,* co będzie skutkowało zmniejszeniem wyjściowej mapy cech, wynoszącej teraz zgodnie z wcześniej wspomnianym wzorem: 12x12x32.
4. Warstwa MaxPooling, której wynikiem będzie tensor o rozmiarze 6x6x32.
5. Ostatnia warstwa konwolucyjna z liczbą filtrów równą 64 oraz hiperparametrami takimi jak wcześniejsze warstwy z funkcją aktywującą *ReLU*. Wynikiem wyjściowym będzie tensor o rozmiarze 4x4x64.

6. Ostatnia warstwa MaxPooling z rozmiarem filtru 2x2, której wynikiem będzie tensor 2x2x64.
7. Następnie wymagana jest warstwa spłaszczająca otrzymane dane, które będą teraz jednowymiarowym wektorem o rozmiarze 256.
8. Pierwsza warstwa w pełni połączona z liczbą neuronów 64 i funkcją aktywacyjną *ReLU.*
9. Ostatnią warstwą modelu będzie ponownie warstwa w pełni połączona z liczbą neuronów równą 10. Odpowiadają one liczbie klas reprezentujących cyfry w bazie danych. Ta warstwa zawiera funkcję aktywującą *Softmax* w celu poprawnej klasyfikacji prawdopodobieństwa otrzymanego wyniku.

Na listingu 3 została przedstawiona budowa opisanego modelu sekwencyjnego, wykorzystująca wcześniej wspomniane warstwy:

```
model = keras.models.Sequential()

model.add(keras.layers.Conv2D(16, (3, 3), activation='relu',
input_shape=(28, 28, 1), padding='same'))

model.add(keras.layers.MaxPooling2D((2, 2)))

model.add(keras.layers.Conv2D(32, (3, 3), activation='relu'))

model.add(keras.layers.MaxPooling2D((2, 2)))

model.add(keras.layers.Conv2D(64, (3, 3), activation='relu'))

model.add(keras.layers.MaxPooling2D((2, 2)))

model.add(keras.layers.Flatten())

model.add(keras.layers.Dense(64, activation='relu'))

model.add(keras.layers.Dense(10, activation='softmax'))
```

Listing 2. Fragment skryptu tworzącego model – deklaracja modelu sekwencyjnego.

Źródło: Opracowanie własne

Tak zbudowany model należy poddać kompilacji i określić wykorzystywany optymalizator oraz funkcję straty. Porównany został optymalizator *adam (Adaptive Moment Estimation)* oraz SGD *(Stochastic Gradient Descent)*. Optymalizatory wykorzystuje się w celu aktualizacji wag modelu podczas treningu oraz ułatwiają one znalezienie minimalnej wartości funkcji straty. SGD jest prostszy i intuicyjny, natomiast Adam jest znacznie bardziej skomplikowany. Aktualizacja wag w SGD opisana jest wzorem:

$$\theta_{t+1} = \theta_t - \eta \nabla_\theta J(\theta_t)$$

gdzie:

θ_t – to wektory wag w czasie t,

η – to współczynnik uczenia,

$\nabla_\theta J(\theta_t)$ – to gradient funkcji kosztu $J(\theta_t)$ względem wag θ_t

Optymalizator adam jest o wiele bardziej zaawansowanym algorytmem optymalizacji, który łączy zalety AdaGrad i RMSProp. Adam utrzymuje ruchome średnie pierwszego rzędu (średnia gradientów) i drugiego rzędu (średnia kwadratów gradientów), a następnie używa tych średnich do skalowania współczynnika uczenia. Algorytm jest opisany kilkoma wzorami, które odpowiadają za konkretne wartości:

1. Aktualizacja pierwszego momentu (średnia gradientów):

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1)\nabla_0 J(\theta_t)$$

2. Aktualizacja drugiego momentu (średnia kwadratów gradientów):

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2)\left(\nabla_0 J(\theta_t)\right)^2$$

3. Korekta biasu dla pierwszego momentu:

$$\widehat{m}_t = \frac{m_t}{1 - \beta_1^t}$$

4. Koretka biasu dla drugiego momentu:

$$\hat{v}_t = \frac{v_t}{1 - \beta_2^t}$$

5. Aktualizacja wag:

$$\theta_{t+1} = \theta_t - \eta \frac{\widehat{m_t}}{\sqrt{\widehat{v_t}} + \epsilon}$$

gdzie:

θ_t – to wektory wag w czasie t,

η – to współczynnik uczenia, zazwyczaj ustawiony na 0,001,

β_1 – parametr kontrolujący eksponentę pierwszego momentu,

β_2 – parametr kontrolujący eksponentę drugiego momentu,

ϵ – mała liczba zapobiegająca dzieleniu przez zero, zazwyczaj 10^{-8}.

Podczas porównania optymalizatorów została również wykorzystana funkcja straty *sparse_categorical_crossentropy*, która jest stosowana do trenowania modeli klasyfikacyjnych z wieloma klasami (np. w klasyfikacji obrazów). Jest ona szczególnie użyteczna, gdy etykiety klas są podane jako liczby całkowite. Dla modelu klasyfikacyjnego z N próbkami, C klasami, oraz etykietami y_i będącymi liczbami całkowitymi, *sparse_categorical_crossentropy* jest zdefiniowana jako:

$$L = -\frac{1}{N}\sum_{i=1}^{N} \log(p_{y_i})$$

gdzie:

L – wartość funkcji straty,

p_{y_i} – to prawdopodobieństwo przypisane prawidłowej klasie y_i dla próbki i.

Podczas testów liczba epok została określona dla obu przypadków na 20. Najpierw badaniu został poddany algorytm SGD. Przetestowano go dla 4 wartości współczynniku uczenia: 0.1, 0.01, 0.001 oraz 0.15. W przypadku 0.1 sieć CNN uległa przetrenowaniu, w związku z tym należało zmniejszyć współczynnik. Pojęcie trafności modelu opisuje współczynnik ilorazu wszstkich poprawnych predykcji, przez wszystkie predykcje wykonane przez model, wartość ta jest wyrażona w procentach i poszukiwana podczas tego badania. Podsumowując procent trafności uzyskany w tych 4 przypadkach, najwyższą wartość osiągnięto dla 0.15 i wynosiła ona 99%. Optymalizator *adam* automatycznie dostosowuje współczynnik uczenia, co przyspiesza i stabilizuje ten proces, czyniąc go bardziej efektywnym w praktyce. W przypadku 20 epok, dla tego optymalizatora został osiągnięty wynik trafności równy 99,19%.

Podsumowując wynik testu w modelu zostanie zaimplementowany optymalizator *adam,* ponieważ osiągnięty został lepszy wynik trafności. Skompilowany model został poddany trenowaniu na bazie danych MNIST przez 20 epok, ponieważ taka liczba okazała się optymalna pod względem nauki. Po wykonaniu tej operacji wykonano testy na wcześniejszych danych testowych:

```
model.compile(optimizer='adam',
    loss='sparse_categorical_crossentropy',
    metrics=['accuracy'])

history = model.fit(x_train, y_train, epochs=20)

loss, accuracy = model.evaluate(x_test,y_test)
```

Listing 3. Fragment skryptu tworzącego model – kompilacja, trenowanie oraz testowanie modelu.

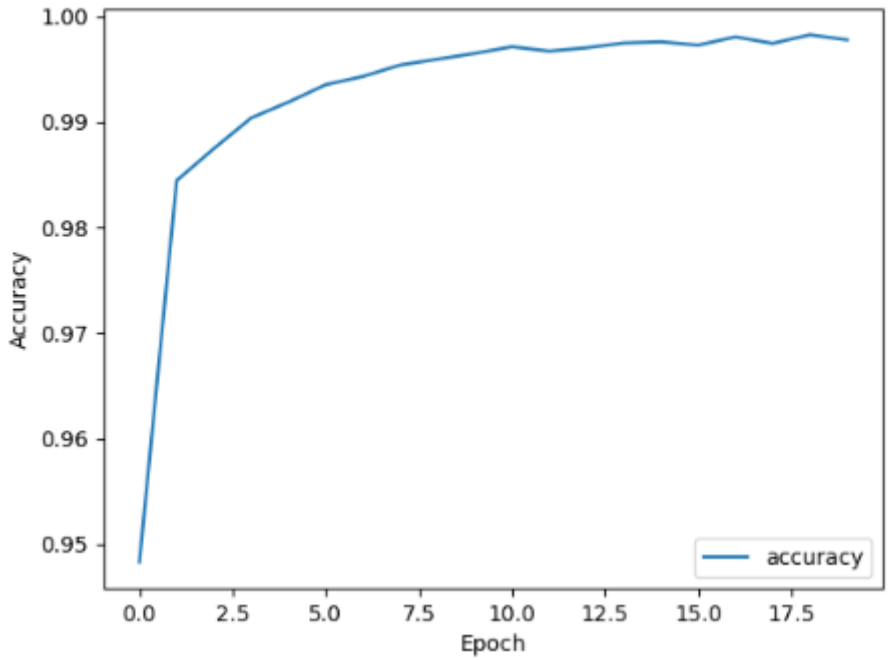
Źródło: Opracowanie własne

Wynik testu został zapisany w zmiennych *loss* oraz *accuracy*, po ich wypisaniu można zobaczyć wynik poprawności modelu, który oscyluje w okolicach ***99,19%,*** natomiast błąd osiągnął ***0,039***:

```
Loss:  0.0399827621877193 45
Accuracy:  0.9919000267982483
```

Rysunek 4. Zdjęcie konsoli wynikowej po uruchomieniu programu nauki modelu sieci CNN.

Źródło: Opracowanie własne

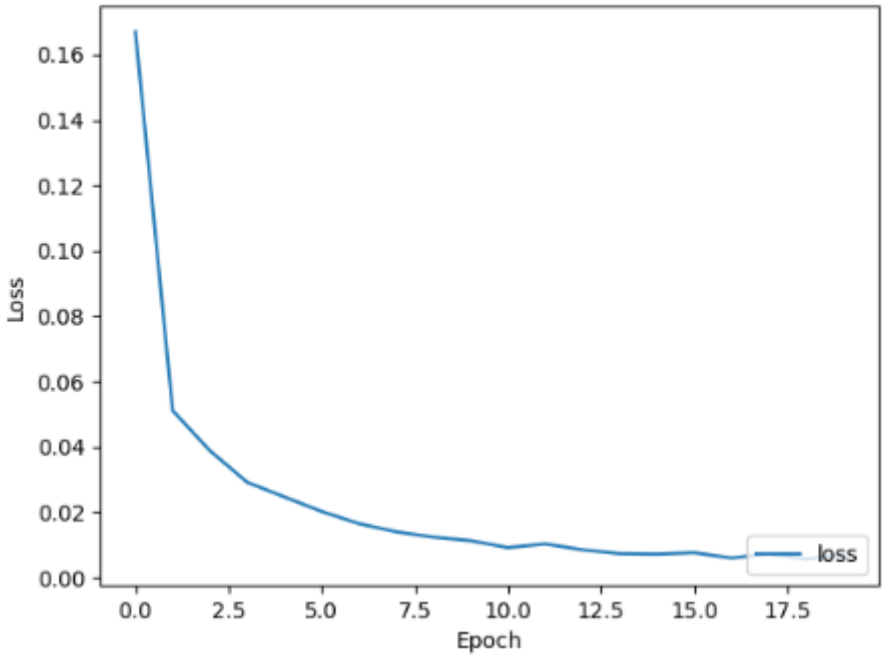


Rysunek 5. Zdjęcie wykresu - procent trafności modelu w kolejnych epokach.

Źródło: Opracowanie własne

Na rysunku 5 widoczny jest wykres trafności modelu podczas jego procesu nauki. Można zauważyć poszczególne wartości w kolejnych epokach, zaczynając od osiągniętych 94.8%

w pierwszej, a kończąc na wcześniej wspomnianych 99.19% na ostatniej. Rysunek 6 opisuje stratę modelu w kolejnych epokach, osiągnęła ona wartość 3,9%:



Rysunek 6. Zdjęcie wykresu - procent straty modelu w kolejnych epokach.

Źródło: Opracowanie własne

Wytrenowany model na koniec programu został zapisany w pliku o nazwie ***cnn.h5*** z rozszerzeniem, które obowiązuje wyłącznie w języku Python z biblioteką Keras, jest to istotne, ponieważ podczas implementowania modelu na stronę internetową konieczna jest konwersja na plik z rozszerzeniem **.json**.

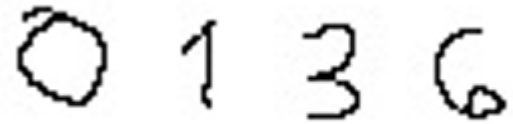
```
model.save('cnn.h5')
```

Listing 4. Fragment skryptu tworzącego model – zapisanie modelu do pliku.

Źródło: Opracowanie własne

4.1. Badanie wydajności modelu

Zanim model zostanie zaimplementowany na stronie internetowej, należy poddać go testom na prawdziwych rysunkach. W tym celu utworzone zostały rysunki w programie Paint, czterech cyfr, odpowiednio: 0, 1, 3, 6, widocznych poniżej (Rysunek 7):



Rysunek 7. Cyfry, na których zostanie przeprowadzony test.

Źródło: Opracowanie własne

Program został zmodyfikowany tak, aby po skompilowaniu modelu wczytywał zdjęcia cyfr, a następnie poddawał je analizie, wypisując wynik. Rysunek poniżej (Rysunek 8) przedstawia otrzymany wynik w kolejności od góry 0, 1, 3, 6:

```
1/1 [==============================] - 0s 90ms/step
Predicted number: 0
1/1 [==============================] - 0s 20ms/step
Predicted number: 1
1/1 [==============================] - 0s 20ms/step
Predicted number: 3
1/1 [==============================] - 0s 20ms/step
Predicted number: 4
```

Rysunek 7. Zdjęcie konsoli z wynikiem badania na własnych cyfrach.

Źródło: Opracowanie własne

Jak można zauważyć ostatni przypadek dla rysunku z cyfrą 6 został zinterpretowany niepoprawnie. Powodem takie analizy może być różnica w piśmie lub położenie w obszarze roboczym cyfry numer 6 z tymi cyframi w bazie danych MNIST, natomiast poprawne rozpoznanie trzech pozostałych cyfr odniosło sukces. Wysuwanie wniosków na temat poprawnego działania modelu na tym etapie jest jeszcze przedwczesne, jednak juz teraz widać problem, który może wystąpić w przyszłym etapie projektu. Utworzony model można zaakceptować, spełnia on wymagania i można przejść do jego implementowania na stronie internetowej.

5. Implementacja modelu na stronie internetowej

Aplikacja webowa zbudowana jest za pomocą biblioteki React wraz z frameworkiem Next.js. Zawiera dodatkowo bibliotekę Tensorflow.js służącą do implementacji modelu oraz

inne dodatkowe biblioteki wymagane do poprawnego działania aplikacji. Zasada działania polega na narysowaniu przez użytkownika cyfry, a następnie wciśnięcia przycisku *„Predict”,* który powoduje przeanalizowanie rysunku poprzez utworzony model oraz zwrócenie na ekran wyniku tej operacji. Po poprawnym przekonwertowaniu modelu na język JavaScript, za pomocą komendy widocznej na listingu 5, można przejść do implementacji.

```
tensorflowjs_converter --input_format keras ./model/cnn.h5
./model/przekonwertowany
```

Listing 5. Komenda konwertująca model z formatu .h5 na format .json.

Źródło: Opracowanie własne

Uruchomienie programu powoduje zbudowanie strony oraz wywołanie jednorazowo hooka *useEffect*, który wczytuje utworzony model z katalogu *‘/public/model/model.json’.* Hook *useEffect* w kontekście biblioteki React służy do wywołania jednorazowo funkcji podczas renderowania strony, która jest wpisana jako argument. Rozwiązanie problemu w taki sposób nie jest zbyt optymalne, ponieważ model jest wczytywany po stronie klienta, a to powoduje problemy z bezpieczeństwem i możliwość wykradzenia modelu. Natomiast na potrzeby samych badań oraz wobec późniejszego udostępniania modelu takie rozwiązanie jest dopuszczalne.

```
useEffect(function () {

    loadModel();

    async function loadModel() {
        const model =
           await tf.loadLayersModel("/model/model.json");

        setModel(model);
    }

}, []);
```

Listing 6. Fragment react hooka wczytującego utworzony model.

Źródło: Opracowanie własne

Działanie powyższego fragmentu kodu (Listing 6) polega na wywołaniu funkcji asynchronicznej wczytującej *model.json* i zapisanie go do zmiennej stanu hooka *useState.* Tak przygotowany model jest juz gotowy do analizowania przekazanych do niego obrazów. Przycisk *„Predict”* powoduje wywołanie asynchronicznej funkcji *handlePredict,* która wczytuje narysowany obraz oraz poddaje go analizie.

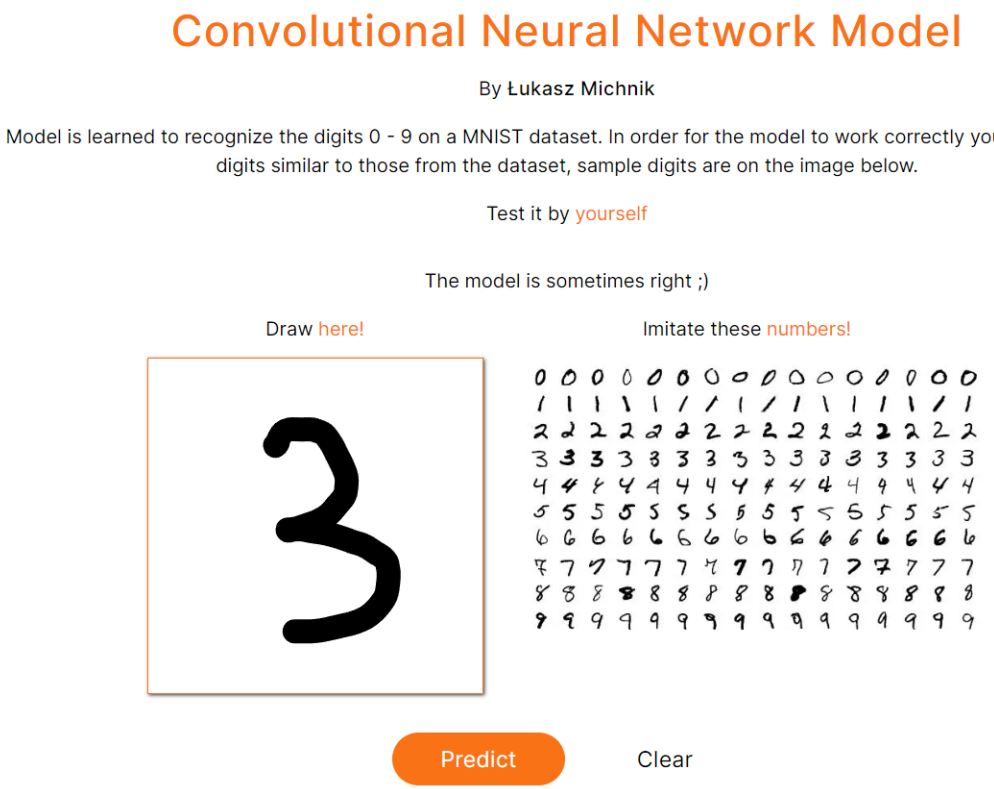
```
async function handlePredict() {
    const canvas = canvasRef.current;

    if (canvas && model) {
        const dataUrl = await canvas.exportImage("jpeg");
        const image = new Image();

        image.src = dataUrl;
        await image.decode();

        let tensor = tf.browser
            .fromPixels(image)
            .mean(2)
            .expandDims(0)
            .expandDims(-1)
            .resizeNearestNeighbor([28, 28])
            .mul(tf.scalar(-1))
            .add(tf.scalar(255));

         const prediction = model.predict(tensor);
         const predictionArray = await prediction.array();
  8      const predictedDigitIndex = predictionArray[0]
            .indexOf( Math.max(...predictionArray[0]) );

         setResult(predictedDigitIndex);
    }
}
```

Listing 7. Funkcja asynchroniczna powodująca predykcję modelu na podstawie podanego obrazu.

Źródło: Opracowanie własne

Pierwsza część funkcji wczytuje rysunek do zmiennej *dataUrl*, ustawia go jako źródło obiektu *Image*, a następnie dekoduje, czyli optymalizuje do późniejszej operacji. Dalej tworzona jest zmienna przechowująca tensor, za pomocą *let tensor = tf.browser.fromPixel(image)* wczytywany jest do niego rysunek cyfry oraz dokonywane są na nim operacje przystosowujące tensor do poprawnego przekazania do modelu. Tensor jest konwertowany na skale szarości, następnie dodawane są nowe wymiary oraz rozmiar obrazu jest przekształcony na tablice dwuwymiarową w rozmiarze 28x28. Na koniec otrzymana macierz jest normalizowana w celu ułatwienia przeanalizowania rysunku przez model. Tak przystosowana zmienna jest poddawana predykcji za pomocą *model.predict(tensor)*, w wyniku tej operacji otrzymany został tensor wynikowy, który należy poddać konwersji na tablicę o rozmiarze 10, gdzie indeks każdego elementu oznacza neuron w ostatniej warstwie sieci CNN. Na koniec w wyniku otrzymania tablicy neuronów należy wyznaczyć najbardziej aktywny z nich, czyli wybrać indeks pod którym znajduje się najwyższa wartość. Indeks oznacza cyfrę napisaną przez użytkownika, ustawiana jest ona jako zmienna stanu, a następnie wyświetlana na ekranie. Na rysunku poniżej

(Rysunek 8) przedstawiony jest widok interfejsu utworzonej aplikacji, wraz z narysowaną cyfrą 3, która została przeanalizowana przez model, a ten zwrócił poprawny wynik:

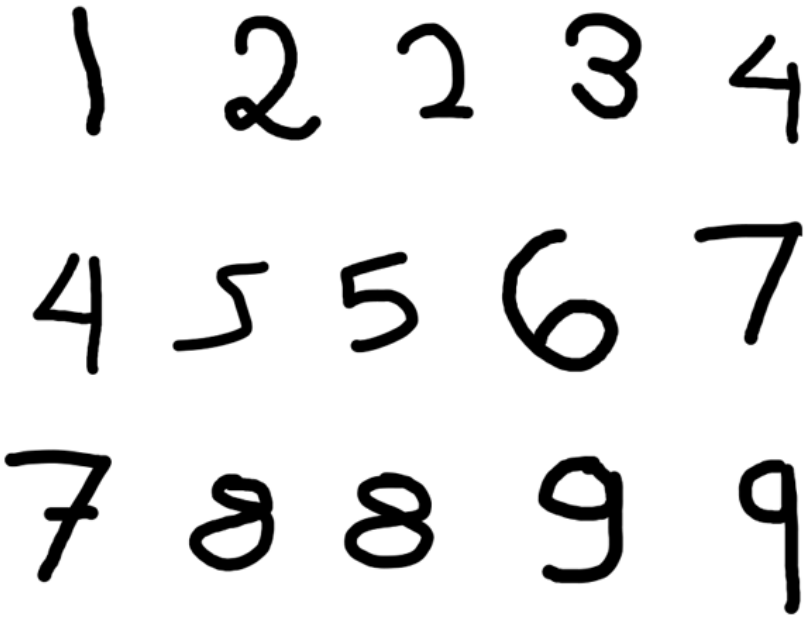


Rysunek 8. Interfejs utworzonej aplikacji.

Źródło: Opracowanie własne

6. Testowanie aplikacji

Ten rozdział artykułu poświęcony jest testowaniu utworzonej aplikacji, test będzie polegał na rysowaniu poszczególnych cyfr oraz analizowaniu poprawności wyników zwracanych przez zaimplementowany model sieci CNN. Model otrzyma po 5 różnych rysunków każdej z cyfr, od 0 do 9, a następnie zwrócone wyniki będą zapisywane i poddane dalszej analizie. Podczas rysowania cyfr, niektóre z nich będą specjalnie zniekształcone lub przesunięte w przestrzeni w celu utrudnienia predykcji.



Rysunek 9. Przykładowe cyfry rysowane w przestrzeni aplikacji.

Źródło: Opracowanie własne

Poprawność utworzonego modelu wypadła dobrze, wystąpiły nieliczne błędy, natomiast jest to spowodowane pismem z jakim użytkownicy mogą pisać cyfry. Najgorzej wypadła cyfra 7, która nagminne jest mylona z 2 oraz 3 w przypadku pisania tej cyfry z charakterystyczną kreską pośrodku. Model najlepiej analizuje cyfry 2, 3 oraz 5. Podczas testów można zauważyć zależność pomiędzy położeniem cyfry w przestrzeni, dla przykładu cyfry pisane po przesunięciu, praktycznie ani razu nie były zinterpretowane poprawnie. Pisanie cyfr w sposób inny niż te w bazie MNIST również mylą model, który oczekuje charakterystycznie pisanych cyfr, dla przykładu cyfry 1, 4, 7, oraz 9 mogą być pisane na różne sposoby, a pisanie ich w sposób inny niż jest nauczony model nie wypada poprawnie.

Podsumowując testy, położenie cyfry w przestrzeni roboczej oraz charakter pisma, mają ogromne znaczenie na wynik sieci CNN. Pisanie zniekształconych cyfr również jest ciężkim wyzwaniem dla modelu, chociaż czasem rysunek zostanie zinterpretowany poprawnie.

7. Podsumowanie

Convolutional Neural Network jest powszechnie wykorzystywana w różnych dziedzinach wymagających analizy obrazów, ich skuteczność w automatycznym wyodrębnianiu cech i klasyfikacji sprawia, że jest niezastąpniona. Projekt opisany w tym artykule dostarczył dogłębnego zrozumienia mechanizów działania sieci konwolucyjnych, poprzez analizę, opis warstw, budowę własnego modelu oraz zaimplementowanie go w aplikacji webowej, w celu

żywego przykładu wykorzystania sieci. Budowa własnego modelu sieci CNN przy użyciu języka Python i biblioteki Keras pozwoliła wykorzystać wcześniej zdobytą wiedzę teoretyczną w praktyce. Tworzenie modelu w opisanej bibliotece jest znacznie uproszczone, ponieważ zawiera ona juz gotowe funkcję poszczególnych warstw. Budowa polega głównie na zadeklarowaniu jego sekwencji i dodaniu warstw w odpowiedniej kolejności oraz dobraniu właściwości, które pomogą osiągnąć jak największą skuteczność. Model opisany w artykule okazał się bardzo skuteczny osiągając 99,19% trafności, co jest wynikiem zadowalającym. Głównym celem utworzenia aplikacji webowej było umożliwienie każdemu użytkownikowi własnoręcznego przetestowania modelu z użyciem stworzonych przez niego rysunków. Działanie utworzonej aplikacji można ocenić jako dobre, nie jest ona idealna z powodu zbyt wielu możliwości pisania cyfr przez użytkowników. Każdy człowiek posiada inny charakter pisma oraz niektóre cyfry mogą być pisane na różne sposoby, również położenie cyfry na przestrzeni roboczej ma ogromne znaczenie. Takie działania i tak wiele możliwości powoduje, że zaimplementowany model nie jest w stanie poprawnie analizować wszystkich rysunków i czasami zwraca błędne wyniki. Najgorzej wypada cyfra 7, która jest dla niego najbardziej problematyczna z powodu podobieństwa do cyfry 2, natomiast pisanie jej z charakterystyczną kreską pośrodku, praktycznie za każdym razem powoduje odbieranie rysunku jako cyfrę 3. Ten problem jest spowodowany bazą MNIST na której model był trenowany. Nie jest ona prawdopodobnie przystosowana do takiego zadania, które zostało dla niej utworzone. Rozwiązaniem poprawności modelu mogłoby być powiększenie wspomnianej bazy danych o kolejne wartości cyfr, które są zniekształcone, pisane na różne sposoby lub przesunięte w przestrzeni. Należy również wspomnieć, że w przypadku pisania prostych cyfr, które są równe i wyśrodkowane oraz nie mają na celu zmylić model, praktycznie zawsze zwracany jest poprawny wynik, więc zamierzony cel został osiągnięty. Utworzona aplikacja internetowa w przyszłości zostanie udostępniona dla wszystkich użytkowników w ramach większego projektu działalności koła GEST, którego opisany w tym artykule projekt jest jedynie częścią.

Literatura

1. Pradeep P., *Practical Convolutional Neural Networks Models,* Packt Publishing, Luty 2018.

Źródła internetowe

1. https://keras.io/api/ (dostęp: 05.06.2024).
2. https://medium.com/latinxinai/convolutional-neural-network-from-scratch-6b1c856e1c07 (dostęp: 05.06.2024).
3. https://github.com/JudiJudi6/Ai-handwritten-digits-classification (dostęp: 05.06.2024).
4. https://ai-handwritten-digits-classification.vercel.app/ (dostęp: 05.06.2024).
5. https://www.datacamp.com/tutorial/cnn-tensorflow-python (dostęp: 05.06.2024).

KOŁO

NAUKOWE

INFORMATYKÓW

„KOD”

Adam Krawczyk, Jakub Jucha, Hubert Kraus, Sebastian Cwynar, Maciej Karczmarz
SKNI „Kod”

dr inż. Bartosz Trybus
Opiekun Koła Naukowego

Zastosowanie frameworku Next.js do tworzenia stron internetowych i aplikacji webowych

Streszczenie

Celem niniejszego artykułu jest przedstawienie zastosowania frameworku Next.js do tworzenia nowoczesnych, wydajnych i dynamicznych aplikacji webowych. Obiektem badań jest Next.js, oparty na React, który umożliwia tworzenie zarówno prostych, jak i zaawansowanych aplikacji internetowych. Metoda badawcza obejmuje praktyczne kroki od podstawowej instalacji Next.js, przez zrozumienie Server Components i Client Components, aż po techniki routingu. Artykuł opisuje również integrację z API przy użyciu PocketBase oraz demonstrację aplikacji z funkcjami tworzenia i wyświetlania artykułów. Wnioski wskazują, że Next.js, dzięki zaawansowanym funkcjom oraz łatwości użycia, jest idealnym narzędziem do tworzenia nowoczesnych i responsywnych aplikacji internetowych, spełniających oczekiwania użytkowników.

Słowa kluczowe: Next.js, React, framework, Server Components, Client Components, routing, integracja z API, PocketBase, frontend.

1. Wprowadzenie

W obecnym dynamicznym środowisku technologicznym, rośnie zapotrzebowanie na szybkie i efektywne tworzenie oprogramowania. Użytkownicy oczekują coraz bardziej zaawansowanych i responsywnych aplikacji internetowych, co stanowi wyzwanie dla programistów. W poszukiwaniu narzędzi, które przyspieszą i ułatwią proces tworzenia aplikacji, wielu deweloperów zwraca się ku nowoczesnym frameworkom.

Next.js jest jednym z takich narzędzi, oferującym szeroki wachlarz funkcji, które umożliwiają szybkie tworzenie wydajnych i skalowalnych aplikacji internetowych. Dzięki integracji z biblioteką React, Next.js umożliwia korzystanie z zaawansowanych technik renderowania, co pozwala na tworzenie aplikacji spełniających wysokie oczekiwania użytkowników.

W odpowiedzi na rosnące zapotrzebowanie rynkowe, Next.js oferuje szereg rozwiązań, które ułatwiają życie programistom. Automatyczne generowanie routingu, obsługa dynamicznych tras, a także możliwość renderowania po stronie serwera i statyczne generowanie stron to tylko niektóre z funkcji, które przyczyniają się do jego popularności.

W niniejszym artykule przyjrzymy się bliżej, jak Next.js odpowiada na wyzwania współczesnego rynku deweloperskiego, jakie korzyści przynosi jego używanie oraz jak wykorzystać go do stworzenia nowoczesnej strony internetowej. Artykuł ten ma na celu nie tylko przybliżenie możliwości Next.js, ale również dostarczenie praktycznych wskazówek, które pomogą w efektywnym wykorzystaniu tego narzędzia do tworzenia nowoczesnych i wydajnych aplikacji internetowych.

2. Next.js w porównaniu z konkurencyjnymi rozwiązaniami

Framework Next.js, oparty na React, wyróżnia się na tle innych rozwiązań dostępnych na rynku. Konkurencyjne narzędzia, takie jak Gatsby, Nuxt.js (dla Vue) czy Angular Universal, również oferują zaawansowane funkcje do tworzenia nowoczesnych aplikacji internetowych. Niemniej jednak, Next.js ma kilka unikalnych cech i zalet, które sprawiają, że jest wyjątkowym wyborem dla deweloperów.

Jednym z głównych atutów Next.js jest wyjątkowa elastyczność w wyborze technik renderowania. Deweloperzy mogą łatwo przełączać się między statycznym generowaniem stron (SSG), renderowaniem po stronie serwera (SSR) i renderowaniem po stronie klienta (CSR). Taka wszechstronność umożliwia tworzenie aplikacji dostosowanych do specyficznych potrzeb projektu, co nie zawsze jest możliwe w konkurencyjnych frameworkach. Na przykład, Gatsby jest bardziej skoncentrowany na statycznym generowaniu stron, co może ograniczać jego zastosowanie w dynamicznych aplikacjach.

Dzięki integracji z Vercel, Next.js oferuje niezwykle szybki proces wdrażania aplikacji. Wbudowane narzędzia do optymalizacji, takie jak automatyczne dzielenie kodu, prefetching zasobów czy optymalizacja obrazów, znacząco przyspieszają ładowanie stron i poprawiają ogólną wydajność aplikacji. Chociaż inne frameworki również oferują narzędzia optymalizacyjne, Next.js wyróżnia się łatwością ich użycia oraz głęboką integracją z platformą Vercel, co przyspiesza i upraszcza proces wdrażania.

Framework rozwijany przez Vercel cieszy się ogromnym wsparciem społeczności oraz regularnymi aktualizacjami. Dokumentacja Next.js jest bardzo szczegółowa i łatwa do zrozumienia, co znacząco ułatwia naukę i rozwiązywanie problemów. W porównaniu do Nuxt.js czy Angular Universal, społeczność Next.js jest jedną z najbardziej aktywnych, co przekłada się na szybsze rozwiązywanie problemów oraz bogatszą bazę wiedzy i narzędzi.

Jako rozszerzenie React, Next.js czerpie pełne korzyści z ekosystemu React. Deweloperzy mają dostęp do ogromnej liczby komponentów, bibliotek i narzędzi, które mogą być łatwo zintegrowane z aplikacją. W przeciwieństwie do Angular Universal, który ma własny ekosystem, czy Nuxt.js, który opiera się na Vue, Next.js umożliwia deweloperom korzystanie z szerokiej gamy zasobów dostępnych dla React, co zwiększa możliwości i przyspiesza rozwój aplikacji.

Next.js oferuje również wbudowane rozwiązania do zarządzania danymi, takie jak możliwość łatwej integracji z różnymi API czy wsparcie dla funkcji getStaticProps i getServerSideProps. W porównaniu do innych frameworków, te funkcje upraszczają proces pobierania danych i zarządzania stanem aplikacji, co może być bardziej skomplikowane w innych narzędziach, takich jak Gatsby czy Angular Universal.

3. Tworzenie projektu i konfiguracja środowiska

Pierwszym krokiem w tworzeniu aplikacji Next.js jest konfiguracja środowiska programistycznego. Wymaga to zainstalowania Node.js oraz pakietu npm (Node Package Manager). W celu utworzenia projektu wystarczy użyć kreatora, który uruchamia się w wierszu poleceń za pomocą frazy “npx create-next-app@latest”. Sufiks “@latest” stosuje się w celu inicjalizacji procesu tworzenia nowego projektu przy użyciu najnowszej wersji Next.js. W wyniku zastosowania tej komendy, w konsoli pojawi się menu konfiguracji projektu. Poniżej przedstawiono proces tworzenia projektu:

```
PS D:\Javascript_Projects\React\NextJs> npx create-next-app@latest
√ What is your project named? ... example
√ Would you like to use TypeScript? ... No / Yes
√ Would you like to use ESLint? ... No / Yes
√ Would you like to use Tailwind CSS? ... No / Yes
√ Would you like to use `src/` directory? ... No / Yes
√ Would you like to use App Router? (recommended) ... No / Yes
√ Would you like to customize the default import alias (@/*)? ... No / Yes
Creating a new Next.js app in D:\Javascript_Projects\React\NextJs\example.
```

Rysunek 1. Menu konfiguracji projektu
Źródło: opracowanie własne.

Po zakończeniu konfiguracji projekt zostanie utworzony w folderze o przypisanej przez użytkownika nazwie. Teraz może on zostać uruchomiony za pomocą środowiska programistycznego. Domyślnie projekt zawiera szereg gotowych elementów, takich jak pliki konfiguracyjne, demonstracyjny arkusz stylów i szablon strony. Na potrzeby artykułu arkusz stylów i szablon strony zostały usunięte.

4. Routing, nawigacja i struktura folderów

Routing w Next.js jest jednym z kluczowych elementów, które wyróżniają ten framework na tle innych narzędzi do tworzenia aplikacji opartych na React. Next.js automatycznie obsługuje routing na podstawie struktury folderów w projekcie, co znacznie upraszcza proces tworzenia i zarządzania trasami w aplikacji.

Nowy system routingowy wprowadzony w Next.js 13 wykorzystuje strukturę folderów src/app do automatycznego generowania tras. Każdy folder i plik w src/app jest mapowany na unikalny URL. Na przykład, mając poniższą strukturę folderów:

Rysunek 2. Struktura folderów projektu uzyskana przy pomocy komendy “tree” w windows powershell
Źródło: opracowanie własne.

- “src/app/page.tsx” będzie dostępny pod adresem URL “/”,
- “src/app/articles/page.tsx” będzie dostępny pod adresem /articles,
- “src/app/articles/[id]/page.tsx” będzie dostępny pod dynamicznym adresem “/articles/[id]”, gdzie “[id]” jest dynamicznym segmentem URL.

Nazwy plików w strukturze Next.js nie są przypadkowe, gdyż pełnią specyficzne funkcje w kontekście routingu i struktury aplikacji, na przykład:

1. Pliki o nazwie “page.tsx” definiują nowe strony w aplikacji. Są one automatycznie mapowane na odpowiednie URL-e na podstawie ich lokalizacji w strukturze folderów. Na przykład “src/app/about/page.tsx” będzie dostępny pod adresem “/about”.
2. Pliki “layout.tsx” definiują układ strony, który może być współdzielony między różnymi podstronami. Umożliwia to tworzenie wspólnych elementów takich jak nagłówki, stopki, czy paski boczne, które pojawiają się na wielu stronach. Na przykład, src/app/layout.tsx może zawierać wspólną nawigację dla całej aplikacji.
3. Plik “error.tsx” jest odpowiedzialny za renderowanie strony błędów. Jest to strona, która jest wyświetlana w przypadku napotkania błędu podczas renderowania strony lub podczas wykonywania żądania. Można go dostosować, aby wyświetlać komunikaty o błędach i instrukcje dla użytkowników.
4. Plik “loading.tsx” jest używany do renderowania stanu ładowania. Jest to komponent, który jest wyświetlany podczas ładowania danych lub generowania strony. Dzięki temu użytkownicy widzą interfejs ładowania zamiast pustej strony, co poprawia doświadczenie użytkownika.
5. Plik middleware.tsx umożliwia dodawanie globalnych funkcji middleware, które mogą przetwarzać żądania HTTP przed ich dotarciem do odpowiednich tras. Middleware może być używane do autoryzacji użytkowników, logowania aktywności, ustawiania nagłówków odpowiedzi i wielu innych zadań.

Do nawigacji między stronami w Next.js używamy komponentu Link. Umożliwia on tworzenie wewnętrznych linków w aplikacji, które korzystają z funkcji bezprzeładowaniowej nawigacji. Dzięki temu, kiedy użytkownik klika w link, aplikacja dynamicznie zmienia stronę bez pełnego odświeżania przeglądarki, co znacząco poprawia wydajność i doświadczenie użytkownika. Najprostszym sposobem użycia komponentu Link jest importowanie go z biblioteki next/link i osadzenie go w aplikacji w miejscu, gdzie chcemy utworzyć link do innej strony.

```
import Link from "next/link";

export default function Header() {
  return (
    <div>
      <Link href="/">
        <span className="text-white hover:text-gray-300">Home</span>
      </Link>
      <Link href="/articles">
        <span className="text-white hover:text-gray-300">Articles</span>
      </Link>
    </div>
  );
}
```

Rysunek 3. Przykład użycia komponentu Link
Źródło: opracowanie własne.

W tym przykładzie komponent Link jest używany do stworzenia dwóch linków: jeden prowadzi do strony "Home" (/), a drugi do strony "Articles" (/articles). Gdy użytkownik kliknie na jeden z tych linków, Next.js przeprowadzi nawigację bez pełnego przeładowania strony. Atrybut “href” jest wymagany w komponencie Link i określa adres URL, do którego użytkownik zostanie przekierowany. Może to być względny lub bezwzględny URL. Możemy również używać dynamicznych wartości w href, co jest szczególnie przydatne w przypadku dynamicznych tras. Next.js automatycznie włącza prefetching linków, które są widoczne w widoku przeglądarki. Oznacza to, że Next.js pobiera dane dla strony docelowej, zanim użytkownik na nią kliknie, co jeszcze bardziej przyspiesza nawigację. Możemy również ręcznie włączyć lub wyłączyć prefetching za pomocą atrybutu “prefetch”.

5. Server Components i Client Components

Next.js wprowadza nowy sposób budowania aplikacji internetowych za pomocą komponentów serwerowych (Server Components) i komponentów klienckich (Client Components). Ta sekcja wyjaśni, czym są te komponenty, jakie mają zastosowania oraz jakie korzyści przynoszą programistom.

Server Components to komponenty React, które są renderowane na serwerze. Dzięki temu podejściu można zredukować ilość JavaScriptu przesyłanego do klienta, co prowadzi do szybszego ładowania strony i lepszej wydajności aplikacji.

Zalety Server Components:

- Mniejsze obciążenie po stronie klienta: Ponieważ większość pracy związanej z renderowaniem odbywa się na serwerze, klient otrzymuje w pełni wyrenderowany HTML, co zmniejsza czas potrzebny na ładowanie i renderowanie strony,
- Lepsze SEO: Ponieważ zawartość jest renderowana na serwerze, boty wyszukiwarek mogą łatwiej indeksować strony, co może poprawić pozycję strony w wynikach wyszukiwania,
- Bezpieczeństwo: Server Components mogą wykonywać bezpieczne operacje, takie jak dostęp do baz danych i przetwarzanie poufnych danych, bez konieczności przesyłania tych danych do klienta.

Client Components (Komponenty Klienckie) to komponenty React, które są renderowane po stronie klienta. Oznacza to, że przeglądarka użytkownika musi wykonywać kod, aby wygenerować zawartość strony. Te komponenty są niezbędne, gdy wymagane są interaktywne funkcje, które reagują na działania użytkownika, takie jak klikanie przycisków, wprowadzanie danych do formularzy itp.

Zalety Client Components:

- Interaktywność: Umożliwiają tworzenie interaktywnych aplikacji, które mogą reagować na akcje użytkownika bez konieczności odświeżania strony,
- Stan lokalny: Mogą przechowywać i zarządzać stanem lokalnym bez konieczności komunikowania się z serwerem.

Zastosowanie Client Components i Server Components wpływa na sposób, w jaki aplikacja zarządza zasobami, w szczególności pamięcią. Przyjrzymy się różnicom w zużyciu pamięci przez oba rodzaje komponentów na przykładzie renderowania kodu z użyciem biblioteki “react-syntax-highlighter".

```
import SyntaxHighlighter from "react-syntax-highlighter";
export default function Home() {
  return (
    <main>
      <SyntaxHighlighter>{`print("Hello world")`}</SyntaxHighlighter>
    </main>
  );
}
```

Rysunek 4. Przykład komponentu serwerowego
Źródło: opracowanie własne.

Zużycie pamięci dla tego komponentu serwerowego wynosi 6,5 MB. Ponieważ renderowanie odbywa się na serwerze, klient nie musi ładować pełnych zależności i bibliotek, a przeglądarka użytkownika otrzymuje już gotowy HTML. To eliminuje potrzebę przetwarzania JavaScriptu po stronie klienta, co znacznie zmniejsza zużycie pamięci w przeglądarce.

```
"use client";
import SyntaxHighlighter from "react-syntax-highlighter";
export default function Home() {
  return (
    <main>
      <SyntaxHighlighter>{`print("Hello world")`}</SyntaxHighlighter>
    </main>
  );
}
```

Rysunek 5. Przykład komponentu klienckiego
Źródło: opracowanie własne.

W przypadku powyższego komponentu klienckiego zawierającego taki sam kod jak poprzedni komponent, zużycie pamięci wyniosło 10,8 MB. Wyższe zużycie pamięci wynika z tego, że wszystkie operacje renderowania muszą być przetworzone po stronie klienta. Ponadto w przypadku bibliotek, wszystkie zależności muszą być załadowane i przechowywane w pamięci przeglądarki.

Wybór między Server Components a Client Components zależy od specyfiki aplikacji i jej wymagań. Komponenty serwerowe mogą być bardziej wydajne pod względem zużycia pamięci i szybkości ładowania, podczas gdy komponenty klienckie są niezbędne do tworzenia dynamicznych i interaktywnych interfejsów użytkownika.

6. Integracja z API i przykładowa aplikacja

Integracja z API (Application Programming Interface) jest kluczowym elementem każdej nowoczesnej aplikacji internetowej, umożliwiającym komunikację z serwerem oraz dostęp do danych i funkcjonalności. W Next.js zarządzanie żądaniami HTTP jest uproszczone dzięki wbudowanym funkcjom, które ułatwiają pracę z API. W tej sekcji omówimy, jak zintegrowano aplikację Next.js z API, na przykładzie projektu do zarządzania artykułami.

Środowisko Next.js zostało skonfigurowane zgodnie z opisem we wcześniejszych sekcjach artykułu. Głównym celem było dodanie funkcji pobierania, wyświetlania i tworzenia artykułów z wykorzystaniem API. Backend został stworzony przy pomocy PocketBase – lekkiego backendu opartego na Go, który oferuje funkcjonalności takie jak bazy danych, uwierzytelnianie użytkowników, oraz obsługę plików. PocketBase umożliwił szybkie i efektywne stworzenie aplikacji dzięki łatwej konfiguracji i intuicyjnemu interfejsowi.

PocketBase został skonfigurowany do zarządzania kolekcją artykułów. Utworzono kolekcję o nazwie "articles", która przechowuje dane artykułów takie jak tytuł, opis, treść oraz data utworzenia. Kolekcja ta jest dostępna poprzez interfejs RESTful API, co umożliwia komunikację z frontendem aplikacji Next.js.

Na potrzeby tego projektu utworzono strukturę plików jak na rysunku 2 oraz następujące komponenty:

- Header: nawigacja po stronie,
- Article: wyświetlanie pojedynczego artykułu,
- ArticlesPage: lista artykułów,
- CreateArticle: formularz do tworzenia artykułów.

Komponent Header (widoczny na rysunku 3) został zaprojektowany jako nawigacja strony, umożliwiając użytkownikom łatwe przechodzenie między stroną główną a stroną zawierającą artykuły. Został zamieszczony w pliku layout.tsx, co gwarantuje jego stałą obecność na górze strony, niezależnie od aktualnie wyświetlanej podstrony. Komponent został dodatkowo dostosowany za pomocą stylów CSS, co zapewnia spójny wygląd i styl nawigacji.

Strona powitalna zawiera prosty kod HTML z krótkim opisem strony, przywitanie użytkowników oraz przedstawienie celu aplikacji. Kolejna strona jest odpowiedzialna za komunikację z API PocketBase, pobieranie listy artykułów oraz ich wyświetlanie w formie siatki.

Na poniższym rysunku widoczny jest kod odpowiadający za funkcję, która wykonuje żądanie do API PocketBase w celu pobrania listy artykułów. Funkcja korzysta z wbudowanej funkcji fetch Next.js, która umożliwia wykonanie żądania HTTP. Funkcja wysyła żądanie GET do API PocketBase, pobiera dane artykułów i zwraca je jako tablicę obiektów. Opcja cache: "no-cache" zapewnia, że zawsze pobierane są najnowsze dane.

```
async function getArticles() {
  const res = await fetch(
    `http://127.0.0.1:8090/api/collections/Articles/records`,
    { cache: "no-cache" }
  );
  const data = await res.json();
  return data?.items as any[];
}
```

Rysunek 6. funkcja pobierająca dane z API
Źródło: opracowanie własne.

Na poniższym rysunku widoczny jest kod odpowiadający za komponent, który jest odpowiedzialny za renderowanie listy artykułów. Wykorzystuje on funkcję getArticles do pobrania artykułów i wyświetla je w formie siatki. Komponent używa pętli map do iteracji przez listę artykułów i wyrenderowania każdego artykułu w osobnym kontenerze. Każdy kontener zawiera komponent wyświetlający szczegóły artykułu.

```
export default async function ArticlesPage() {
  const articles = await getArticles();
  return (
    <div>
      <div className=" grid grid-cols-3">
        {articles.map((article) => {
          return (
            <div className="border-4 border-gray-600 m-5 h-64 bg-gray-300">
              <Article key={article.id} article={article}></Article>
            </div>
          );
        })}
      </div>
      <CreateArticle></CreateArticle>
    </div>
  );
}

function Article({ article }: any) {
  const { id, title, description, created } = article || {};
  const formattedDate = new Date(created).toLocaleDateString("en-US", {
    year: "numeric",
    month: "long",
    day: "numeric",
  });
  return (
    <Link href={`/articles/${id}`}>
      <div className="flex flex-col justify-between h-full">
        <div>
          <h2 className="flex justify-center text-2xl">{title}</h2>
          <h4>Description: {description}</h4>
        </div>
        <p className="flex justify-start">Creation date: {formattedDate}</p>
      </div>
    </Link>
  );}
```

Rysunek 7. fragment kodu odpowiadający za wyświetlanie artykułów na stronie
Źródło: opracowanie własne.

Kolejna strona, której kod widoczny jest na poniższym rysunku, zawiera kod odpowiedzialny za wyświetlanie szczegółów pojedynczego artykułu na podstawie jego ID. Widoczna jest funkcja, która pobiera dane pojedynczego artykułu z API PocketBase. Funkcja ta wysyła żądanie GET do API PocketBase z podanym ID artykułu, pobiera dane artykułu i zwraca je.

```tsx
// src\app\articles\[id]\page.tsx
async function getArticle(articleID: string) {
  const res = await fetch(
    `http://127.0.0.1:8090/api/collections/Articles/records/${articleID}`,
    {
      cache: "no-cache",
    }
  );
  const data = await res.json();
  return data;
}

export default async function Article({ params }: any) {
  const article = await getArticle(params.id);
  return (
    <div className="flex justify-center">
      <div className="flex flex-col justify-between h-full border-4 border-gray-400 bg-gray-300 w-3/6 ">
        <div>
          <h1 className="flex justify-center text-3xl">{article.title}</h1>
          <h4>{article.text}</h4>
        </div>
        <p className="flex justify-start">{article.created}</p>
      </div>
    </div>
  );
}
```

Rysunek 8. Fragment kodu odpowiadający za wyświetlenie artykułu na podstronie
Źródło: opracowanie własne.

Na poniższym rysunku widoczny jest kod odpowiadający za komponent będący formularzem umożliwiającym użytkownikom dodawanie nowych artykułów. Komponent używa hooków stanu useState do zarządzania wartościami wprowadzonymi przez użytkownika. Formularz zawiera pola tekstowe do wprowadzania tytułu, opisu i treści artykułu. Po przesłaniu formularza dane są wysyłane do API PocketBase za pomocą żądania POST, co skutkuje utworzeniem nowego artykułu w bazie danych. Po pomyślnym dodaniu artykułu pola formularza są resetowane.

```tsx
// src\app\components\article-form.tsx
"use client";

import { useState } from "react";

export default function CreateArticle() {
  const [title, setTitle] = useState("");
  const [description, setDescription] = useState("");
  const [text, setText] = useState("");

  const add_article = async () => {
    await fetch("http://127.0.0.1:8090/api/collections/Articles/records", {
      method: "POST",
      headers: {
        "Content-Type": "application/json",
      },
      body: JSON.stringify({
        title,
        description,
        text,
      }),
    });
    setTitle("");
    setDescription("");
    setText("");
  };
  return (···
  );
}
```

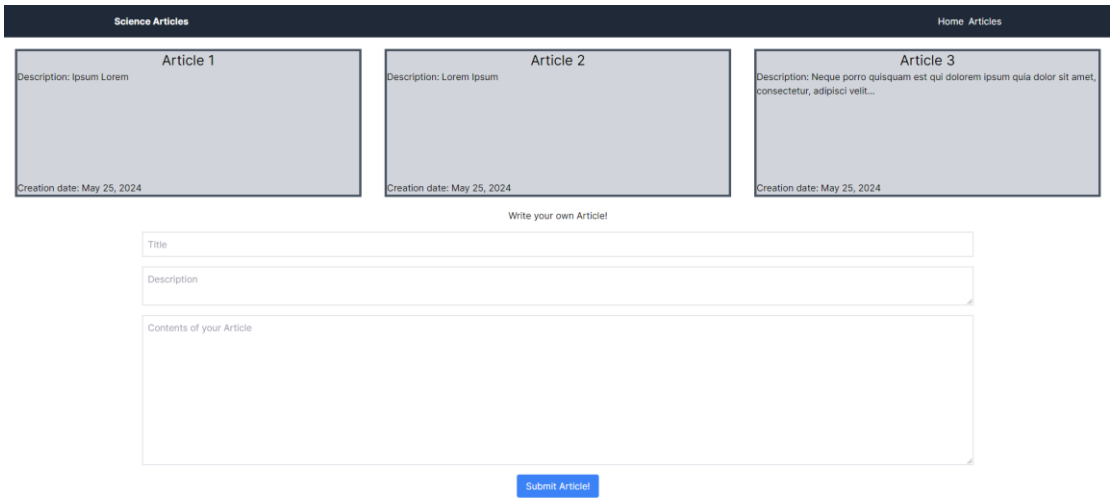
Rysunek 9. Fragment kodu odpowiadający za dodanie artykułu do bazy danych
Źródło: opracowanie własne.

Integracja Next.js z PocketBase umożliwia zarządzanie artykułami w aplikacji. Wykorzystując wbudowane funkcje Next.js oraz intuicyjne API PocketBase, proces tworzenia, pobierania i wyświetlania artykułów staje się prosty i efektywny. Struktura komponentów oraz

logika komunikacji z API zapewniają elastyczność i skalowalność aplikacji, co jest kluczowe w nowoczesnym tworzeniu aplikacji internetowych.

Warto również zwrócić uwagę na asynchroniczność w zarządzaniu żądaniami HTTP. Funkcje asynchroniczne umożliwiają wykonywanie operacji sieciowych w tle bez blokowania głównego wątku aplikacji, co jest szczególnie ważne dla zachowania płynności interfejsu użytkownika. W Next.js funkcja fetch umożliwia wysyłanie żądań HTTP i obsługę odpowiedzi, co jest kluczowe w integracji z zewnętrznymi API. Asynchroniczność i fetch umożliwiają efektywną i bezproblemową komunikację z serwerem, co przekłada się na lepsze doświadczenia użytkowników końcowych.

W ten oto sposób utworzona została prosta aplikacja webowa przy użyciu Next.js. Aplikacja demonstruje podstawowe funkcje frameworku, takie jak routing, renderowanie komponentów serwerowych i klienckich, oraz integrację z zewnętrznym API. Dzięki elastyczności Next.js możliwe było stworzenie nowoczesnej i responsywnej strony, która pobiera i wyświetla artykuły oraz pozwala na ich dodawanie za pomocą prostego formularza. To pokazuje, jak Next.js może być wykorzystany do szybkiego i efektywnego tworzenia aplikacji internetowych, które spełniają współczesne standardy i oczekiwania użytkowników.



Rysunek 10. Wygląd fragmentu gotowej aplikacji
Źródło: opracowanie własne.

7. Podsumowanie

Next.js jest frameworkiem umożliwiającym proste i szybkie tworzenie stron internetowych. Dzięki swoim zaawansowanym funkcjom, takim jak wbudowane routingowanie, obsługa SSR i SSG, oraz integracja z React, Next.js staje się coraz popularniejszym wyborem wśród deweloperów. Jego elastyczność i wydajność sprawiają, że jest idealnym narzędziem do tworzenia nowoczesnych, responsywnych aplikacji internetowych. Wraz z nowymi funkcjami, takimi jak Server Components i Client Components, Next.js pozostaje na czele innowacji w świecie front-end developmentu, umożliwiając deweloperom tworzenie aplikacji, które spełniają oczekiwania użytkowników.

Literatura

Michele Riva, *Real-World Next.js*, Packt Publishing, 2022

Źródła internetowe

https://nextjs.org/docs (dostęp: 19.05.2024).
https://vercel.com/blog (dostęp: 19.05.2024).
https://pocketbase.io/docs (dostęp:20.05.2024)
https://github.com/Kemnaz/ArticleWebPage (dostęp: 01.06.2024)

Aleksandra Rokita, Katarzyna Maternia, Magdalena Matuła, Aleksandra Sawicka, Łukasz Książek
SKNI „Kod”

dr inż. Bartosz Trybus
Opiekun Koła Naukowego

Algorytmy i artyści – rola sztucznej inteligencji w sztuce i muzyce

Streszczenie

Sztuczna inteligencja od lat rozwija się i ewoluuje. Widząc na ulicy plakat lub reklamę, możemy nie zdawać sobie sprawy, że wykonało je właśnie AI. Dziś obrazy i instalacje artystyczne wykonane przez sztuczną inteligencję możemy spotkać nawet w muzeach i na wystawach sztuki. Algorytm uczy się tworzyć dzieła coraz bardziej zbliżone do ludzkich i dąży do tego, aby były od nich nieodróżnialne.

Artykuł skupia się na coraz większym wpływie sztucznej inteligencji na obszary sztuki i muzyki. Analizuje sposoby wykorzystania AI do generowania obrazów, zdjęć i muzyki. Omawia między innymi różne modele uczenia maszynowego, takie jak GAN i VAE, oraz ich zastosowania w praktyce artystycznej. Przytacza artystów korzystających ze sztucznej inteligencji w swojej pracy, oraz wspomina moralne i prawne aspekty korzystania z niej.

Artykuł podkreśla, że pomimo postępów technologicznych, sztuczna inteligencja nie zastąpi ludzkiego talentu i wyobraźni.

Słowa kluczowe: sztuka, sztuczna inteligencja, generowanie obrazów, muzyka, algorytmy uczenia maszynowego.

1. Wprowadzenie

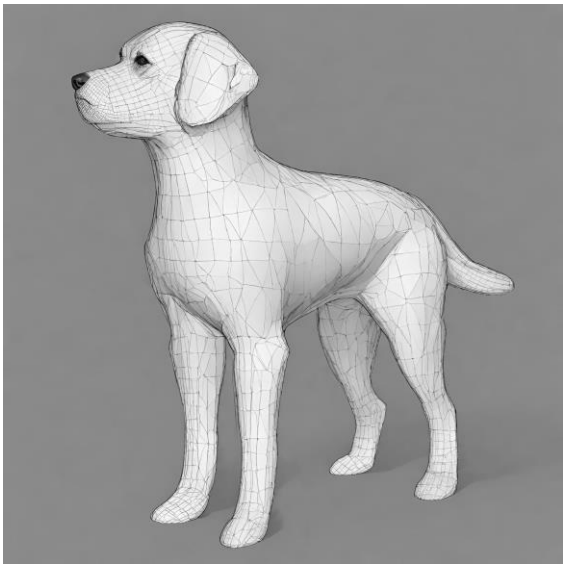
Algorytmy sztucznej inteligencji znajdują zastosowanie w wielu różnych dziedzinach naszego życia. Sztuka nie jest tu wyjątkiem. Na przestrzeni setek lat zmieniała się ona wraz z panującymi trendami. Powstawały nowe narzędzia i techniki tworzenia, aż doszliśmy do momentu, kiedy do stworzenia dzieła wystarczy nam tylko komputer. Algorytmy są w stanie tworzyć coraz więcej, od generowania zdjęć i grafiki, przez pisanie krótkich tekstów literackich aż po komponowanie i wykonywanie utworów.

AI rozszerza granice znane nam do tej pory i otwiera nowe możliwości w tworzeniu obrazów, muzyki a nawet filmów. Zmienia ona sposób, w jaki do tej pory patrzyliśmy na sztukę i sprawia, że każdy z nas może stać się artystą.

2. Jak powstają obrazy generowane przez sztuczną inteligencję?

W przypadku zapytania AI o wygenerowanie obrazu psa, uzyskane wyniki mogą zawierać różne interpretacje tego pojęcia, włączając w to fotografie, modele 3D, oraz rysunki w różnych

stylach artystycznych. Ostateczny wynik jest rezultatem procesu uczenia maszynowego, który może interpretować pojęcie “pies” na wiele różnych sposobów. Algorytmy sztucznej inteligencji nauczyły się interpretować i rozpoznawać grafiki na podstawie zapamiętanych wzorców. Po zapytaniu AI o wygenerowanie psa, wykorzystało ono wszystkie obrazy psów na których było trenowane. Nie były to więc tylko fotografie żywych zwierząt, ale także ilustracje, rysunki i grafiki, często w różnych stylach artystycznych. Sztuczna inteligencja nie miała nigdy do czynienia z żywym zwierzęciem, a jedynie z jego różnymi reprezentacjami w formie graficznej. Zarówno rysunek psa w stylu komiksowym, jak i jego fotografia są dla niej autentycznymi źródłami informacji.



Rysunek 1 Pies, źródło: opracowanie własne

Dla uzyskania większej dokładności generowanych obrazów, musimy bardziej precyzyjnie zdefiniować kryteria grafiki, tak aby algorytm uczył się tylko na obrazach, odpowiadających naszym oczekiwaniom.

Rysunek 2 Pies biegający po łące, źródło: opracowanie własne

Istnieje wiele metod nauki sztucznej inteligencji. Kilka z nich zostanie przedstawionych w dalszej części artykułu.

Na przełomie lat 50 i 60 XX wieku amerykański informatyk Arthur Samuel spopularyzował zwrot “uczenie maszynowe. Stworzył on jeden z pierwszych samouczących się programów- Samuel Checkers. Zapoczątkowało to szereg metod nauki sztucznej inteligencji, które rozwijają się do dziś.

Modele uczenia maszynowego możemy podzielić na trzy grupy: nadzorowane, nienadzorowane oraz uczenie ze wzmocnieniem.

Uczenie nadzorowane wyróżnia się tym, że algorytm wykorzystuje etykietowane zbiory danych. Zestawy danych łączone są w pary, składające się z obiektu wejściowego (uczącego) oraz obiektu wyjściowego. Na ich podstawie algorytm wyszukuje wzorce, np. litera napisana odręcznie i jej odpowiednik, w języku maszynowym. W praktyce, użytkownik wprowadza poprawne dane, aby algorytm ucząc się ich mógł ostatecznie sam zwracać poprawne odpowiedzi. Dane dzielą się na uczące i testowe. Na początku algorytm jest trenowany wyłącznie przy użyciu danych uczących. Następnie wykorzystywane są dane testowe. Dzięki nim wytrenowany algorytm może zostać wykorzystany do przewidywania wyników wyjściowych na podstawie danych testowych i porównywać przewidywania z prawidłowymi wynikami wyjściowymi (np. przyszłe ceny produktów żywnościowych).

W uczeniu nienadzorowanym algorytm nie dostaje etykiet ani poprawnego wyniku wyjściowego, jedynie same dane. Grupuje ze sobą podobne elementy szukając wzorów i generalizacji lub redukuje dane do niewielkiej ilości najbardziej istotnych. W praktyce, w uczeniu nienadzorowanym prawidłowe dane wyjściowe nie są znane, więc model nie może być nauczony przez zmuszenie go do dopasowania prawidłowych danych wejściowych z danymi uczącymi. Metody uczenia nienadzorowanego obejmują m.in wizualizacje, w której im bardziej dane są do siebie podobne, tym bliżej siebie się znajdują lub grupowanie, w którym na podstawie danych wskazujemy grupy podobnych do siebie elementów.

Uczenie przez wzmacnianie stosowane jest w sytuacjach, gdy sztuczna inteligencja musi działać w otoczeniu w którym informacja zwrotna na temat prawidłowości wyborów dostępna jest z opóźnieniem, np. w pojazdach autonomicznych. Proces jest podobny do uczenia się ludzi i zwierząt, algorytm uczy się w oparciu otrzymanie pozytywnej lub negatywnej informacji zwrotnej.

3. Modele uczenia

Generatywne sieci kontradyktoryjne (GAN) są przykładem uczenia bez nadzoru. Technologia sieci generatywnych przeciwników (ang. generative adversarial network) powstała w 2014 roku. Jako pierwszy opisał ją amerykański informatyk Ian Goodfellow. Polega ona na tym, że dwie niezależne głębokie sieci neuronowe przeciwstawiają się sobie i rywalizują ze sobą, jednak tylko jedna z nich wygrywa.

Pierwsza sieć to generator. Jej zadanie polega na uczeniu się i generowaniu dźwięków lub obrazów. Pobiera ona wektor z wybranego rozkładu losowego (np.Gaussa) i generuje z niego obrazy próbne. Są one następnie wykorzystywane jako fałszywe przykłady treningowe dla dyskryminatora. Na początku procesu nie są one realistyczne, jednak zmienia się to wraz z czasem uczenia się. O tym, jak długo trwa ten proces decyduje druga sieć neuronowa czyli dyskryminator, który ocenia dane pod względem autentyczności. Klasyfikuje on przykłady przekazane przez generator jako prawdziwe lub fałszywe. W kolejnych iteracjach parametry generatora są aktualizowane pod kątem tego, jak dobrze obrazy oszukały dyskryminator, a dyskryminatora tak, aby lepiej rozróżniał fałszywe obrazy. Generator tworzy do momentu, w którym dyskryminator nie uzna, że dźwięki lub obrazy są na tyle dobre, że nie da się ich odróżnić od prawdziwych. Najczęściej występującym problemem dla modeli GAN jest sytuacja, w której generator produkuje niewielką ilość próbek z powodu nieznacznie różniących się wartości mediany.

Wariant cGAN (conditional generative network), rozszerza możliwości GAN o warunkowe generowanie obrazów. Uzależnia on generator i dyskryminator od pewnego rodzaju informacji pomocniczych, które są dodatkowymi danymi wejściowymi dla modelu. Taką informacją może być np. płeć lub wiek w przypadku generowania twarzy.

CycleGAN służy do przenoszenia obrazu z jednej domeny do innej. CycleGAN korzysta z dwóch rodzajów funkcji straty. Strata przeciwna pomaga w generowaniu obrazów, które będą lepiej oszukiwać dyskryminator, jednak nie daje gwarancji, że dane wejściowe i wyjściowe będą do siebie podobne. Dzięki stracie spójności cyklicznej obraz przekształcony do innej domeny, i z powrotem do oryginalnej, będzie podobny do obrazu wejściowego.

Autoeknokodery wariacyjne (VAE) generują nowe dane, przypominające te w zbiorze treningowym. Autoenkodery składają się z dwóch elementów- kodera, który przekształca dane wejściowe do postaci reprezentacji w przestrzeni wektorowej (tzw. przestrzeni ukrytej), oraz dekodera który uczy się przekształcać dane z powrotem do oryginalnej formy, bazując na zawartej w przestrzeni ukrytej kodowaniach. Algorytm przekształca warstwy wejściowe w

mniejszą ilość warstw ukrytych, aby wyłapać pewne zależności. Następnie wylicza błąd (różnicę między danymi wejściowymi, a tym co otrzymaliśmy) i poprawia wagi w sieci wykorzystując algorytm propagacji wstecznej.

Rysunek 3 Przykładowe użycie CycleGAN, źródło:https://ai.stackexchange.com/questions/20035/before-gan-what-are-the-commonly-used-techniques-for-image-to-image-translation

Wyróżniane są dwa rodzaje autoenkoderów-niedopełniony i przepełniony. W rodzaju niedopełnionym liczba warstw ukrytych jest mniejsza niż danych wejściowych. Eliminuje to możliwość skopiowania danych i przekazania ich dalej, i zmusza do poszukiwania innego sposobu na odtworzenie danych. Autoenkoder przepełniony ma więcej warstw ukrytych niż wejściowych. Trzeba jednak wtedy “oszukać” autoenkoder, aby nie nauczył się przekazywać wartości do kolejnego węzła ukrytego.

Model VAE wprowadza modyfikację dzięki której zamiast zwykłej rekonstrukcji możliwe jest generowanie nowych obrazów. Dane wejściowe są kodowane jako parametry wektora zmiennych losowych. Oznacza to, że zamiast bezpośredniego kodowania obrazu, koder zwraca wektor średnich i wektor odchyleń standardowych. Dzięki temu, ograniczone jest nadmierne dopasowanie przestrzeni ukrytej. Kodowania próbkowane są losowo z wybranego rozkładu, najczęściej rozkładu Gaussa, określonego przez parametry specyficzne dla tego rozkładu. Dzięki temu, nawet dla tych samych danych wejściowych, dane wyjściowe będą różne.

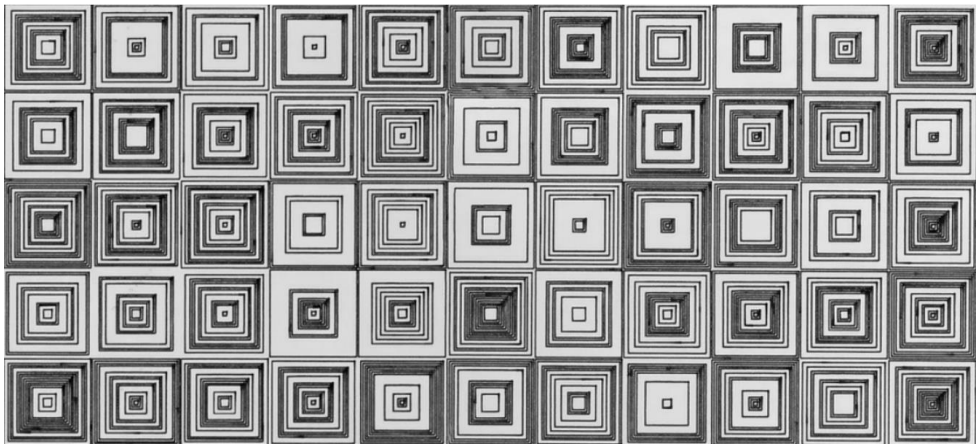
Najsilniejszym argumentem za wykorzystaniem autoenkoderów do uczenia maszynowego jest prostota modeli opartych na tej strukturze oraz szybkość trenowania. Nie sprawdzają się one jednak przy generowaniu wielu szczegółów na obrazie.

Transfomery działają podobnie do autoenkoderów. Główną różnicą między nimi jest zależność, między poszczególnymi wartościami wektorów. Początkowo były stosowane tylko

dla danych tekstowych, jednak wraz ze wzrostem ich popularności znaleziono sposób, aby wykorzystać je do generowania obrazów.

4. Przykłady wykorzystania sztucznej inteligencji przez artystów

Węgierską artystkę Vere Molnár uznaje się za pionierkę w dziedzinie tworzenia grafiki z wykorzystaniem sztucznej inteligencji. W 1968 roku rozpoczęła pracę nad programem komputerowym losowo generującym kompozycje. Jej nowatorskie i wyjątkowe w owym czasie podejście wpłynęło na rozwój sztuki cyfrowej. Dziś jej obrazy możemy podziwiać w muzeach.



Rysunek 4 Praca Very Molnár,
źródło: https://www.tekedia.com/tag/groupe-de-recherche-dart-visuel/

Kolejnym artystą wykorzystującym sztuczną inteligencję do tworzenia obrazów jest Refik Anadol. Tworzy on gigantyczne ekrany, prezentowane jako wygaszacze komputerowe. Za pomocą algorytmów AI generuje kolorowe, ruchome obrazu. Turecko-amerykański artysta korzysta z danych z zakresu botaniki i zoologii zebranych za pomocą technologii LiDAR, która opiera się na działaniu laserów. System służy do mapowania i skanowania przestrzeni, mierząc odległości między obiektami na podstawie zmian długości fal emitowanych przez laser oraz różnic w czasie powrotu wiązki. Prace Anadola były prezentowane m.in w Museum of Modern Art w Nowym Jorku. Artysta twierdzi, że jego celem jest nadanie znaczenia informacjom, poprzez ich przekształcenie na formę wizualną.

Rysunek 5 Praca Refika Anadola,
źródło: https://events.umich.edu/event/109991

5. Narzędzia wykorzystywane do tworzenia grafiki

DALL-E to system stworzony przez OpenAI, który umożliwia generowanie obrazów na podstawie opisu. Po raz pierwszy został zaprezentowany w 2021 roku. Jest on oparty na architekturze GPT-3. DALL-E wykorzystuje modele uczenia bez nadzoru, co oznacza, że dostaje duże ilości par danych tekstowych oraz graficznych, jednak nie otrzymuje informacji o poprawności swoich odpowiedzi. Używa on procesu optymalizacji, aby zminimalizować błąd między przewidywaniami, a rzeczywistymi wynikami.

DALL-E może kontrolować cechy obiektów, ich ilość oraz ułożenie. Potrafi także zmieniać punkt widzenia sceny oraz wywnioskować szczegóły, które nie są dokładnie opisane w tekście.

Rysunek 6 Przykład użycia DALL-E,
źródło: https://akademiabioetyki.pl/nauka/sztuczna-inteligencja-zaskoczyla-naukowcow-tworzac-wlasny-jezyk/

Innym, bardzo popularnym generatorem grafiki AI jest Midjourney. Algorytm tworzy obraz na podstawie tzw. promptów, czyli krótkich fraz tekstowych. Tworząc zapytanie określamy takie elementy jak obiekt, czyli jaka postać ma znajdować się na obrazku, technikę np. obraz olejny lub zdjęcie oraz otoczenie. Następnie należy wybrać także kolor, nastrój oraz kompozycje.

Do stworzenia grafiki możemy także korzystać z takich narzędzi jak DeepArt, RunwayML, GANPaint Studio lub Artbreeder

6. Muzyka tworzona z wykorzystaniem AI

Dzięki sztucznej inteligencji jesteśmy w stanie odtworzyć głos dowolnie wybranej osoby. Coraz popularniejsze, zwłaszcza w Korei, staje się również tworzenie z pomocą AI wirtualnych artystów lub nawet całych zespołów. Jedną z takich grup jest koreański zespół k-popowy IITERNITI. Grupa składa się z 11 członkiń, wygenerowanych przy pomocy technologii Deep Real AI, opracowanej przez Pulse9. Zespół został utworzony w marcu 2021 i od tamtej pory wydał kilka singli.

Rysunek 7 Grupa IITERNITI, źródło: https://pabii.com/pl/news/282422/

W dzisiejszych czasach istnieje wielu artystów stworzonych z pomocą sztucznej inteligencji. Często wykorzystywana jest technologia deep fake, czyli technika manipulacji dźwiękiem i obrazem wideo. Algorytm analizuje nagrania, i na ich podstawie uczy się imitować zachowanie konkretnych osób. Technologia pozwala na stworzenie trudnych do odróżnienia, fałszywych materiałów. Algorytmy deepfake mogą dopasowywać m.in. ruchy ust i mimiki

twarzy. Dzięki swojej dokładności, technologia może być niestety wykorzystywana nie tylko do rozrywki, ale także do nieetycznych celów.

7. Prawa autorskie do dzieł sztucznej inteligencji, etyka i moralność

W 2023 roku niemiecki fotograf Boris Eldagsen został wyróżniony w prestiżowym konkursie Sony World Photography awards, w kategorii fotografii kreatywnej. Nie przyjął jednak nagrody. Dlaczego? Ponieważ zdjęcie nie zostało wykonane przez niego. Fotografię w stylu połowy XX wieku przedstawiającą dwie kobiety i enigmatycznie wyglądające dodatkowe dłonie w całości wygenerowała sztuczna inteligencja.

Rysunek 8 Zwycięska fotografia,
źródło:https://www.fotopolis.pl/newsy-sprzetowe/branza/36636-ai-to-nie-fotografia-autor-zdjecia-ai-ktore-nagrodzono-na-swpa-odmowil-przyjecia-nagrody

Podobna sytuacja miała miejsce rok wcześniej, w Kolorado. Na konkursie sztuk pięknych organizowanym przez Colorado State Fair pierwsze miejsce zajął projektant gier Jason Allen. Zwycięski obraz “Théâtre D'opéra Spatial” przedstawijący kosmiczny teatr operowy został stworzony z pomocą sztucznej inteligencji. Decyzja sędziów spotkał się z jednak z nieprzychylną opinią odbiorców. Krytycy zarzucają Allenowi oszustwo.

Komu więc przysługują prawa autorskie do obrazów tworzonych przez AI?

Rysunek 9 Zwycięska grafika,
źródło:https://www.dobreprogramy.pl/sztuczna-inteligencja-pokonala-czlowieka-lepsza-nawet-w-sztuce,6807609411148288a

Zgodnie z polskim prawem, aby dzieła były objęte prawami autorskimi, muszą być efektem pracy twórczej człowieka. Wynika z tego, że utwory wytworzone przez AI nie podlegają ochronie prawnoautorskiej. Ponieważ prawo nie nadąża za rozwojem sztucznej inteligencji, nie istnieją prawne regulacje dotyczące praw autorskich i AI. Prowadzi to do niejasności i luk prawnych. Obecnie trwają prace nad regulacjami prawnymi takimi jak AI Act, mającymi uregulować te kwestie. Oznacza to, że możemy wykorzystywać dzieła wygenerowane przez sztuczną inteligencję do użytku własnego, np. w reklamach, animacjach czy prezentacjach. Należy jednak uważać, by nie naruszyć przy tym praw autorskich innych osób np. przez wykorzystanie do treningu sztucznej inteligencji własności intelektualnej innych osób.

Należy również pamiętać, aby korzystając ze sztucznej inteligencji nie tworzyć dzieł obraźliwych, rasistowskich i szkodliwych dla społeczeństwa. Ważne jest, aby dane treningowe, z których korzystamy, zawierały materiały poprawne politycznie i etycznie. Należy unikać uprzedzeń i krzywdzących stereotypów. Nie powinno się rozpowszechniać treści obraźliwych stworzonych przez AI.

8. Różnice między artystą a sztuczną inteligencją

Twórcy sztucznej inteligencji dążą do tego, aby dzieła przez nią tworzone były nie do odróżnienia od tych wykonanych przez ludzi. Poprzez odpowiednie trenowanie algorytmów, mogą one naśladować style różnych artystów i epok. W części przypadków możemy jednak rozpoznać, czy wykonawcą jest sztuczna inteligencja, czy człowiek. Główną wskazówka mogą

być niereallistyczne detale, takie jak zaburzone i nienaturalne proporcje, niedoskonałości czy niejasne tła. Algorytmy nie posiadają emocji ani osobowości więc dzieła przez nie wykonane mogą budzić wątpliwości co do intencji twórcy. Dodatkowo, twarze wykonane przez AI są zlepkiem wielu twarzy różnych ludzi, więc emocje które pokazują mogą być ciężkie do odczytania.

Twórcy generatorów AI zaczynają wprowadzać na swoich grafikach różne oznaczenia i znaki wodne, aby były one łatwiejsze do odróżnienia.

9. Podsumowanie

Sztuczna inteligencja wkracza coraz głębiej w obszary sztuki i muzyki. Z biegiem czasu odgrywa coraz większą rolę w procesie twórczym. Wiele dziedzin takich jak grafika, muzyka czy fotografia zyskuje nowe oblicze dzięki algorytmom, które są w stanie dać artystom niewyobrażalne możliwości. Różnego rodzaju generatory i boty są w stanie otworzyć drzwi do świata artystów osobom nie mającym do tej pory do czynienia z żadną dziedziną sztuki. Jednakże, wraz ze wzrostem popularności sztucznej inteligencji pojawiają się kontrowersje i problemy natury moralnej, takie jak prawa autorskie do dzieł stworzonych przez AI.

Artykuł podkreśla rolę sztucznej inteligencji w świecie artystów oraz opisuje możliwości i metody działania algorytmów.

Źródła internetowe:

1. https://biletprosze.pl/technologie-w-kulturze/ai-i-sztuka/
2. https://stationof.art/jak-sztuczna-inteligencja-zmienia-rynek-sztuki/
3. https://businessinsider.com.pl/technologie/digital-poland/sztuczna-inteligencja-w-sztuce-szansa-czy-zagrozenie/7lq70sx
4. https://aioai.pl/ai-w-sztuce-kiedy-technologia-spotyka-kreatywnosc/
5. https://stationof.art/kontrowersje-wokol-sztuki-generowanej-przez-ai/
6. https://www.capcut.com/pl-pl/resource/use-ai-to-create-art
7. https://cyberdefence24.pl/cybermagazyn/cybermagazyn-czy-sztuka-generowana-przez-sztuczna-inteligencje-moze-zastapic-artystow
8. https://digitalmasterinstitute.com/generatory-obrazow-ai/
9. https://mindboxgroup.com/pl/uczenie-nadzorowane-i-nienadzorowane-rodzaje-nauczania-maszynowego/
10. https://neonshake.pl/blog/jak-midjourney-usprawnia-generowanie-grafik/

Krystian Kiełbasa, Hubert Futoma, Oskar Niedziałek
Studenckie Koło naukowe informatyków „KOD”

dr inż. Bartosz Trybus
Opiekun Koła Naukowego

Zastosowanie Kubernetes w tworzeniu i zarządzaniu aplikacjami

Streszczenie

Artykuł koncentruje się na opisie możliwości i zastosowań platformy Kubernetes w procesie tworzenia aplikacji przez programistów oraz zarządzanie nimi. Określona zostanie przede wszystkim ocena jej efektywności w automatyzacji, skalowaniu, czy zarządzaniu cyklem życia kontenerów, odzwierciedlających działające aplikacje.

W treści zostanie przedstawiona zasada działania platformy jak również ważne podczas jej działania najważniejsze elementy, takie jak m.in.: kontener, pod, węzeł, klaster, a nawet składowe komponenty. Omówione zostaną również zalety i wady korzystania z platformy Kubernetes, jak również przedstawione przykłady praktycznego wykorzystania platformy pokazujące skuteczne użycie podstawowych jej możliwości. Na potrzeby artykułu przykłady bazować będą o podstawową aplikację stworzonej w Spring Boot (Java). Artykuł również przedstawi porównanie platformy Kubernetes z innymi platformami, takimi jak Docker Swarm, czy Nomad.

Słowa kluczowe: kubernetes, platforma, kontener, pod, wezęł, klaster, serwis, docker, swarm, nomad.

1. Wprowadzenie

Wraz z dynamicznym rozwojem technologii, programiści muszą nie tylko efektywnie kodować, ale także zarządzać całym cyklem życia ich aplikacji, od tworzenia po wdrożenie i utrzymanie. Kubernetes, stworzony przez Google i przekazany do Cloud Native Computing Foundation, stał się jednym z najpopularniejszych narzędzi, które wspierają programistów na każdym etapie pracy nad aplikacją. Platforma ta pozwala na tworzenie spójnych lokalnych środowisk deweloperskich, automatyzację procesów wdrażania oraz testowanie ich w izolacji, dzięki czemu można szybko i efektywnie tworzyć, testować i integrować nowe funkcje w aplikacjach. Dodatkowo umożliwianie łatwego zarządzania zależnościami i skalowanie zasobów podczas testów obciążeniowych pozwalala zapewnić jakość i niezawodność tworzonych aplikacji. W miarę gdy aplikacje stają się coraz bardziej złożone i wymagające, tradycyjne metody zarządzania nimi często okazują się niewystarczające i to właśnie wtedy, użycie Kubernetesa, staje się uproszczeniem całego procesu.

Ze względu na to niniejszy artykuł przede wszystkim skupi się na przedstawieniu Kubernetesa jako efektywnego narzędzia do zarządzania aplikacjami. Zostaną omówione jakie korzyści płyną z jego stosowania, takie jak możliwość automatyzacji procesów, zapewnienia

skalowalności i łatwego zarządzania cyklem życia aplikacji, a też przedstawione zostaną elementy samego Kubernetesa omawiając je poprzez praktyczne przykłady opartych na projekcie napisanym w Spring Boot (Java), co pozwoli zobrazować jego możliwości w rzeczywistych scenariuszach.

W artykule zostanie również porównany Kubernetes z innymi popularnymi platformami do zarządzania, takimi jak Docker Swarm i Nomad, podkreślając różnice w podejściu oraz funkcjonalności każdej z nich. Celem jest dostarczenie kompleksowego przeglądu i oceny Kubernetesa jako narzędzia, które może znacząco poprawić efektywność oraz elastyczność zarządzania aplikacjami w nowoczesnych środowiskach IT.

2. Podstawy i architektura Kubernetes

Kubernetes, znany również jako K8s, został opracowany przez Google i po raz pierwszy udostępniony jako open-source w 2014 roku. Celem jego stworzenia było zapewnienie platformy do automatyzacji wdrażania, skalowania i operacji aplikacji kontenerowych (tj. uruchomionych w tzw. kontenerach). W 2015 roku Kubernetes został przekazany do Cloud Native Computing Foundation (CNCF), umożliwiającym tym przyspieszenie jego rozwoju i adopcję przez społeczność oraz przedsiębiorstwa. Od tego czasu stał się standardem w zarządzaniu kontenerami, zdobywając szerokie wsparcie i integracje z wieloma narzędziami DevOps (tj. praktyki łączącej rozwój oprogramowania i operacje IT). Dzięki wsparciu CNCF Kubernetes ewoluował, oferując coraz bardziej zaawansowane funkcje, co w jeszcze większym stopniu pozwoliło odpowiadać na potrzeby nowoczesnych aplikacji.

K8s jako narzędzie składa się z kilku kluczowych komponentów, ściśle ze sobą współpracujących, aby zapewnić skuteczne, a przede wszystkim efektywne zarządzanie aplikacjami kontenerowymi:

- *kontenery* - lekkie, samowystarczalne środowiska uruchomieniowe zawierające wszystko, co jest potrzebne do uruchomienia aplikacji, w tym kod, zależności, biblioteki systemowe i narzędzia. Ich zadaniem jest izolacja aplikacji od systemu operacyjnego, zwiększając ich przenośność i bezpieczeństwo. Zaletą takiego rozwiązania jest to, że programiści mogą być pewni indetyczności działania niezależnie od środowiska, w którym są uruchamiane,
- *pody (eng. Pods)* - są najmniejszymi jednostki w Kubernetes. Zawierają jeden lub więcej kontenerów. Pody są uruchamiane na węzłach i reprezentują jednostkę wdrożeniową, którą Kubernetes zarządza. Każdy pod ma własny adres IP, co umożliwia łatwą komunikację między aplikacjami uruchomionymi w różnych podach,

- *węzły (eng. nodes)* - to maszyny fizyczne lub wirtualne, na których uruchamiane są pody. Każdy węzeł zawiera serwer kubelet, który zarządza podami w danym węźle oraz serwer proxy do obsługi sieci. Węzły mogą być dodawane lub usuwane z klastra w miarę potrzeb, co pozwala na elastyczne skalowanie zasobów,
- *klastry (eng. Clusters)* - to zbiór węzłów zarządzanych przez Kubernetes, które współpracują jako jedna jednostka obliczeniowa. Klaster zapewnia wysoką dostępność i tzw. redundancję, co zwiększa niezawodność uruchomionej aplikacji. Dzięki klastrowi Kubernetes może dynamicznie zarządzać zasobami, optymalizując ich wykorzystanie,
- *Master Node* - węzeł zarządzający klastrem i jego stanem. Składa się z kilku komponentów, takich jak etcd (system przechowywania klucz-wartość), kube-apiserver (interfejs API Kubernetes), kube-scheduler (przydziela pody do węzłów) i kube-controller-manager (zarządza kontrolerami Kubernetes). Węzeł jest odpowiedzialny za koordynację wszystkich działań w klastrze, zapewniając, że aplikacje działają zgodnie z oczekiwaniami.

Kubernetes wykorzystując powyższe komponenty działa jako system orkiestracji, automatyzujący wiele zadań związanych z zarządzaniem aplikacjami uruchomionymi w kontenerach. Do kluczowych jego funkcji możemy zaliczyć:

- automatyczne skalowanie - K8s może automatycznie skalować aplikacje w górę lub w dół w odpowiedzi na zmieniające się jej obciążenie. Dzięki temu zasoby są wykorzystywane efektywnie, oszczędzając koszty i uzyskując jednocześnie lepszą wydajność aplikacji. Skalowanie to może być oparte na metrykach takich jak CPU, pamięć czy specyficzne wskaźniki aplikacyjne,
- równoważenie obciążenia (eng. load balancing) - Kubernetes automatycznie rozdziela ruch sieciowy między różne pody, zapewniając równomierne obciążenie. Zapewnia to lepszą dostępność i wydajność aplikacji, ponieważ żaden pod nie jest przeciążony. Dodatkową zaletą jest to, że równoważenie obciążenia działa również w przypadku niespodziewanych awarii, przekierowując w.w. ruch sieciowy do zdrowych podów,
- auto-naprawa (eng. self-healing) - platforma automatycznie restartuje lub replikuje pody, które uległy awarii, co pozwala na zapewnienie ciągłości działania aplikacji. Jeśli pod lub węzeł przestanie działać poprawnie, Kubernetes automatycznie podejmuje kroki naprawcze, takie jak ponowne uruchomienie poda lub jego migracja na inny węzeł. To zapewnia wysoką dostępność i minimalizuje przestoje,

- automatyczne wdrażanie (eng. auto-deploy) i funkcja przywracania (eng. roll-back) - Kubernetes oprócz wspomnianych podstawowych funkcjonalności umożliwia automatyczne wdrażanie nowych wersji aplikacji, a zarazem możliwość szybkiego jej powrotu do poprzedniej wersji w przypadku problemów. Proces samego wdrażania jest kontrolowany, co minimalizuje ryzyko błędów i zakłóceń w działaniu aplikacji. W przypadku wykrycia problemów podczas wdrożenia Kubernetes automatycznie przeprowadza roll-back, przywracając poprzednią stabilną wersję.

3. Tworzenie i zarządzanie aplikacjami

Tworzenie aplikacji w Kubernetes jest mniej więcej prostym co do trudności procesem i zaczyna się od konteneryzacji aplikacji za pomocą Dockerfile, który definiuje, jak aplikacja powinna być zbudowana i uruchomiona. Konfiguracja Dockerfile zawiera instrukcje dotyczące instalacji zależności, kopiowania plików aplikacji oraz konfiguracji uruchamiania. Po zbudowaniu obrazu Docker aplikacja jest gotowa do wdrożenia w środowisku Kubernetes.

```
FROM gradle:7.4.0-jdk17 AS GRADLE_BUILD
COPY --chown=gradle:gradle . /home/gradle/project
WORKDIR /home/gradle/project
RUN gradle build

FROM openjdk:17.0.2-slim-buster
EXPOSE 8080
COPY --from=GRADLE_BUILD /home/gradle/project/build/libs/*.jar /app.jar
ENTRYPOINT ["java","-jar","/app.jar"]
```

Rysunek 1. Kod pliku Dockerfile przygotowanej aplikacji
Źródło: opracowanie własne

Konfiguracja przedstawiona na rysunku 1 składa się z dwóch części. W pierwszej części używając obrazu Gradle zostaje zbudowana aplikacja Spring Boot, zapoczątkowana kopiowaniem jej kodu źródłowego do kontenera, po ustawienie odpowiednich uprawnień. W drugiej natomiast poprzez obraz OpenJDK zostaje określone uruchomienie aplikacji wraz z potrzebnymi informacjami.

Kubernetes do działania potrzebuje jednak jeszcze innej konfiguracji (manifestu) takiej, która zdefiniuje specyfikację poda, takie jak obraz Dockera, zmienne środowiskowe, porty oraz zasoby. Ta następnie przesyłana jest do API Kubernetes, który zarządza cyklem życia poda, przydzielając go do odpowiednich węzłów w klastrze. Dzięki temu procesowi aplikacja jest automatycznie uruchamiana i zarządzana przez Kubernetes, zapewniając jej skalowalność i niezawodność.

```yaml
apiVersion: apps/v1 # Wersja API Kubernetes
kind: Deployment # Typ zasobu, który tworzymy
metadata:
  name: spring-boot-deployment
  labels:
    app: spring-boot # Etykieta wdrożenia
spec:
  replicas: 1 # Liczba uruchomionych podów
  selector:
    matchLabels:
      app: spring-boot # Etykieta dla zarządzanych podów
  template:
    metadata:
      labels:
        app: spring-boot # Etykieta poda
    spec:
      containers:
        - name: spring-boot-container # Nazwa kontenera
          image: spring-app:local # Obraz użytego kontenera
          imagePullPolicy: Never # Pobieranie obrazu
          ports:
            - containerPort: 8080 # Port działania w kontenerze
```

Rysunek 2. Kod manifestu wdrożenia dla K8s przygotowanej aplikacji
Źródło: opracowanie własne

Konfiguracja przedstawiona na rysunku 2 pozwala na późniejszą możliwość ręcznego wdrażania aplikacji. Proces ten jest jednak często automatyzowany przez użycie zewnętrznych narzędzi takich jak Jenkins, GitLab CI lub Tekton pozwalający na tzw. automatyzację CI/CD (Continuous Integration/Continuous Deployment), umożliwiający automatyczne budowanie, testowanie i wdrażanie aplikacji po każdej zmianie kodu. W przypadku np. narzędzia Jenkins, gdy jest on skonfigurowany do uruchamiania tzw. pipeline'u (zautomatyzowanego ciągu kroków do budowania, testowania i wdrażania oprogramowania), pozwala budować obraz Dockera, przesyłać go do rejestru kontenerów, a następnie przesłać aktualizację informacji w Kubernetes.

Autorzy Kubernetesa postanowili wesprzeć ten proces, zapewniając narzędzia takie jak Helm do zarządzania wersjami aplikacji, czy ich automatyczne wdrażanie. Helm Chart definiuje wszystkie zasoby Kubernetesa potrzebne do uruchomienia aplikacji, co pozwala na zautomatyzowane i powtarzalne wdrożenia. Dzięki temu programiści mogą skupić się na rozwijaniu kodu, podczas gdy Kubernetes zarządza wdrożeniami i zapewnia ciągłość działania aplikacji.

4. Skalowanie aplikacji

Kubernetes jest narzędziem, który zapewnia zaawansowane mechanizmy skalowania aplikacji, co jest kluczowe dla utrzymania wydajności i dostępności usług w zmieniających się warunkach obciążenia. Aby zobrazować, jak działa skalowanie w K8s, można posłużyć się przykładem aplikacji Spring Boot zbudowanej za pomocą Gradle. Wykorzystując konfigurację ze zmienioną wartością parametru replik, czyli ilości kopii aplikacji i zbudowany wcześniej obraz Dockera, możemy uruchomić kilka instancji tej samej aplikacji, gdzie każda z nich będzie się znajdowała na osobnym podzie jak przedstawiono na rysunku 3.

```
$ eval $(minikube docker-env)
$ docker build -q -t spring-app:local .
sha256:b639def2cad3e8533c501d87eb83a85f0661aeaf34d5269f481fc81186c9dae5
$ kubectl apply -f spring-app-k8s.yml --context=minikube
deployment.apps/spring-boot-deployment created
$ kubectl scale deployment spring-boot-deployment --replicas=5
deployment.apps/spring-boot-deployment scaled
$ kubectl get pods -l app=spring-boot
NAME                                   READY   STATUS    RESTARTS   AGE
spring-boot-deployment-5bf55ccbdd-645l9   1/1     Running   0          80s
spring-boot-deployment-5bf55ccbdd-fxp45   1/1     Running   0          105s
spring-boot-deployment-5bf55ccbdd-jc5hs   1/1     Running   0          80s
spring-boot-deployment-5bf55ccbdd-khmzc   1/1     Running   0          80s
spring-boot-deployment-5bf55ccbdd-n64x2   1/1     Running   0          80s
```

Rysunek 3. Komendy CLI wraz z rezultatem uruchomienia kilka instancji przykładowej aplikacji
Źródło: opracowanie własne

Uruchomienie kilku instancji tej samej aplikacji pozwala na równomierne rozłożenie obciążenia między podami, co zwiększa wydajność i niezawodność systemu. Każdy pod działa jako niezależna jednostka, co pozwala na izolację problemów i minimalizację wpływu awarii na cały system. W przypadku wzrostu obciążenia Kubernetes automatycznie skaluje aplikację w górę, tworząc dodatkowe repliki poda, w przypadku, gdy obciążenie maleje, K8s redukuje liczbę instancji, aby zoptymalizować wykorzystanie zasobów. Takie podejście zapewnia elastyczność i efektywność zarządzania zasobami w dynamicznych środowiskach produkcyjnych.

5. Zarządzanie wieloma instancjami

Skalowanie aplikacji w Kubernetes zwiększa liczbę podów, co pozwala na uruchomienie wielu instancji tej samej aplikacji. Zarządzanie jednak taką aplikacją wymaga efektywnej koordynacji komunikacji między instancjami, równoważenia obciążenia oraz zapewnienia wysokiej dostępności. Kubernetes, dzięki swoim wbudowanym mechanizmom, doskonale wspiera te procesy, umożliwiając automatyczne rozdzielanie ruchu sieciowego i dynamiczne

dostosowywanie zasobów. Dla przykładu w przygotowanej aplikacji zaimplementowano punkt końcowego GET */instance*, który w praktyczny sposób pozwala na ukazanie w.w zjawiska poprzez zwracanie unikalnego ID, będącego prowizorycznym sposobem przedstawienia instancji aplikacji, która procesuje dane żądanie. Funkcjonalność ta, aby działała wymaga zdefiniowania pliku konfiguracyjnego serwisu, który po uruchomieniu umożliwi rozdzielenie ruchu sieciowego pomiędzy różne instancje aplikacji. Serwis jest kluczowym komponentem K8s, który zapewnia dostęp do aplikacji bez konieczności wiedzy o tym, na którym konkretnie podzie jej instancja jest uruchomiona.

```yaml
apiVersion: v1  # Wersja API Kubernetes
kind: Service  # Typ zasobu, który tworzymy
metadata:
  name: spring-boot-service  # Nazwa serwisu
spec:
  selector:
    app: spring-boot  # Etykieta podów dla serwisu
  ports:
    - protocol: TCP  # Protokół serwisu
      port: 80  # Port dostępności serwisu
      targetPort: 8080  # Port aplikacji
  type: NodePort  # Typ serwisu wystawiającego
```

Rysunek 4. Kod manifestu serwisu dla K8s przygotowanej aplikacji
Źródło: opracowanie własne

Wykorzystując konfiguracje serwisu przedstawioną na rysunku 4 uzyskujemy wysoką dostępność aplikacji. Serwis w Kubernetes pełni funkcję load balancera, który dynamicznie przydziela zasoby, zapewniając optymalne wykorzystanie infrastruktury. Dzięki temu możemy efektywnie zarządzać wieloma instancjami aplikacji, minimalizując ryzyko przestojów oraz zwiększając niezawodność systemu. W praktyce, zapytania kierowane do punktu końcowego GET */instance* są równomiernie rozdzielane między dostępne instancje, co pozwala na obserwowanie unikalnych ID zwracanych przez różne pody, potwierdzając skuteczność działania mechanizmu równoważenia obciążenia jak przedstawiono na rysunku 5.

```
$ kubectl apply -f spring-app-k8s.yml --context=minikube
deployment.apps/spring-boot-deployment created
$ kubectl apply -f spring-app-service.yml --context=minikube
service/spring-boot-service created
$ kubectl scale deployment spring-boot-deployment --replicas=5
deployment.apps/spring-boot-deployment scaled
$ kubectl get services
NAME                 TYPE        CLUSTER-IP      EXTERNAL-IP   PORT(S)        AGE
kubernetes           ClusterIP   10.96.0.1       <none>        443/TCP        97m
spring-boot-service  NodePort    10.105.180.70   <none>        80:32657/TCP   2m40s
$ minikube ip
192.168.49.2
$ for i in {1..5}; do curl -s http://$(minikube ip):32657/instance; echo; done
Spring boot app, instance ID: c692d5a1-bf2a-498e-aa3c-992eac41a1f7
Spring boot app, instance ID: cc622280-6a84-4d9f-98ae-20cdc1a69457
Spring boot app, instance ID: 69d06785-00a6-401a-93f9-ccf0bb707e10
Spring boot app, instance ID: c692d5a1-bf2a-498e-aa3c-992eac41a1f7
Spring boot app, instance ID: 0a4a08ee-d0d2-4513-93f1-125f64effed5
```

Rysunek 5. Komendy CLI wraz z rezultatem działania load balancing'u dla przykładowej aplikacji
Źródło: opracowanie własne

6. Optymalizacja zasobów

Kubernetes, jako zaawansowane narzędzie do wdrażania aplikacji kontenerowych, oferuje funkcjonalności optymalizacji zasobów, które pozwalają na efektywne wykorzystanie dostępnych zasobów sprzętowych minimalizując przy tym koszty operacyjne. W K8s dynamiczne przydzielanie zasobów jest konfigurowalne w konfiguracji wdrażania poprzez podanie parametrów requests i limits, które precyzyjnie określają zasoby, jakie każda aplikacja może zużywać. W przypadku parametr requests definiuje on minimalne zasoby niezbędne do poprawnego działania aplikacji, natomiast limits określa maksymalne zasoby, jakie aplikacja może zużyć. Ich kompletną konfigurację na przykładowej aplikacji przedstawiono na rysunku 6.

```yaml
## ... kod z rysunku 2
containers:
  - name: spring-boot-container # Nazwa kontenera
    image: spring-app:local # Obraz użytego kontenera
    imagePullPolicy: Never # Pobieranie obrazu
    ports:
      - containerPort: 8080 # Port działania w kontenerze
    resources:
      requests:
        memory: "512Mi" # Minimalna ilość pamięci
        cpu: "500m" # Minimalna ilość CPU, 500m = 0.5 CPU
      limits:
        memory: "1Gi" # Maksymalna ilość pamięci
        cpu: "1" # Maksymalna ilość CPU, 1 CPU
```

Rysunek 6. Kod manifestu wdrażania z zasobami dla K8s przygotowanej aplikacji
Źródło: opracowanie własne

Konfiguracja przedstawiona na rysunku 6 pozwala zapewnić stabilność wdrażanej aplikacji przy zagwarantowanym optymalnym wykorzystanie dostępnych zasobów. W praktyce oznacza to, że Kubernetes gwarantuje aplikacji określone minimum zasobów, aby mogła działać poprawnie, jednocześnie ograniczając jej maksymalne zużycie, aby zapobiec nadmiernemu obciążeniu klastra. Takie podejście pozwala na efektywne zarządzanie zasobami, ale przede wszystkim lepszą kontrolę nad wydajnością aplikacji.

Precyzyjne określenie parametrów requests i limits pozwala administratorom systemów unikać scenariuszy, w których aplikacje konkurują o zasoby, prowadząc do degradacji wydajności. Optymalizacja zasobów w Kubernetes jest kluczowym elementem zarządzania środowiskiem produkcyjnym, zapewniającym wysoką dostępność i stabilność usług nawet pod zmiennym obciążeniem. Dostosowywanie tych ustawień na podstawie rzeczywistych danych o zużyciu pozwala na dynamiczne reagowanie na zmiany w obciążeniu aplikacji. Wykorzystując zewnętrzne narzędzia takie jak Prometheus, czy Grafana, możliwe jest bieżące śledzenie metryk zużycia i odpowiednie dostosowywanie przyszłych konfiguracji wdrożeń co w przypadku K8s pozwala wspierać proces optymalizacji zasobów w klastrze. W przypadku przykładowej aplikacji stan podów po ustawieniu w.w. parametrów przedstawiono na rysunku 7.

```
$ kubectl apply -f spring-app-k8s.yml --context=minikube
deployment.apps/spring-boot-deployment created
$ kubectl get pods
NAME                                    READY   STATUS    RESTARTS   AGE
spring-boot-deployment-567fb6c4b6-kf2mr   1/1     Running   0          77s
$ kubectl describe pod spring-boot-deployment-567fb6c4b6-kf2mr --context=minikube \
>    | grep -A 7 "Limits"
    Limits:
      cpu:     1
      memory:  1Gi
    Requests:
      cpu:        500m
      memory:     512Mi
    Environment:  <none>
    Mounts:
```

Rysunek 7. Komendy CLI wraz z rezultatem zasobów dla poda K8s przykładowej aplikacji
Źródło: opracowanie własne

7. Zalety i wady rozwiązania

Kubernetes jako jedno z najpopularniejszych narzędzi do zarządzania aplikacjami kontenerowymi oferuje wiele korzyści, ale niesie ze sobą również pewne wyzwania. Przedstawione w poprzednich rozdziałach możliwości dotyczące automatyzacji, skalowalności i niezawodności są kluczowymi aspektami tego narzędzia, lecz w przypadku K8 można jeszczów mówić o innych aspektach jego działania, gdzie do zalet przypisać można także:

- Łatwość zarządzania - Kubernetes oprócz przedstawionej wcześniej deklaratywnej konfiguracji oferuje ujednolicony interfejs zarządzania, który ułatwia operacje na aplikacjach i infrastrukturze,
- Wsparcie społeczności i ekosystem - Kubernetes od początku cieszy się szerokim wsparciem społeczności oraz dużych firm technologicznych, pozwalając na gwarantowany ciągły jego rozwój i wsparcie narzędzi. Bogaty ekosystem narzędzi i rozszerzeń, takich jak Helm, Prometheus i Istio, które integrują się z K8s, ułatwiają zarządzanie, ale i rozszerzają jego możliwości.

W przypadku Kubernetesa można także mówić o wadach, które w pewien sposób w pewnym momencie jego integracji z infrastrukturą utrudniają jego użycie:

- Złożoność konfiguracji - często dla początkujących użytkowników konfiguracja i zarządzanie Kubernetes staje się zbyt złożone, gdyż komponenty składające się na jego działanie w większości przypadków wymagają większego zrozumienia, a sama nauka i opanowanie może być czasochłonne i wymagać znaczących inwestycji w szkolenia,
- Wymagania dotyczące zasobów - Kubernetes ze względu na zasadę swojego działania w pewnym momencie wykorzystania może powodować dużą zasobożerność, zwłaszcza w przypadku dużych klastrów. Wówczas wymaga on znacznej ilości pamięci RAM i mocy obliczeniowej, powodując zwiększenie kosztów infrastruktury, na którym stoi. Często też zasoby, którymi dysponuje K8s musi zostać skonsumowane przez komponenty zarządzające, które się na nie składają, takie jak master nodes, etcd i kubelet.

8. Porównanie z innymi platformami

Kubernetes jest jednym z najpopularniejszych narzędzi do zarządzania aplikacjami kontenerowymi, ale we współczesnym świecie nie jest jedynym rozwiązaniem na rynku. Wybór zatem odpowiedniej platformy do orkiestracji kontenerów zależy od wielu czynników, począwszy od takich jak specyfiki projektu, po wymagania techniczne, zasoby i preferencje zespołu.

Pierwszym możliwym wyborem obok K8s jest Docker Swarm, czyli narzędzie stworzone przez Docker, Inc. Ze względu na to, że został on stworzony bezpośrednio przez twórców Dockera integruje się bezpośrednio z nim, co ułatwia jego wdrożenie dla użytkowników już korzystających z Dockera. Swarm jednak, pomimo bycia częścią ekosystemu Dockera, zapewniając tym samym spójność i łatwość użycia, jest ograniczona w porównaniu do

Kubernetes, gdyż nie oferuje zaawansowanych mechanizmów, takich jak automatyczne skalowanie na podstawie metryk, co ogranicza jego elastyczność i skalowalność w dużych, dynamicznych środowiskach, co powoduje korzystność wybrania akurat jego w przypadku mniejszych zespołów lub projektów, które nie wymagają zaawansowanych funkcji K8s.

Nomad, stworzony przez HashiCorp, jest znany z elastyczności i możliwości zarządzania różnorodnymi typami zadań, nie tylko kontenerami. Zaletą Nomada jest to, że może zarządzać nie tylko kontenerami Dockera, ale także innymi typami zadań, takimi jak aplikacje Java, .NET, po zadania opakowane w skrypty. Niestety mimo swojej elastyczności, Nomad w porównaniu do Kubernetes nie oferuje rozbudowanych funkcji automatyzacji i zarządzania jak Kubernetes, co dzięki integracji z szerokim ekosystemem narzędzi, takich jak Helm, Prometheus i Istio, oferuje bardziej kompleksowe rozwiązania do zarządzania aplikacjami i całą infrastrukturą.

Kubernetes wyróżnia się na tle Docker Swarm i Nomad przede wszystkim posiadaniem zaawansowanych funkcji automatyzacji i skalowalności. Jego zdolność do automatycznego skalowania aplikacji, zarządzania cyklem życia podów oraz integracji z narzędziami do monitorowania i zarządzania zasobami czyni go bardziej wszechstronnym i potężnym narzędziem w porównaniu do innych. Wybór odpowiedniej platformy do orkiestracji kontenerów zależy jednak przede wszystkim od specyficznych potrzeb projektu i preferencji zespołu, a decyzja o wyborze w dużej mierze uwzględnia obecne potrzeby, jak i przyszłe wymagania rozwojowe.

9. Podsumowanie

Kubernetes jest ciekawym, a zarazem mocnym narzędziem do orkiestracji kontenerów, które oferuje szeroki zakres funkcji znacząco ułatwiających zarządzanie aplikacjami kontenerowymi. Jego zdolność do automatyzacji, skalowania i zapewnienia niezawodności powoduje, że staje się on interesującym rozwiązaniem w porównaniu z innymi platformami. Pomimo pewnych wyzwań związanych z konfiguracją i zarządzaniem, Kubernetes oferuje elastyczność i wsparcie, które mogą znacząco poprawić efektywność działań w złożonych środowiskach produkcyjnych. Taka zaleta powoduje to, że nauka Kubernetesa i późniejsze jego wykorzystanie może przynieść znaczące korzyści w danym projektach i zapewnić solidną podstawę do niezawodnego sposobu zarządzania aplikacjami w nowoczesnych środowiskach IT.

Literatura

1. Burns B., Beda J., Hightower K., Kubernetes. *Tworzenie niezawodnych systemów rozproszonych. Wydanie II*, Helion, Gliwice 2020.
2. Burns B., Villalba E., Strebel D., Evenson L., *Najlepsze praktyki w Kubernetes. Jak budować udane aplikacje*, Helion, Gliwice 2020.
3. Espinosa A., McKendrick R., *Docker. Wydajność i optymalizacja pracy aplikacji. Wydanie II*, Helios, Gliwice 2020.

Źródła internetowe

1. https://kubernetes.io/docs/home/ (dostęp: 13.06.2024).
2. https://kubernetes.io/pl/docs/reference/glossary/ (dostęp: 13.06.2024).

Mateusz Fesz, Dominika Fergisz, Maja Jaszowska, Filip Skawiński
Studenckie Koło Naukowe Informatyków „KOD”

dr. inż. Bartosz Trybus
Opiekun Koła Naukowego

Mechanizmy zarządzania pamięcią oraz synchronizacji wątków w języku Rust

Streszczenie

Artykuł przedstawia innowacyjne podejście do procesu zarządzania pamięcią operacyjną zaimplementowane w języku programowania Rust,. W artykule znalazła się charakterystyka rozwiązań takich jak „borrow checker” oraz „memory ownership” oraz ich wpływ na proces wytwarzania oprogramowania. Omówione zostały zalety oraz wady prezentowanego podejścia, wraz z porównaniem do innych popularnych schematów spotykanych w językach takich jak C czy Python.

Słowa kluczowe: rust, programowanie, synchronizacja wątków, pamięć, wzajemne wykluczanie, wyścigi

1. Wprowadzenie

Rust to wieloparadygmatowy język programowania stworzony przez zespół Mozilla, którego pierwsza stabilna wersja została wydana w 2015 roku. Jako jeden z najmłodszych języków, wyróżnia się na tle starszych technologii takich jak C, C++, a nawet Python i JavaScript, będąc jednocześnie w stanie z nimi konkurować zarówno pod względem wydajności, jak i funkcjonalności. Dynamicznie rozwijająca się społeczność programistów nieustannie tworzy nowe biblioteki, a istniejący ekosystem obejmuje już prawie 150 000 powszechnie dostępnych pakietów.

Rust jest najczęściej używany jako alternatywa dla C++ w programowaniu systemowym, skupiając się na tworzeniu wydajnego i niezawodnego oprogramowania, które rozwiązuje problemy związane z zarządzaniem pamięcią, reprezentacją danych oraz współbieżnością. Jego głównym celem jest wyeliminowanie typowych problemów napotykanych przez programistów w tej dziedzinie, oferując jednocześnie liczne wygodne i użyteczne narzędzia znane z innych środowisk. Rust umożliwia kontrolę nad kodem na poziomie charakterystycznym dla najstarszych, najbardziej wydajnych języków programowania, a jego wyróżniającą cechą jest unikalne podejście do zarządzania pamięcią.

2. Mechanizm zarządzania pamięcią

Rust został zaprojektowany tak, aby uniemożliwiać, a przynajmniej znacznie utrudniać programistom popełnianie klasycznych błędów związanych z niepoprawnym zarządzaniem pamięcią operacyjną. Oznacza to konieczność wyeliminowania takich problemów jak use-after-free, double-free czy indeksowanie poza wymiar tablicy, które są częstymi przyczynami awarii i podatności w programach napisanych w innych językach.

W odróżnieniu od większości współczesnych języków programowania, Rust nie korzysta z mechanizmu Garbage Collection (GC), który analizowałby pamięć w trakcie działania programu i usuwał nieużywane obiekty. GC jest powszechnie stosowany w językach takich jak Java, C#, Python czy JavaScript, jednak ma istotne wady. Największym problemem związanym z GC jest jego negatywny wpływ na wydajność programu, spowodowany koniecznością regularnego przerywania pracy aplikacji w celu skanowania pamięci. Te przerwy mogą być szczególnie problematyczne w aplikacjach wymagających wysokiej wydajności i niskich opóźnień, takich jak systemy czasu rzeczywistego, gry komputerowe czy serwery obsługujące duże ilości żądań.

Zamiast GC, Rust wprowadza mechanizm "memory ownership" (własności pamięci), który jest fundamentem jego podejścia do zarządzania pamięcią. Mechanizm ten opiera się na kilku kluczowych zasadach dotyczących każdej zmiennej w programie:

- Każda wartość musi mieć właściciela.
- Wartość może mieć tylko jednego właściciela w danym momencie.
- Kiedy właściciel wartości opuszcza aktualny zakres, wartość zostaje zwolniona.

Zakres (ang. scope) można najprościej określić jako obszar kodu, w którym dana zmienna jest poprawna. Rozpoczyna się on w momencie zdefiniowania zmiennej i trwa aż do końca bieżącego bloku kodu, w którym została zdefiniowana. Pojęcie to jest częściowo tożsame z odpowiednikiem znanym z języka C, obejmując takie elementy jak ciało funkcji, wyrażenia grupujące (odpowiednik instrukcji grupującej w C) czy wyrażenia lambda. Zmienna staje się poprawna w momencie wejścia w swój zakres i pozostaje taka aż do jego opuszczenia. Do tego momentu nie występują jeszcze znaczące różnice względem innych języków programowania.

Zaczynają one pojawiać się, gdy chcemy zapisać lub skopiować wartość typu przechowywanego na stercie (ang. Heap), np. String lub Box<T>. Ponieważ ich rozmiar nie

jest znany na etapie kompilacji (String jest dynamicznie rozrastającą się tablicą literałów UTF-8, zaś Box<T> jest jawnym wskaźnikiem na stertę), konieczne jest dynamiczne przydzielenie pamięci w trakcie działania programu. Niezbędne jest również uzyskanie możliwości zwrotu wcześniej zażądanego bloku pamięci do mechanizmu alokacyjnego (najczęściej dedykowanego API systemu operacyjnego) w momencie, gdy dana zmienna przestanie być wykorzystywana. W językach takich jak C czy C++, odpowiedzialność zarówno za alokację pamięci, jak i jej odpowiednie zwolnienie, spada na programistę.

W przypadku Rust mechanizm alokacji jest wywoływany niejawnie w momencie wywołania odpowiedniego konstruktora (w języku tym wszystkie konstruktory muszą być wywołane jawnie, np. String::new()). To zachowanie jest charakterystyczne dla większości współczesnych rozwiązań. Całkowicie zmienia się natomiast mechanizm zwalniania niewykorzystywanej już pamięci. W Rust odpowiedzialność za to przejmuje kompilator, który automatycznie zarządza zwalnianiem pamięci, gdy zmienna opuszcza swój zakres. Dzięki temu programiści nie muszą ręcznie zwalniać pamięci, co znacząco redukuje ryzyko wycieków pamięci i innych błędów związanych z zarządzaniem pamięcią.

W językach takich jak Python czy JavaScript, a także nowszych jak Go, mechanizm garbage collection automatycznie zarządza pamięcią, ale wiąże się to z dodatkowym zużyciem pamięci i czasu procesora. Mechanizm GC śledzi wykorzystanie zmiennych i usuwa z pamięci dane, do których dostęp nie będzie już możliwy. Ten proces, choć wygodny, wiąże się z kosztami w postaci większego zapotrzebowania na pamięć oraz obciążenia procesora, ponieważ GC musi regularnie przerywać wykonywanie programu, aby skanować pamięć i usuwać nieużywane obiekty. W niektórych sytuacjach konieczne jest nawet wyłączenie garbage collectora, aby zapewnić odpowiednią wydajność programu.

Klasyczne podejście, stosowane w językach takich jak C i C++, polega natomiast na ręcznym zwalnianiu pamięci w momencie, gdy dana wartość nie jest już potrzebna. W przypadku prostego kodu nie stanowi to problemu, jednak w miarę wzrostu rozmiaru projektów ryzyko zapomnienia o zwolnieniu pamięci znacznie wzrasta. Może to doprowadzić do wycieku pamięci, czyli niekontrolowanej alokacji coraz większej ilości pamięci na rzecz programu. W efekcie program zużywa więcej pamięci, niż realnie potrzebuje, co może prowadzić do całkowitego wyczerpania dostępnych zasobów. Co więcej, pomijając szereg obostrzeń, takich jak konieczność wskazania zawsze na początek bloku pamięci, nie można zwalniać tej samej pamięci wielokrotnie. W językach C i C++ takie zabronione operacje określane jako "Undefined Behavior" (zachowanie niezdefiniowane), co oznacza, że może prowadzić do nieprzewidywalnych rezultatów. Program może w takiej sytuacji zignorować operację, w

nieoczekiwany sposób naruszyć wartości zapisane w pamięci lub uruchomić funkcję formatującą dysk twardy. Choć ostatni przykład jest przerysowany, dobrze ilustruje, że zachowanie niezdefiniowane może prowadzić do poważnych problemów.

3. Memory Ownership

Mechanizm Ownership w Rust pełni funkcję automatyzacji procesu zwalniania pamięci, gdy właściciel danej wartości opuszcza swój zakres. Dzięki zasadom, według których każda wartość ma jednego właściciela, uzyskujemy pewność, że pamięć zostanie zwolniona w momencie opuszczenia zakresu przez właściciela. Dodatkowo, unikamy ryzyka ponownego wywołania funkcji zwalniającej pamięć, ponieważ nie istnieje żaden inny właściciel tej pamięci. W momencie zwolnienia pamięci, wywoływana jest funkcja drop dla danego typu, która pełni rolę destruktora, podobnie jak w innych językach programowania. Warto tutaj wspomnieć, że w C++ istnieje podobny wzorzec zarządzania zasobami, znany jako RAII (Resource Acquisition Is Initialization). Mechanizm ten ma znaczący wpływ na sposób pisania kodu w Rust i narzuca pewne zasady, które mogą początkowo wydawać się nieintuicyjne.

Domyślnym sposobem interakcji z wartością w Rust jest przeniesienie (ang. move), czyli zmiana właściciela danego obszaru pamięci bez faktycznego kopiowania danych. W momencie przeniesienia, poprzedni właściciel staje się niepoprawny i nie może być używany – każda próba jego użycia zakończy się błędem kompilatora z dokładnym wskazaniem miejsca, gdzie nastąpiło przeniesienie. Ten mechanizm umożliwia efektywne przekazywanie wartości między funkcjami poprzez czasową zmianę właściciela zmiennej, na przykład poprzez przekazanie wektora, a następnie jego zwrócenie jako wynik funkcji, zarówno w formie pierwotnej, jak i zmodyfikowanej. Chociaż mechanizm przenoszenia jest wygodny i eliminuje zbędne kopiowanie, w niektórych sytuacjach, zwłaszcza gdy funkcja przyjmuje wiele argumentów, może być niepraktyczny. Możliwe jest ręczne wykonanie głębokiej kopii, lecz jest to niewłaściwe rozwiązanie dla funkcji często modyfikujących duże zbiory danych.

Na listingu 1 zaprezentowana została przykładowa funkcja zawierająca prezentację działania operacji przenoszenia własności. Zmienna s1 utworzona jako łańcuch znakowy przypisywana jest do nowo tworzonej zmiennej s2. Próba odczytu s1 podczas wypisywania jej wartości na ekran oznaczałaby skorzystanie z pamięci której już nie posiada, zatem kompilator zwróci błąd, przedstawiony na listingu 2. Informuje on programistę że operacja taka nie jest dozwolona, wskazując przy tym nie tylko zmienną która powoduje problem, ale także w której części wartość została przeniesiona. Ten sam typ błędu wystąpiłby przy próbie wywołania funkcji do której s1 zostałoby przekazane jako parametr. Co więcej, dalsza część komunikatu

podaje jedno z możliwych, lecz zazwyczaj nieoptymalne rozwiązanie – jawne wykonanie głębokiej kopii.

```
fn hello(){
    let s1 = String::from("hello");
    let s2 = s1;

    println!("{s1}, world!");
}
```

Listing 1: Funkcja hello demonstrująca działanie mechanizmu ownership dla typu String

```
error[E0382]: borrow of moved value: `s1`
 --> src/main.rs:8:15
  |
5 |     let s1 = String::from("hello");
  |         -- move occurs because `s1` has type `String`, which does not
implement the `Copy` trait
6 |     let s2 = s1;
  |              -- value moved here
7 |
8 |     println!("{s1}, world!");
  |               ^^^^ value borrowed here after move
  |
  = note: this error originates in the macro `$crate::format_args_nl`
which comes from the expansion of the macro `println` (in Nightly builds,
run with -Z macro-backtrace for more info)
help: consider cloning the value if the performance cost is acceptable
  |
6 |     let s2 = s1.clone();
  |                +++++++
```

Listing 2: Fragment informacji o błędzie zwróconej dla kodu z listingu 1

Aby rozwiązać ten problem, Rust wprowadza drugi element mechanizmu Ownership – system Referencji oraz Pożyczania (ang. Borrowing). Referencja w Rust przypomina klasyczny wskaźnik na zmienną, ale z ważną różnicą: gwarantuje ona, że wskazuje na poprawną wartość konkretnego typu przez cały czas swojego życia. Dane mogą istnieć dłużej niż referencja, ale nie mogą być zwolnione z pamięci, dopóki istnieje do nich choć jedna referencja. Ta gwarancja jest kluczowa dla wielu innych pozytywnych aspektów języka. Referencje dzielą się na mutowalne (edytowalne) i niemutowalne (nieedytowalne).

Najważniejszą zasadą w mechanizmie Borrowing jest to, że w dowolnym momencie do konkretnej wartości w pamięci może istnieć wyłącznie jedna referencja mutowalna lub dowolna liczba referencji niemutowalnych. Zachowanie to jest egzekwowane przez mechanizm nazywany "Borrow Checker", który analizuje referencje na etapie kompilacji i zapobiega wykonaniu jakichkolwiek operacji naruszających te zasady. Dzięki temu Rust oferuje dalsze gwarancje bezpieczeństwa, zarówno w klasycznym kodzie proceduralnym, jak i w podejściu obiektowym, a przede wszystkim w implementacji współbieżności.

Borrow Checker w Rust zapewnia, że każda referencja jest zawsze poprawna i że żadna zmienna nie zostanie zmodyfikowana w sposób, który mógłby naruszyć integralność danych. W praktyce oznacza to, że możemy bezpiecznie pracować z danymi, wiedząc, że Rust automatycznie zapobiegnie typowym błędom związanym z zarządzaniem pamięcią, które mogą wystąpić w innych językach programowania.

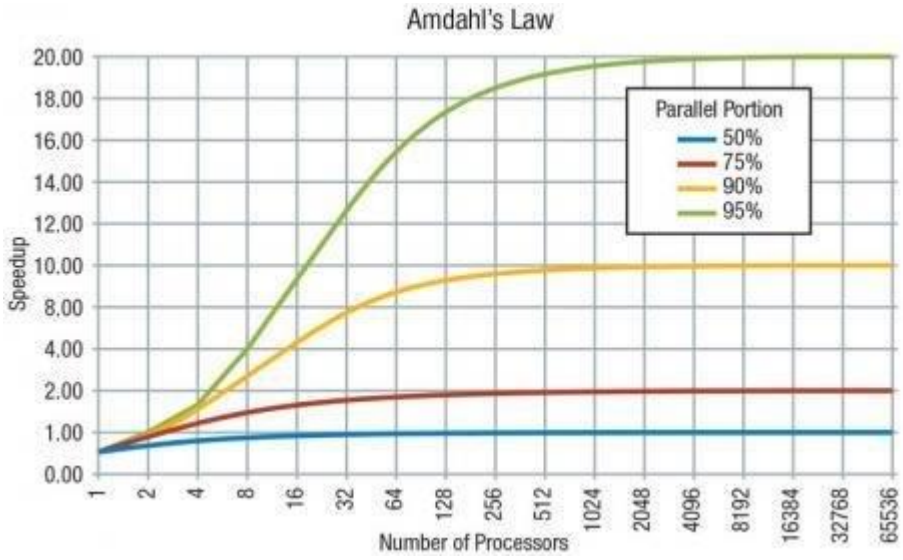
Mechanizm Borrowing pozwala na efektywne wykorzystanie pamięci bez nadmiernego kopiowania danych. Referencje mogą być przekazywane między funkcjami, co umożliwia manipulowanie danymi bez konieczności przenoszenia własności. Mutowalne referencje pozwalają na modyfikację danych, podczas gdy niemutowalne referencje zapewniają, że dane nie zostaną zmienione. To podejście pozwala na elastyczne zarządzanie danymi, dostosowane do różnych potrzeb programistycznych.

4. Wzajemne wykluczanie i synchronizacja wątków

Stworzenie wydajnego, współbieżnego kodu jest wyzwaniem nawet dla doświadczonych programistów. Wymaga to rozwiązania wielu problemów, w tym współdzielenia danych przez niezależne fragmenty kodu, wymiany danych między wątkami oraz dostępu do wspólnych obszarów pamięci przez wiele wątków jednocześnie. Do najczęściej spotykanych problemów należą:

- Wyścigi: sytuacje, gdy wiele wątków próbuje uzyskać dostęp do tych samych danych bez określonej kolejności.
- Zakleszczenia: sytuacje, gdy dwa wątki wzajemnie oczekują na zwolnienie zasobów przez siebie.
- Niejasne błędy: błędy występujące tylko w określonych okolicznościach, trudne do wyizolowania i naprawy.

Zgodnie z prawem Amdahla, przyspieszenie wynikające z podziału kodu na wiele równocześnie działających procedur jest proporcjonalne do udziału tych procedur w całym kodzie i ograniczone przez część, która musi być wykonywana sekwencyjnie. Wobec rosnącej liczby jednostek obliczeniowych w procesorach domowych oraz trendu zrównoleglania obliczeń chmurowych, ignorowanie tej ścieżki rozwoju oprogramowania jest niepraktyczne.



Rysunek 32. Graficzna reprezentacja prawa Amdahla. Górna granica wzrostu wydajności uzyskiwanej przez zrównoleglenie kodu jest dyktowana ilością operacji możliwych do zrównoleglenia [https://en.wikipedia.org/wiki/Amdahl%27s_law]

Istnieje wiele technik ułatwiających tworzenie bezpiecznego, równoległego kodu, takich jak semafory, przejmowanie zasobów na wyłączność czy operacje atomiczne. Problemy pojawiają się jednak, gdy zdamy sobie sprawę, że mechanizmy te są opcjonalne. W C++ możemy używać std::mutex dla zapewnienia wyłączności zasobów, ale zwykłe zmienne i wskaźniki są równie poprawne z punktu widzenia kompilatora. Python natomiast rezygnuje z wielowątkowości – pozwala na utworzenie wielu wątków, ale tylko jeden może być wykonywany w danym momencie. Alternatywą jest wieloprocesowość, gdzie tworzone są odrębne procesy, nie mogące współdzielić danych w większości systemów operacyjnych.

Rust oferuje pełne wsparcie dla mechanizmów wielowątkowości, zapewnianych przez systemy operacyjne, jednocześnie gwarantując bezpieczeństwo danych współdzielonych między wątkami. Mechanizm ownership w Rust blokuje jednoczesny mutowalny dostęp do danych z wielu wątków. Rodzi to pytanie, czy wątki mogą jedynie odczytywać dane spoza swojego zakresu? Odpowiedź brzmi: nie, dostęp mutowalny jest możliwy poprzez mechanizm zwany "mutowalnością wewnętrzną" (ang. Interior Mutability). Pozwala on na przekazywanie niemutowalnych referencji do obiektów synchronizacyjnych, które poprzez swoje wewnętrzne mechanizmy udostępniają funkcjom w wątku możliwość edytowania danych.

Na listingu 3 zaprezentowany został minimalny kod wykorzystujący Mutex. Struktura ta może przechowywać dowolny typ danych, w poniższym przykładzie jest to liczba całkowita. Dostęp do wartości odbywa się poprzez pozyskanie tzw. blokady otrzymywanej po wywołaniu metody lock(). Jeśli blokadę mutexu w danej chwili posiada inny wątek, metoda ta zatrzyma wykonywanie programu aż do odblokowania zmiennej. Dostęp jest możliwy do końca

aktualnego bloku kodu, a po jego osiągnięciu mutex zostaje automatycznie odblokowany – nie ma potrzeby wywoływania metod ręcznie.

```
use std::sync::Mutex;

fn main() {
    let m = Mutex::new(5);

    {
        let mut num = m.lock().unwrap();
        *num = 6;
    }

    println!("m = {m:?}");
}
```

Listing 3: Sposób użycia wzajemnego wykluczania przy wykorzystaniu struktur dostępnych w bibliotece standardowej.

W normalnych warunkach takie działanie naruszałoby gwarancje kompilatora, wywołując błąd kompilacji. Rust pozwala jednak na użycie "unsafe code", specjalnego typu bloku kodu, który umożliwia wykonywanie operacji naruszających te gwarancje:

- Wyłuskiwanie wartości ze wskaźnika.
- Wywoływanie funkcji oznaczonych jako unsafe.
- Uzyskiwanie mutowalnego dostępu do globalnej zmiennej statycznej.
- Implementacja niebezpiecznego interfejsu.
- Uzyskiwanie dostępu do pola obiektu typu union.

Te operacje, dostępne wewnątrz bloku unsafe, mogą prowadzić do "undefined behavior" lub problemów z synchronizacją, dlatego to na programiście spoczywa odpowiedzialność za zapewnienie dodatkowych mechanizmów zabezpieczających przed i po użyciu unsafe. Rust jednak nie wyłącza mechanizmu borrow checkera ani innych gwarancji języka. Mechanizmy zawarte w standardowej bibliotece Rust są bezpieczne i oferują API, które można używać bez ryzyka, a większość mechanizmów synchronizacji jest zaimplementowana jako typy otaczające, co pozwala na bezproblemowe korzystanie z nich w kodzie, jedynie z niewielkim narzutem wydajnościowym.

Dodatkowo, Rust posiada dwa specjalne interfejsy, służące do informowania kompilatora w jaki sposób dany typ danych może być obsługiwany w środowisku wielowątkowym. Nie posiadają one żadnych skojarzonych metod, zamiast tego stanowią jedynie marker informujący o dostępnych możliwościach. Pierwszy z nich, Send informuje że ownership danej zmiennej może zostać przeniesiony pomiędzy wątkami. Jest on domyślnie zaimplementowany dla

większości typów w bibliotece standardowej, poza konkretnymi wyjątkami, np. wskaźnikami. Jednym z nich jest licznik referencji, typ Rc<T>. W jego przypadku wykonanie operacji głębokiej kopii a następnie przesłanie jej do innego wątku wiązałoby się z istnieniem dwóch odniesień do tego samego licznika referencji. Bez dodatkowego mechanizmu synchronizacji, mogłoby dojść do jednoczesnego zwiększenia go przez oba wątki (na co pozwoliłby mechanizm mutowalności wewnętrznej i kod unsafe). Dzięki temu że Send nie jest zaimplementowane, przypadkowa próba przekazania tego typu wartości do innego wątku zakończy się błędem kompilacji – informującym o problemie. W przypadku licznika referencji można go łatwo naprawić stosując dedykowany w tym celu typ licznika atomowego – Arc<T>. Warto także wspomnieć, że jeśli dowolny typ złożony, np. struktura, będzie składać się wyłącznie z pól implementujących Send, to on sam zaimplementuje go automatycznie.

Drugi z opisywanych markerów, Sync, informuje że do zmiennej danego typu można bezpiecznie odnosić się w środowisku wielowątkowym, a zatem możliwe jest przekazywanie między wątkami referencji do niego. Analogicznym do implementacji Sync na typie T, byłoby zaimplementowanie Send na referencji &T. Przykładem takiej zmiennej jest wspomniany już wcześniej Mutex. Posiada on wbudowane mechanizmy które powodują że nie jest możliwy dostęp do zawartej w jego wnętrzu wartości, bez wykonania odpowiednich procedur.

Literatura

1. https://doc.rust-lang.org/book (dostęp 13.06.2024r.)
2. Jon Gjengset., *Rust for Rustaceans: Idiomatic Programming for Experienced Developers,* No Starch Press, 2021.
3. https://doc.rust-lang.org/rust-by-example/ (dostęp: 13.06.2024r.)
4. https://blog.tchatzigiannakis.com/undefined-behavior-can-literally-erase-your-hard-disk/ (dostęp 13.06.2024r.)
5. Shiang, Wong & Izzatdin, Abdul & Haron, Nazleeni & Jaafar, Jafreezal & Ismail, Norzatul & Mehat, Mazlina. (2016). The high performance linpack (HPL) benchmark evaluation on UTP high performance cluster computing. Jurnal Teknologi. 78. 21-30. 10.11113/jt.v78.9715.

Wiktor Kuczek, Katarzyna Maternia, Magdalena Matuła, Aleksandra Rokita, Aleksandra Sawicka
SKNI „Kod”

dr inż. Bartosz Trybus
Opiekun Koła Naukowego

Zastosowanie sztucznej inteligencji w grach komputerowych

Streszczenie

Artykuł omawia historię oraz rozwój sztucznej inteligencji w branży gier komputerowych. SI odgrywa kluczową rolę w dostarczaniu wyzwań, immersji i realistycznych doświadczeń dla graczy. Od sterowania zachowaniem przeciwników, przez generowanie proceduralnych światów, aż po systemy rankingowe - SI stała się nieodłącznym elementem nowoczesnych gier komputerowych. AI odpowiada za realistyczne zachowania NPC, prowadzenie dialogów i podejmowanie decyzji taktycznych w czasie rzeczywistym. Jest kluczowa w grach RPG, MMORPG i RTS. Służy również do tworzenia treści generuje unikalne światy, poziomy i wyposażenie za każdą rozgrywką, zwiększając replayability. Ponadto, technologie jak DLSS od NVIDII wykorzystują sztuczną inteligencję do skalowania obrazów, umożliwiając płynną rozgrywkę w wysokich rozdzielczościach. Wraz z rosnącą rolą AI pojawiają się jednak pytania dotyczące etyki, wpływu na graczy i prywatności danych.

Słowa kluczowe: Sztuczna inteligencja, AI, SI, NPC, immersja, gry rankingowe, RPG, RTS, DLSS.

1. Wprowadzenie

Sztuczna inteligencja (SI), z angielskiego Artificial Intelligence (AI), odgrywa kluczową rolę w branży gier komputerowych, przyczyniając się do tworzenia bardziej angażujących, realistycznych i innowacyjnych doświadczeń dla graczy. Od prostych algorytmów sterujących zachowaniem przeciwników w grach zręcznościowych, po skomplikowane systemy uczące się w czasie rzeczywistym - SI przekształca sposób, w jaki gry są projektowane oraz jak doświadczają ich gracze. Sztuczna inteligencja w grach komputerowych nie tylko zwiększa realizm, ale także wprowadza nowe możliwości w zakresie interakcji i personalizacji rozgrywki.

2. Historia sztucznej inteligencji w grach

Pierwsze próby wprowadzenia sztucznej inteligencji do gier komputerowych sięgają lat 50. XX wieku. W 1951 roku powstała gra “Nimrod”, która była jednym z pierwszych przykładów gry komputerowej wykorzystującej algorytmy do podejmowania decyzji. W latach 70. i 80. XX wieku, wraz z pojawieniem się pierwszych konsol i komputerów osobistych, zaczęły pojawiać się bardziej zaawansowane gry, takie jak “Pong” i “Space Invaders”, które również

wykorzystywały podstawowe formy SI. Inną z takich gier, o której należy wspomnieć jest “PacMan” z 1980 roku. Pomimo swojej prostoty, żółty bohater to tej pory może pochwalić się popularnością. Przyczynił się do tego poziom trudności narzucany przez duchy, które były sterowane przez SI, a ich zadaniem było dotrzeć do gracza i powstrzymać go w dążeniu do swojego celu. Jednak prawdziwy przełom nastąpił dopiero w latach 90., kiedy to gry takie jak “Doom” i “Warcraft” wprowadziły bardziej zaawansowane systemy SI, które umożliwiały przeciwnikom podejmowanie bardziej złożonych decyzji taktycznych. W kolejnych dekadach, rozwój mocy obliczeniowej oraz technik uczenia maszynowego pozwolił na tworzenie jeszcze bardziej realistycznych i złożonych systemów SI. Zaobserwować to można dzięki powstaniu na przełomie tysiącleci gry pod tytułem “The Sims’’, która była kamieniem milowym w historii gier z gatunku symulator.

3. Zastosowania sztucznej inteligencji w grach komputerowych

Sztuczną inteligencję, która odpowiada za dostarczenie dobrego doświadczenia dla gracza można podzielić na trzy główne kategorie:

- sztuczna inteligencja, która gra,
- sztuczna inteligencja, która odpowiada za tworzenie,
- tworzenie modeli za pomocą sztucznej inteligencji oraz uczenie jej.

3.1. Grająca sztuczna inteligencja

Ta kategoria AI ma również swoje podkategorie, z których każda na swój sposób wpływa znacząco na rozgrywkę oraz to jak ją odbiera gracz. Jednym z wyróżniających się tego typu systemów jest podział NPC (Non-Playable Character), na postacie, które są tłem gry, częścią świata przedstawionego lub takie, które stanowią przeszkodę dla gracza na drodze na jego celu.

Pierwszymi z nich są zwykli przechodnie, mieszkańcy wiosek czy miast, handlarze, sklepikarze i inni guślarze. Z reguły ich rolą jest “bycie”. Oznacza to, że stanowią część tła, aby odczuwał, że świat, do którego się przeniósł nie jest pusty i funkcjonuje autonomicznie. Pozwala to na stworzenie immersji, dzięki której gracz może poczuć się jakby też był jego częścią. Takie postacie są kluczową częścią gier z gatunku RPG (Role Playing Game) czy też MMORPG (Massively Multiplayer Online Role-Playing Game), które jak sama nazwa implikuje różnią się głównie tym, że ten drugi gatunek opiera się na rozgrywce z wieloma graczami. Zwykli NPC w takich grach reagują na działania gracza na różne sposoby, na przykład mogą skomentować jego działanie dialogiem tekstowym, który gracz może przeczytać

lub komputer zapamięta działanie gracza a konsekwencje tego czynu dotkną go znacznie później. W podobnych sytuacjach może zmieniać się ekonomia. Jeśli gracz zrobi coś moralnie nieetycznego w trakcie rozgrywki, handlarze mogą zwiększyć ceny a nawet odmówić wymiany. Przykładami gier znacząco opierających się na tego typu mechanikach sztucznej inteligencji są:

- **“Wiedźmin 3 - Dziki gon”** - Gdy gracz wcielający się w wiedźmina pomoże pewnej wiosce z wypędzeniem niechcianych potworów, zostanie nagrodzony, a następnie po powrocie do tej wioski wołany będzie mianem bohatera, w przeciwieństwie do poprzednich wyzwisk.
- **“Cyberpunk 2099”** - Działania głównego bohatera mają wpływ na świat go otaczający. Wpływa to na główny wątek fabularny, jak i jego zakończenie oraz na każdy aspekt rozgrywki, na przykład jak go traktują ludzie ze slumsów, miasta albo z półświatka.
- **Seria gier “Grand Theft Auto”** - W tej serii gier bardzo dobrze została zaimplementowana symulacja życia miejskiego. NPC prowadzą swoje życie, wykonując codzienne czynności, takie jak chodzenie do pracy, robienie zakupów czy odpoczywanie. Ich zachowania są sterowane przez zaawansowane systemy SI, które symulują realistyczne wzorce aktywności.

Innym rodzajem NPC są przeciwnicy, którzy sterowani przez sztuczną inteligencje są jednym z jej najbardziej podstawowych zastosowań w grach komputerowych. Algorytmy nimi sterujące są projektowane tak, aby reagować na działania gracza, podejmować decyzje taktyczne w czasie rzeczywistym lub w podziale turowym oraz dostosowywać swoją strategię w zależności od pozostałych czynników wpływających na rozgrywkę. Idealnym gatunkiem gier, które wykorzystują zaawansowane AI dla przeciwników jest RTS (Real Time Strategy). Gracz w nich nie ma czasu się zatrzymać i przeanalizować sytuację na polu bitwy, gdyż sztuczna inteligencja może to wykorzystać.

W serii gier **Warcraft**, zwłaszcza w “Warcraft III”, sztuczna inteligencja przeciwników odgrywa kluczową rolę w kształtowaniu rozgrywki. AI jest zaprogramowana, aby realizować różne cele i strategii, takie jak:

- **Zbieranie zasobów** - AI zarządza jednostkami do wydobywania złota i drewna, co jest podstawą ekonomii w grze,
- **Budowa bazy** - AI planuje i rozwija swoją bazę, budując struktury obronne, produkcyjne oraz technologiczne,

- **Produkcja jednostek** - AI tworzy jednostki wojskowe odpowiednio do sytuacji na polu bitwy, balansując pomiędzy ofensywnymi i defensywnymi strategiami,
- **Ataki i obrona** - AI analizuje sytuację na mapie, podejmując decyzje o ataku na wrogie bazy, obronie własnej oraz reakcjach na działania gracza. AI może prowadzić zwiad, organizować rajdy na zasoby przeciwnika i inicjować pełnowymiarowe ataki.

Analogicznie sytuacja wygląda w grach z serii **Warhammer**, szczególnie w strategiach takich jak **"Warhammer 40,000: Dawn of War"** oraz **"Total War: Warhammer"**, AI przeciwników jest kluczowym elementem, który wpływa na realizm i wyzwanie rozgrywki.

- **Zarządzanie armią** - AI kontroluje ruchy i formacje wojsk, podejmując decyzje strategiczne w czasie rzeczywistym. Na przykład, w **"Total War: Warhammer"**, sztuczna inteligencja zarządza dużymi armiami, biorąc pod uwagę jednostki piechoty, kawalerii i artylerii.
- **Strategia kampanii** - AI w grach Warhammer prowadzi kampanię, zarządzając zasobami, dyplomacją, rozbudową imperium oraz planowaniem długoterminowych strategii wojennych. AI potrafi zawierać sojusze, prowadzić wojnę, i realizować cele kampanii.
- **Taktyka na polu bitwy** - AI w czasie bitwy podejmuje decyzje dotyczące użycia jednostek, wykorzystania terenu oraz manewrów flankujących. W **"Dawn of War"**, AI stosuje różne taktyki ataku, obrony oraz używa specjalnych zdolności jednostek.

3.2. Generowanie zawartości

Proceduralne generowanie terenu i światów to technika wykorzystywana w grach komputerowych do tworzenia dużych, zróżnicowanych i dynamicznych środowisk bez potrzeby ręcznego projektowania każdego elementu przez twórców. Proces ten opiera się na algorytmach matematycznych i losowości, co pozwala na uzyskanie nieskończonej liczby unikalnych rezultatów. Poniżej przedstawiono szczegółowy opis tego, jak działa proceduralne generowanie terenu i światów, wraz z technikami i przykładami zastosowania. Przykłady obejmują:

- **"No Man's Sky" -** Gra ta wykorzystuje algorytmy generowania proceduralnego do tworzenia ogromnych, unikalnych wszechświatów. Każda planeta w grze jest generowana algorytmicznie, co oznacza, że każda rozgrywka może być inna.

- **“Minecraft” -** W tej grze proceduralne generowanie terenu tworzy unikalne krajobrazy dla każdego nowego świata. Gracze mogą eksplorować nieskończone, losowo generowane światy, co zapewnia niepowtarzalne doświadczenia w każdej sesji gry.
- **“Terraria” -** Powszechnie nazywana jako Minecraft 2D, posiada generowanie terenu analogiczne do tego z gry od Mojag. To i budowanie z klocków to jedyne rzeczy łączące obie gry.

W bliźniaczy sposób sztuczna inteligencja wykorzystywana jest w grach z gatunku **Roguelike**. Nazwa odnosi się do tytułu **“Rogue”**, który zapoczątkował wprowadzanie losowości do generowania pomieszczeń oraz rynsztunku, którym może wojować gracz.

- **“Spelunky”** - Gra ta wykorzystuje generowanie proceduralne do tworzenia unikalnych poziomów dla każdej nowej rozgrywki, co zwiększa jej replayability (możliwość wielokrotnego grania bez znudzenia),
- **“Deadcells”** - Generowane za każdym razem lokacje składają się z różnych planszy jak budynki, mosty, platformy. Za każdym razem gracz zaczyna z innymi broniami oraz może trafić na wszelkie ulepszenia podczas swojej rozgrywki, co skutkuje świeżością każdej rozgrywki.

3.3. Tworzenie modeli sztucznej inteligencji do rozgrywek rankingowych

Tworzenie modeli sztucznej inteligencji (SI) do rozgrywek rankingowych jest jednym z najbardziej zaawansowanych i wymagających zastosowań SI w branży gier komputerowych. Rozgrywki rankingowe to tryby gry, w których gracze rywalizują ze sobą o miejsce w rankingu, a ich umiejętności są oceniane na podstawie wyników meczów. Modele SI w takich systemach mają za zadanie zapewnić uczciwość, równowagę i dokładność w ocenie umiejętności graczy. Poniżej przedstawiono szczegółowy opis tworzenia i stosowania modeli SI w rozgrywkach rankingowych.

Kluczowymi elementami modeli AI w rozgrywkach rankingowych są:

Systemy Rankingowe:

- **Elo** - Jeden z najstarszych i najbardziej znanych systemów rankingowych, pierwotnie stworzony dla szachów. Elo ocenia graczy na podstawie ich wyników w meczach przeciwko innym graczom, dostosowując ich ranking w oparciu o wynik i oczekiwaną trudność przeciwnika.

- **TrueSkill** - System opracowany przez Microsoft do gier online, takich jak Halo i Gears of War. TrueSkill ocenia zarówno umiejętności graczy, jak i niepewność tych ocen, pozwalając na dokładniejsze dopasowanie przeciwników.
- **Glicko** - Rozszerzenie systemu Elo, które wprowadza ocenę wariancji umiejętności gracza, co pozwala na bardziej dynamiczne i precyzyjne dopasowania.

Algorytmy Dopasowywania (Matchmaking)

- Algorytmy te mają za zadanie tworzenie zrównoważonych drużyn i dobieranie przeciwników o podobnym poziomie umiejętności. Wykorzystują one modele SI do analizy rankingu, wyników i stylu gry graczy.

Analiza Danych Graczy

- Modele AI analizują ogromne ilości danych zebranych z gier, takich jak statystyki graczy, wyniki meczów, czas reakcji, decyzje strategiczne i wiele innych.
- **Uczenie nadzorowane**: Trening modeli SI na oznakowanych danych, gdzie wynik (wygrana/przegrana) jest znany, pomaga w przewidywaniu przyszłych wyników i ocenianiu umiejętności graczy.
- **Uczenie nienadzorowane**: Wykorzystywane do identyfikowania wzorców i grupowania graczy o podobnym stylu gry, co pomaga w bardziej precyzyjnym dopasowywaniu.

Systemy Anty-Cheatingowe

- Modele SI są również wykorzystywane do wykrywania i zapobiegania oszustwom w grach rankingowych. Analizują one zachowania graczy, wykrywając nietypowe wzorce, które mogą wskazywać na używanie nieuczciwych praktyk.
- **Analiza zachowań**: Monitorowanie i analiza nietypowych działań, takich jak nagłe skoki w umiejętnościach, nienaturalnie wysokie wskaźniki celności itp.
- **Uczenie głębokie**: Sieci neuronowe uczą się rozpoznawać subtelne wzorce oszustw na podstawie ogromnych zbiorów danych zebranych z rozgrywek.

3.4. Uczenie maszynowe i adaptacyjne sztucznej inteligencji

Uczenie maszynowe i adaptacyjne systemy SI stają się coraz bardziej popularne w branży gier. Techniki te pozwalają przeciwnikom i systemom SI na uczenie się i adaptowanie w czasie rzeczywistym, co prowadzi do bardziej wyrafinowanych i wymagających doświadczeń dla graczy. Przykłady obejmują:

- **“AlphaGo”** - Chociaż nie jest to gra komputerowa w tradycyjnym sensie, sukces AlphaGo w pokonaniu najlepszych graczy w Go pokazuje potencjał uczenia maszynowego w grach. Podobne techniki mogą być stosowane w grach komputerowych do doskonalenia strategii przeciwników.
- **“Middle-earth: Shadow of Mordor”** - Gra ta wykorzystuje system Nemesis, który pozwala wrogom zapamiętywać wcześniejsze spotkania z graczem i dostosowywać swoje zachowanie w przyszłych konfrontacjach. Dzięki temu każda interakcja z wrogiem staje się unikalna i dynamiczna.
- **“Dota 2”** - Boty w tej grze są trenowane za pomocą technik głębokiego uczenia, co pozwala im na konkurowanie z najlepszymi ludzkimi graczami. Uczenie maszynowe pozwala botom na ciągłe doskonalenie swoich umiejętności i strategii.

4. Techniki i algorytmy sztucznej inteligencji w grach komputerowych

4.1 Algorytmy Minimax i Alfa-Beta

Algorytmy Minimax oraz jego ulepszenie, Alfa-Beta, są często stosowane w strategiach turowych i grach planszowych, takich jak szachy. Algorytmy te pozwalają na analizowanie możliwych ruchów i wybranie optymalnego działania na podstawie przewidywanych odpowiedzi przeciwnika. Minimax ocenia wartość każdego możliwego ruchu poprzez symulację wielu kroków naprzód, podczas gdy Alfa-Beta wprowadza optymalizację, która eliminuje niepotrzebne gałęzie drzewa decyzyjnego, co zwiększa efektywność obliczeń.

4.2 Sieci neuronowe i głębokie uczenie

Sieci neuronowe i techniki głębokiego uczenia zyskują na popularności w grach, gdzie wymagane są skomplikowane decyzje i adaptacja. Głębokie sieci neuronowe są wykorzystywane do trenowania botów, które mogą analizować ogromne ilości danych z gier i uczyć się optymalnych strategii. Przykłady zastosowań obejmują:

- **Boty w “Dota 2”** - Wcześniej już wspomniane stworzone zostały przez OpenAI. Ich zadaniem jest wykorzystywanie głębokiego uczenia do analizy milionów gier i doskonalenia swoich umiejętności. Te boty są w stanie rywalizować z profesjonalnymi graczami, co pokazuje potencjał głębokiego uczenia w grach.

- **“StarCraft II”** - Algorytmy SI, takie jak AlphaStar od DeepMind, są trenowane przy użyciu głębokich sieci neuronowych do rywalizacji z ludzkimi graczami na najwyższym poziomie. AlphaStar jest w stanie analizować i adaptować się do strategii przeciwników, co czyni go niezwykle skutecznym.

4.3 Algorytmy genetyczne

Algorytmy genetyczne są stosowane do optymalizacji i tworzenia strategii w grach. Polegają one na mechanizmach ewolucji biologicznej, takich jak selekcja naturalna i mutacja, aby znaleźć optymalne rozwiązania w złożonych problemach. Przykłady obejmują:

- **“Galactic Civilizations III”** - Gra ta wykorzystuje algorytmy genetyczne do optymalizacji projektów statków kosmicznych, umożliwiając tworzenie bardziej efektywnych i potężnych jednostek.
- **“Creatures”** - W tej grze algorytmy genetyczne są używane do symulacji ewolucji stworzeń, które mogą uczyć się i adaptować w swoim środowisku, co prowadzi do unikalnych i zróżnicowanych zachowań.
- **“Civilization VI”** - wykorzystuje algorytmy genetyczne do optymalizacji projektów jednostek wojskowych i budynków. Algorytmy te naśladują procesy ewolucji biologicznej, takie jak selekcja naturalna i mutacja, aby znaleźć optymalne rozwiązania dla złożonych problemów projektowych.

5. Wpływ SI na doświadczenia graczy

SI znacząco wpływa na doświadczenia graczy, oferując bardziej złożone, dynamiczne i realistyczne interakcje. Gry takie jak “Red Dead Redemption 2” czy “Grand Theft Auto V” wykorzystują zaawansowane systemy SI, aby tworzyć immersyjne światy, w których gracze mogą swobodnie eksplorować i wchodzić w interakcje z otoczeniem. SI umożliwia tworzenie NPC, którzy reagują na zachowania gracza w realistyczny sposób, co zwiększa poczucie zanurzenia w wirtualnym świecie.

- **“Red Dead Redemption 2”** - NPC w tej grze prowadzą swoje życie niezależnie od działań gracza, co tworzy wrażenie żyjącego, dynamicznego świata. Gracze mogą obserwować, jak NPC wchodzą w interakcje ze sobą, reagują na zmieniające się warunki pogodowe i porę dnia.
- **“Grand Theft Auto V”** - Gra ta wykorzystuje zaawansowaną SI do sterowania ruchem ulicznym, zachowaniem przechodniów oraz interakcjami policji z graczami, co tworzy realistyczne i dynamiczne środowisko miejskie.

6. Wyzwania i przyszłość SI w grach

Chociaż SI w grach komputerowych osiągnęła już wiele, wciąż istnieje wiele wyzwań do pokonania. Jednym z nich jest zrównoważenie realizmu i wydajności, aby zapewnić płynne działanie gier na różnych platformach. Innym wyzwaniem jest tworzenie sztucznej inteligencji, która nie tylko jest inteligentna, ale również zabawna i wciągająca dla graczy. AI musi być w stanie dostarczać wyzwań, ale jednocześnie unikać frustrujących sytuacji, które mogłyby zniechęcić graczy.

Przyszłość SI w grach komputerowych wygląda obiecująco. Gracz może spodziewać się jeszcze bardziej zaawansowanych systemów uczących się, realistycznych symulacji oraz interaktywnych światów, które dostosowują się do działań gracza. Techniki takie jak głębokie uczenie i sieci neuronowe będą odgrywać kluczową rolę w dalszym rozwoju tej technologii. W miarę jak sprzęt komputerowy staje się coraz potężniejszy, możliwości sztucznej inteligencji w grach będą się tylko zwiększać.

7. Zastosowanie Sztucznej Inteligencji w DLSS i Narzędziach Ułatwiających Produkcję Gier

Sztuczna inteligencja odgrywa kluczową rolę nie tylko w rozgrywce, ale także w technologii poprawiającej jakość grafiki. Poniżej przedstawiono, jak SI jest wykorzystywana w technologii DLSS.

DLSS (Deep Learning Super Sampling) to technologia opracowana przez firmę NVIDIA, która wykorzystuje sztuczną inteligencję do poprawy wydajności grafiki w grach komputerowych. DLSS wykorzystuje sieci neuronowe do skalowania obrazów z niższej rozdzielczości do wyższej, co pozwala na uzyskanie wysokiej jakości grafiki przy niższym obciążeniu sprzętowym.

- **Jak działa DLSS** - Technologia DLSS wykorzystuje głębokie sieci neuronowe, które są trenowane na superkomputerach NVIDIA za pomocą ogromnych zestawów danych. Sieć neuronowa uczy się, jak przekształcać obrazy z niższej rozdzielczości w obrazy o wysokiej rozdzielczości, zachowując szczegóły i ostrość. W czasie rzeczywistym, DLSS stosuje te nauczone wzorce, aby poprawić jakość obrazu w grach bez znaczącego spadku wydajności,
- **Zalety DLSS** - Dzięki DLSS gracze mogą cieszyć się wyższą jakością grafiki, taką jak 4K, bez konieczności posiadania najnowszego i najdroższego sprzętu. DLSS pozwala

na zwiększenie liczby klatek na sekundę (FPS), co przekłada się na płynniejszą i bardziej responsywną rozgrywkę,

- **Przykłady zastosowania** - DLSS jest wykorzystywane w wielu nowoczesnych grach, takich jak “Cyberpunk 2077”, “Control”, “Death Stranding’, czy “Call of Duty: Warzone”. Dzięki DLSS te gry mogą działać płynnie w wysokich rozdzielczościach nawet na średniej klasy kartach graficznych.

8. Etyka i odpowiedzialność w projektowaniu SI

Wraz z rosnącą rolą SI w grach, pojawiają się również pytania dotyczące etyki i odpowiedzialności. Twórcy gier muszą zastanowić się nad wpływem swoich produktów na graczy, w tym na to, jak realistyczne i interaktywne systemy SI mogą wpływać na ich doświadczenia i zachowania. Istnieje ryzyko, że zbyt realistyczne SI mogłoby prowadzić do nieprzewidzianych konsekwencji, takich jak promowanie nieetycznych zachowań lub wywoływanie stresu u graczy bądź też uzależnienia.

Inną stroną tego problemu jest prywatność danych, otóż modele uczące w grach mogą w różnych przypadkach pobierać dane, których gracz mógłby sobie nie życzyć. Deweloperzy mogą zastrzegać, że takie incydenty nie mają prawa bytu, jednak w rzeczywistości sytuacja jest w stanie wyglądać zupełnie inaczej.

9. Podsumowanie

Artykuł podkreśla, że sztuczna inteligencja odgrywa kluczową rolę w kształtowaniu przyszłości gier komputerowych. Od sterowania przeciwnikami, przez generowanie zawartości, aż po symulację realistycznych zachowań - SI nieustannie zmienia sposób, w jaki doświadczamy gier. Pomimo licznych wyzwań, rozwój tej technologii niesie ze sobą ogromny potencjał, który z pewnością będzie kontynuowany w nadchodzących latach. W miarę jak technologia będzie się rozwijać, możemy spodziewać się jeszcze bardziej zaawansowanych i immersyjnych doświadczeń, które przekroczą nasze dzisiejsze wyobrażenia.

Ponadto, sztuczna inteligencja jest wykorzystywana w technologiach poprawiających jakość grafiki, takich jak DLSS od NVIDII, które skalują obrazy z niższej rozdzielczości, umożliwiając płynną rozgrywkę w wysokich rozdzielczościach.

Jednak pomimo wszystkich zalet stosowania sztucznej inteligencji w grach należy pamiętać, że wraz z rosnącą jej rolą w tej branży pojawiają się pytania dotyczące etyki, wpływu

na graczy, promowania nieetycznych zachowań oraz prywatności danych. Deweloperzy powinni mieć na uwadze ten problem i pracować nad jego rozwiązaniem.

Źródła internetowe:

1. https://en.wikipedia.org/wiki/Artificial_intelligence_in_video_games (dostęp: 14.06.2024)
2. https://appinventiv.com/blog/ai-in-gaming/ (dostęp: 14.06.2024)
3. https://sii.pl/blog/metody-sztucznej-inteligencji-w-grach-komputerowych/ (dostęp: 14.06.2024)
4. https://www.nvidia.com/pl-pl/geforce/technologies/dlss/ (dostęp: 14.06.2024)
5. https://developer.nvidia.com/rtx/dlss (dostęp: 14.06.2024)

Aleksandra Sawicka, Łukasz Książek, Katarzyna Maternia, Magdalena Matuła, Aleksandra Rokita
SKNI „Kod”

dr inż. Bartosz Trybus
Opiekun Koła Naukowego

Etyka w technologicznej rewolucji: Zagadnienia moralne w kontekście rozwoju sztucznej inteligencji

Streszczenie

W miarę postępu technologicznego, coraz bardziej zaawansowane systemy sztucznej inteligencji (SI) mogą stwarzać ryzyko naruszenia praw jednostek, dyskryminacji i pogłębiania nierówności społecznych. Dlatego kluczowe jest, aby rozwój SI uwzględniał nie tylko aspekty techniczne, lecz również etyczne i społeczne. Analiza różnych dziedzin etyki, takich jak etyka technologiczna, społeczna czy zawodowa, staje się niezbędna w takim procesie. Konieczne jest zrozumienie, jakie mogą być potencjalne skutki społeczne i jak uniknąć negatywnych konsekwencji. Warto więc podejmować środki zaradcze, takie jak zwiększanie transparentności algorytmów czy zapewnienie uczciwego dostępu do technologii, aby zapobiec ewentualnym szkodom dla społeczeństwa. Ostatecznie, równowaga między innowacjami technologicznymi a etycznymi aspektami ich wykorzystania jest kluczowa dla zapewnienia, że postęp technologiczny przynosi korzyści dla wszystkich, niezależnie od różnic społecznych.

Słowa kluczowe: sztuczna inteligencja, etyka, zasady społeczne, zagrożenia, programiści

1. Wprowadzenie

Sztuczna Inteligencja jest obecnie dziedziną rozwijającą się w tempie najszybszym. Dzieje się to głównie przez ogromne zgłębienie informatyki i elektroniki w ostatnich latach, szczególnie w i po czasach pandemii w 2019 roku. Dzięki temu rozwojowi posiadamy dostęp do różnych form sztucznej inteligencji nawet w naszym codziennym życiu, mimo że nie zawsze zdajemy sobie z tego sprawę. Wpadła ona w dziedziny takie jak medycyna, motoryzacja, marketing, handel czy bankowość. Tak jak bardzo jest to ułatwieniem naszego codziennego życia, tak również niesie to ze sobą negatywne aspekty, takie jak mniejsze zapotrzebowanie ludzi na rynku pracy w wielu obszarach, uzależnienie się od wspomagania maszyną czy też nawigacja naszego życia. Czym więc jest dla ludzkości sztuczna inteligencja? Czy jest to nosząca komfort rewolucja czy też zguba na drodze naszej moralności?

Człowiek od lat kierował się zasadami etyki które pomagały nam ustanawiać reguły życia codziennego, m.in. prawo. Jest ważne żeby nie zapominać o niej również w tym aspekcie tworzenia nowej dla człowieka rewolucji. Etyka odgrywa kluczową rolę w kształtowaniu odpowiedzialnego i zrównoważonego rozwoju technologii poprzez zapewnienie, że postępy technologiczne są podejmowane w sposób moralnie uprawniony i zgodny z wartościami społecznymi. Poprzez uwzględnianie aspektów etycznych, takich jak prywatność, uczciwość, bezpieczeństwo i równość,

możemy minimalizować negatywne skutki rozwoju technologicznego oraz maksymalizować korzyści dla społeczeństwa i środowiska. Bez uwzględnienia aspektów moralnych, postęp technologiczny może prowadzić do niekontrolowanych konsekwencji społecznych i indywidualnych. Brak odpowiedniej refleksji nad etyką w kontekście rozwoju technologicznego może doprowadzić do powstania naruszeń prywatności, wzrostu nierówności społecznych, utraty pracy, wykluczenia cyfrowego oraz powstania technologii wykorzystywanych do celów szkodliwych lub dyskryminujących. Społeczne zaufanie jest fundamentem, na którym opiera się akceptacja i adopcja nowych technologii przez społeczeństwo. Bez zaufania ludzie mogą być niechętni do korzystania z nowych technologii lub mogą czuć się narażeni na ich potencjalne negatywne skutki. Zachowanie zaufania społecznego wobec technologii ma kluczowe znaczenie dla zapewnienia ich trwałego sukcesu oraz pozytywnego wpływu na społeczeństwo jako całość. Dlatego też promowanie etycznych standardów powinno być priorytetem dla wszystkich zaangażowanych w rozwój i wdrażanie technologii. Celem tego artykułu jest identyfikacja kluczowych aspektów etycznych które należy uwzględnić w trakcie zarówno tworzenia jak i korzystania ze sztucznej inteligencji aby jej rozwój był zgodny z zasadami ludzkiej moralności tworzonej i modyfikowanej na przestrzeni lat. Kolejny aspekt godny poruszenia to zidentyfikowanie potencjalnych rozwiązań które mogą przyczynić się do etycznego poniesienia tej ogólnospołecznej odpowiedzialności którą ponosimy wszyscy korzystając z takich technologii jak sztuczna inteligencja.

2. Zasady etyczne w kontekście SI

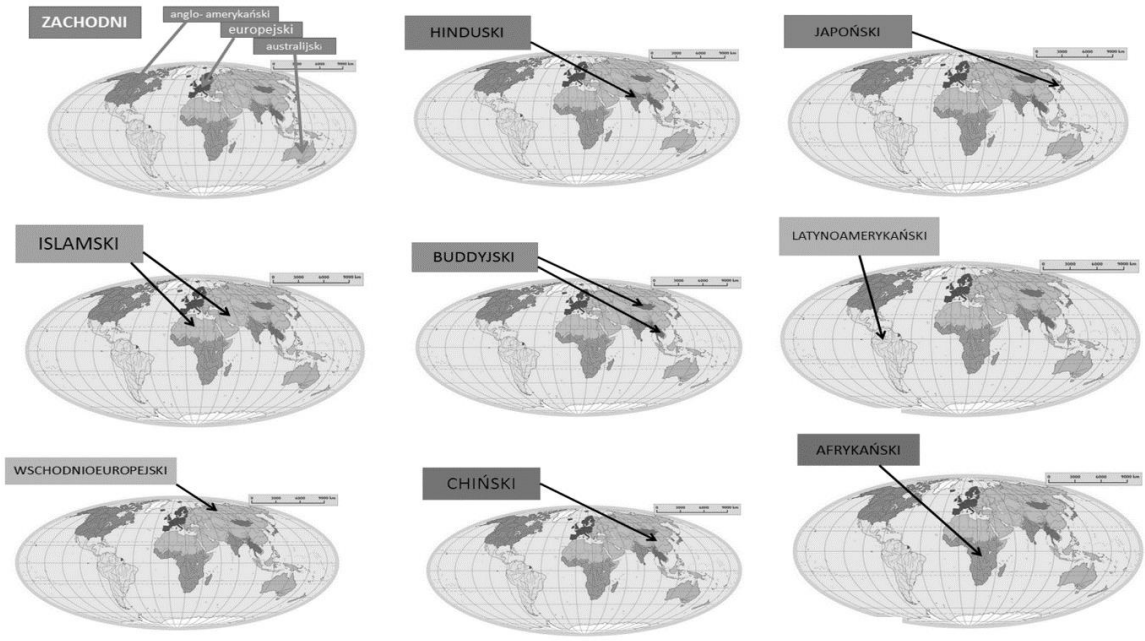
Istnieje wiele zasad etycznych które na przestrzeni lat zostały obrane przez ludzi jako ich kierunkowskazy postępowań społecznych. Etyka jako bardzo obszerna dziedzina składa się z wielu części. Autonomia jednostki jest kluczową zasadą etyczną, która zakłada, że każda osoba ma prawo do samostanowienia i podejmowania decyzji dotyczących swojego życia oraz swojej osoby. Jest to fundamentem szacunku dla godności jednostki oraz wyrazem poszanowania dla indywidualności i wolności każdej osoby. W praktyce zasada autonomii jednostki jest często balansowana z innymi zasadami etycznymi, takimi jak dobro wspólne czy sprawiedliwość społeczna. Ta druga wspomniana zakłada, że wszyscy ludzie powinni mieć równy dostęp do korzyści społecznych oraz zasobów, a także równy udział w obowiązkach i odpowiedzialnościach społecznych. W kontekście stosowania sztucznej inteligencji sprawiedliwość społeczna staje się kluczową kwestią, ponieważ SI (Sztuczna Inteligencja) może mieć wpływ na różnorodne aspekty życia społecznego i gospodarczego, a nierówności w dostępie do korzyści mogą prowadzić do pogłębiania się istniejących dysproporcji społecznych. Kolejną wartością jest Dobroczynność, która promuje działania mające na celu maksymalizowanie dobra społecznego poprzez pomoc innym ludziom oraz wsparcie dla potrzebujących. Z samej definicji można wywnioskować jak kluczowe jest zachowanie tej wartości tworząc kolejne postępy w Sztucznej Inteligencji. Jest to jednak tylko część tego co powinniśmy uwzględnić.

Zapewnienie zgodności z wartościami etycznymi podczas projektowania sztucznej inteligencji stanowi złożone wyzwanie, które wymaga holistycznego podejścia do procesu tworzenia i wdrażania

systemów SI. W praktyce oznacza to, że programiści i inżynierowie muszą mieć świadomość etycznych implikacji swoich decyzji programistycznych. To wymaga głębokiego zrozumienia kontekstu społecznego i kulturowego, w którym będą działać te systemy. Konieczne jest przeprowadzanie analiz, aby zrozumieć, jakie mogą być konsekwencje dla różnych grup społecznych oraz czy istnieje ryzyko naruszania praw człowieka czy powstawania nierówności. Odpowiedzialność za decyzje algorytmów jest niezbędna. Systemy SI muszą być zaprojektowane w taki sposób, aby można było prześledzić procesy decyzyjne i zrozumieć, jakie czynniki wpływają na podejmowane decyzje. To umożliwia identyfikację potencjalnych błędów oraz podejmowanie odpowiednich działań naprawczych.

Aby zapewnić wspomnianą zgodność wymagane jest również uwzględnienia różnorodności społecznej i kulturowej. Zrozumienie różnych perspektyw, potrzeb i wartości społecznych jest kluczowe, aby uniknąć nierówności i dyskryminacji. W kontekście projektowania sztucznej inteligencji, uwzględnienie różnorodności społecznej oznacza, że programiści muszą rozumieć, jak różne grupy społeczne mogą być różnorodnie reprezentowane lub dotknięte przez działania systemów SI. Wdrażanie systemów SI powinno uwzględniać te różnice i zapewnić, że narzędzia i interfejsy są dostosowane do różnorodności użytkowników, biorąc pod uwagę różnice językowe, kulturowe czy umiejętności techniczne. Poprzez uwzględnienie różnorodności społecznej i kulturowej w procesie projektowania i implementacji systemów SI można przyczynić się do tworzenia bardziej sprawiedliwych technologii, które służą wszystkim użytkownikom w sposób odpowiedni i zgodny z ich potrzebami oraz wartościami społecznymi.

Rysunek poniżej pozwoli nam lepiej zobrazować, jak bardzo zróżnicowane jest nasze społeczeństwo aby wiedzieć jak bardzo ważna jest branie ich pod uwagę programując.



Rysunek 33 Zróżnicowanie kulturowe.
Źródło: https://www.naukowiec.org/wiedza/geografia/kregi-kulturowe-swiata_3215.html

Kolejnym kluczowym aspektem w radzeniu sobie z etycznymi wyzwaniami w projektowaniu sztucznej inteligencji jest współpraca z ekspertami z różnych dziedzin. Etycy mogą pomóc w identyfikacji potencjalnych zagrożeń dla wartości moralnych i zasad etycznych, które mogą wynikać z konkretnych zastosowań SI. Filozofowie mogą przyczynić się do zrozumienia głębszych implikacji etycznych i filozoficznych związanych z rozwojem technologii SI. Ekspertów nauk społecznych można zaangażować w analizę wpływu sztucznej inteligencji na społeczeństwo, w tym na struktury społeczne, relacje międzyludzkie i równość. Ich wiedza może pomóc w zrozumieniu, jakie mogą być skutki społeczne i jak uniknąć negatywnych konsekwencji. Specjaliści prawa mogą z kolei wspomóc w opracowywaniu odpowiednich ram regulacyjnych, które mogą zapobiec nadużyciom, zapewnić ochronę praw człowieka oraz zagwarantować zgodność z obowiązującymi przepisami prawnymi. Współpraca z ekspertami z różnych dziedzin może również przyczynić się do edukacji społeczeństwa na temat etycznych i społecznych aspektów sztucznej inteligencji. Poprzez wspólne działania, takie jak organizowanie konferencji, warsztatów czy publikowanie artykułów naukowych, eksperci mogą promować świadomość i zrozumienie tych kwestii w szerszym gronie odbiorców.

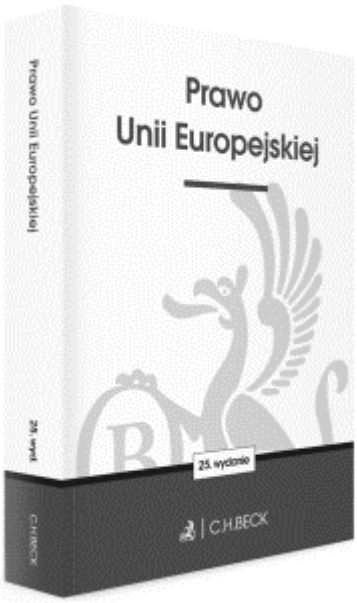
3. Kwestie odpowiedzialności

W obliczu szybkiego rozwoju sztucznej inteligencji (SI) i jej coraz szerszego zastosowania, kwestie identyfikacji odpowiedzialności stają się niezmiernie istotne. Jednym z kluczowych wyzwań jest niejasność w ustalaniu, kto właściwie ponosi odpowiedzialność za działania podejmowane przez systemy SI. Czy jest to ich twórca, użytkownik, czy może same algorytmy? Brak jednoznacznej odpowiedzi na to pytanie stwarza poważne problemy zarówno w kontekście etycznym, jak i prawnym. W dzisiejszym świecie, gdzie algorytmy SI są coraz bardziej zaawansowane i autonomiczne, tradycyjne modele odpowiedzialności stają się niewystarczające. Pojawia się więc konieczność zdefiniowania nowych ram prawnych i normatywnych, które jasno określą zakres i sposób odpowiedzialności za działania SI. Jest to istotne nie tylko dla ochrony użytkowników, ale także dla zapewnienia zgodności z wartościami społecznymi i etycznymi.

Skutki braku jasno określonej odpowiedzialności mogą być poważne i dotknąć zarówno jednostki, jak i całe społeczeństwo. Bez jasno określonej odpowiedzialności istnieje także ryzyko nadużyć, dyskryminacji czy naruszenia praw człowieka przez systemy SI. Aby zapobiec takim negatywnym konsekwencjom, konieczne jest opracowanie skutecznych mechanizmów monitorowania i oceny działań SI. Takie mechanizmy powinny umożliwić śledzenie działań systemów SI, identyfikację potencjalnych zagrożeń oraz szybką interwencję w przypadku wystąpienia nieprawidłowości. Wdrażanie tych mechanizmów będzie kluczowe dla zapewnienia odpowiedzialnego i bezpiecznego użytkowania sztucznej inteligencji.

Wreszcie, w kontekście identyfikacji odpowiedzialności, istotne jest również uwzględnienie aspektu edukacyjnego. Użytkownicy, twórcy i decydenci muszą być świadomi swoich obowiązków i roli w procesie stosowania SI. Edukacja na temat etycznych i prawnych aspektów sztucznej inteligencji może

przyczynić się do budowy świadomego społeczeństwa, które potrafi odpowiedzialnie korzystać z nowych technologii. W związku z powyższym, jasne określenie odpowiedzialności oraz implementacja skutecznych mechanizmów monitorowania i oceny działania systemów SI są kluczowe dla zapewnienia, że sztuczna inteligencja będzie służyć społeczeństwu w sposób bezpieczny, sprawiedliwy i zgodny z jego wartościami.



Rysunek 2 Prawo Unii Europejskiej jako przykładowy kodeks do przystosowania się
Źródło: https://www.ksiegarnia.beck.pl/21033-prawo-unii-europejskiej

4. Bezpieczeństwo i prywatność danych

W kontekście rozwijających się systemów sztucznej inteligencji (SI) i zaawansowanych algorytmów, bezpieczeństwo danych oraz ochrona prywatności stają się coraz istotniejsze. Zagrożenia związane z bezpieczeństwem danych mogą prowadzić do poważnych konsekwencji, w tym naruszenia prywatności użytkowników oraz potencjalnych luk w zabezpieczeniach. Rosnąca ilość przechowywanych informacji stwarza ryzyko naruszenia prywatności i bezpieczeństwa danych użytkowników. Konieczne jest zatem zapewnienie bezpiecznego przechowywania, przetwarzania i transmisji danych w systemach SI. Mechanizmy anonimizacji i pseudonimizacji danych mogą stanowić skuteczną ochronę prywatności użytkowników, jednocześnie umożliwiając wykorzystanie danych w celach badawczych i analitycznych. Transparencyjne polityki prywatności oraz procedury zarządzania danymi są kluczowe dla budowania zaufania użytkowników do systemów SI. Konieczne jest promowanie przejrzystości w zakresie zbieranych danych, sposobu ich przetwarzania oraz praw użytkowników do kontroli nad swoimi danymi.

5. Uprzedzenia i dyskryminacja

Problem uprzedzeń w algorytmach sztucznej inteligencji jest istotny, może prowadzić do nierównego traktowania użytkowników na podstawie różnorodnych cech, takich jak rasa, płeć, wiek czy orientacja seksualna. Przykłady takich uprzedzeń można znaleźć w systemach rekrutacyjnych czy oceny kredytowej, gdzie algorytmy wykazują tendencję do faworyzowania pewnych grup i dyskryminacji innych. Aby zapobiec dyskryminacji w systemach SI, istnieją różne propozycje metodologii, w tym podejście fair AI, które ma na celu eliminację uprzedzeń z algorytmów i zapewnienie uczciwego traktowania wszystkich użytkowników.

Rysunek 3 Wskaźnik równouprawnienia płci
Źródło: https://pl.wikipedia.org/wiki/Wska%C5%BAnik_r%C3%B3wnouprawnienia_p%C5%82ci

Metody te mogą obejmować zastosowanie technik fairness-aware machine learning oraz aktywną kontrolę nad danymi treningowymi, aby uniknąć wprowadzania uprzedzeń. Jednakże, oprócz działań na poziomie technologicznym, istotną rolę w zapobieganiu dyskryminacji odgrywają również regulacje prawne. Regulacje te mogą wymuszać sprawiedliwe standardy w systemach SI oraz narzucać odpowiednie procedury audytów, które mają na celu identyfikację i eliminację uprzedzeń. Ponadto,

regulacje prawne mogą stworzyć ramy prawne dla odpowiedzialności za skutki działań systemów SI, co dodatkowo wspiera walkę z dyskryminacją. W ten sposób, łącząc działania na poziomie technologicznym z odpowiednimi regulacjami prawny, można skutecznie radzić sobie z problemem uprzedzeń i dyskryminacji w systemach sztucznej inteligencji, tworząc środowisko, które zapewnia uczciwe i równomierne traktowanie wszystkich użytkowników. Rysunek 3 pokazuje jak dalej zmagamy się z nierównością płci co jest przykładem dyskryminacji.

6. Transparentność i przejrzystość

Wartość transparentności w działaniu algorytmów sztucznej inteligencji (AI) jest niezwykle istotna dla użytkowników i społeczności. Udostępnianie informacji na temat sposobów podejmowania decyzji przez systemy SI pozwala użytkownikom lepiej zrozumieć, jak działają te systemy oraz jakie kryteria są brane pod uwagę przy podejmowaniu decyzji. To z kolei buduje zaufanie społeczne i umożliwia ocenę etyczności działań SI. Transparentne i zrozumiałe algorytmy przynoszą wiele korzyści, między innymi umożliwiają użytkownikom śledzenie procesu podejmowania decyzji oraz sprawdzanie, czy decyzje podejmowane przez systemy SI są zgodne z ich oczekiwaniami. Ponadto, budując zaufanie społeczne, transparentność może przyczynić się do większej akceptacji i adopcji technologii AI.

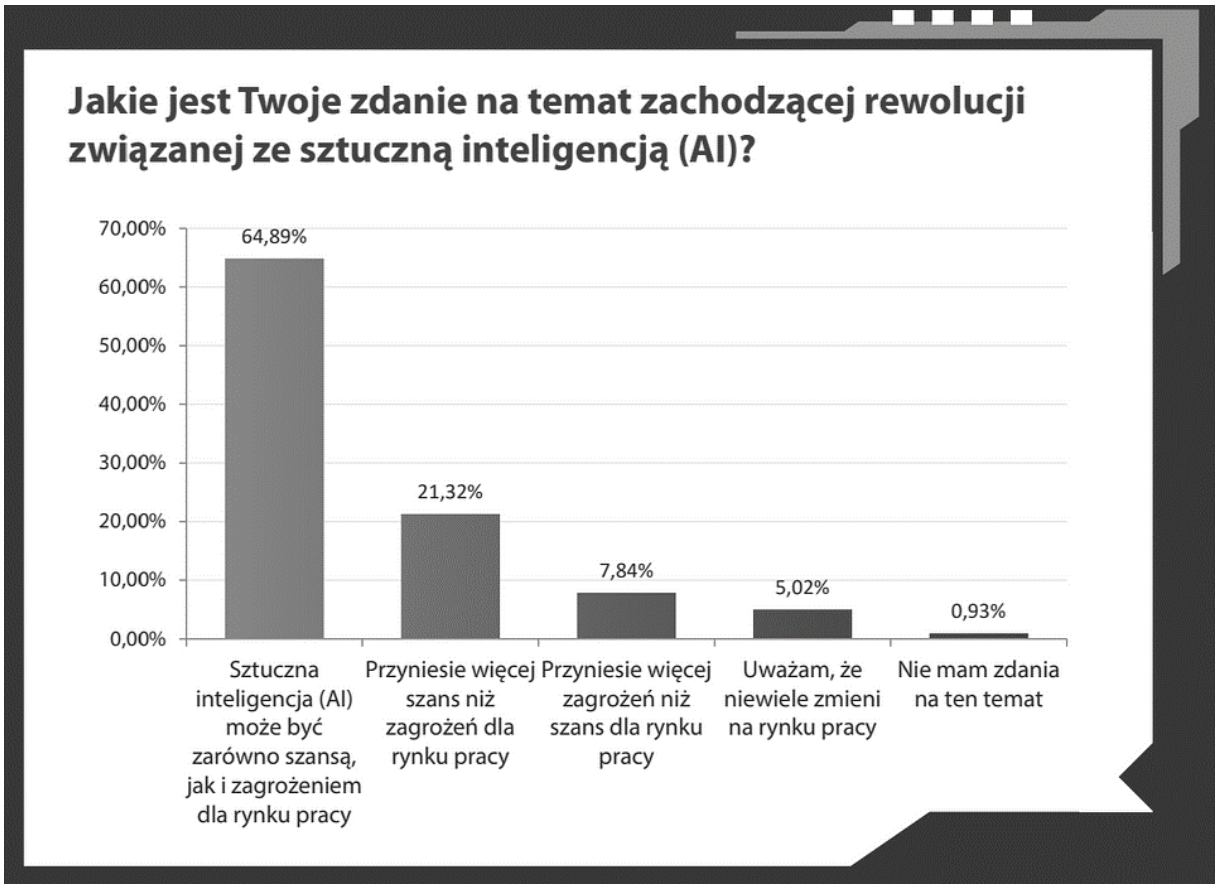
Zapewnienie przejrzystości algorytmów SI może być osiągnięte poprzez różne metody i techniki. Jedną z nich jest interpretowalność modeli, czyli umożliwienie zrozumienia, jak algorytmy dochodzą do swoich wniosków. Dodatkowo, udostępnianie otwartych danych i kodu źródłowego umożliwia społeczności analizę oraz ocenę działania algorytmów. No i oczywiście, audytowane procesy decyzyjne pozwalają użytkownikom śledzić, jakie kryteria są brane pod uwagę przy podejmowaniu decyzji przez systemy SI. Wprowadzenie transparentności w zaawansowanych systemach uczenia maszynowego, takich jak sieci neuronowe głębokiego uczenia, może jednak stanowić wyzwanie. Ze względu na złożoność tych algorytmów, interpretowanie ich działania może być trudne. Jednak rozwój technik interpretowalności oraz otwarty dialog między twórcami algorytmów a społecznością mogą przyczynić się do zwiększenia przejrzystości tych systemów.

Podsumowując, zapewnienie przejrzystości działania algorytmów AI jest kluczowe dla budowania zaufania społecznego oraz umożliwienia użytkownikom oceny etyczności działań SI. Dążenie do transparentności powinno być integralną częścią procesu projektowania i wdrażania systemów SI, co przyczyni się do bardziej zrównoważonego i odpowiedzialnego rozwoju sztucznej inteligencji.

7. Wpływ na rynki pracy

Rozwój sztucznej inteligencji i automatyzacja mają znaczący wpływ na rynki pracy, co skutkuje zmianami w strukturze zatrudnienia i potencjalnymi zagrożeniami dla różnych sektorów i zawodów. Prognozy dotyczące automatyzacji wskazują na to, że niektóre rutynowe i powtarzalne zadania mogą zostać zautomatyzowane, co może prowadzić do redukcji miejsc pracy w tych obszarach. Jednak

równocześnie innowacje technologiczne mogą prowadzić do powstania nowych miejsc pracy, zwłaszcza w sektorach związanych z technologią, badaniami naukowymi i usługami cyfrowymi. Rysunek poniżej pokazuje zdanie na temat sztucznej iteligencji na rynku pracy pytanych w ankiecie Polaków.



Rysunek 4 Sztuczna inteligencja a rynek pracy ankieta wśród Polaków
Źródło: https://media.justjoin.it/266966-sztuczna-inteligencja-a-rynek-pracy

Przykłady takich innowacji obejmują rozwój sztucznej inteligencji w medycynie, gdzie systemy SI wspierają diagnostykę i leczenie chorób, tworząc nowe możliwości zatrudnienia dla specjalistów medycznych i inżynierów biomedycznych. Jednakże, istnieją także obszary, w których automatyzacja może prowadzić do znaczącej redukcji miejsc pracy, np. w sektorze usług finansowych czy produkcji przemysłowej. W obliczu tych zmian, istotne jest opracowanie strategii adaptacji społecznej, które pomogą społeczeństwu przystosować się do zmian w rynku pracy spowodowanych przez rozwój SI. Reorientacja zawodowa i rozwój umiejętności cyfrowych są kluczowymi elementami tych strategii.

Elastyczne modele pracy, takie jak praca zdalna czy umowy na czas określony, mogą także umożliwić dostosowanie się do zmieniających się warunków na rynku pracy. Ponadto, wsparcie ze strony rządu i instytucji edukacyjnych jest istotne dla zapewnienia, że społeczeństwo posiada niezbędne umiejętności i kompetencje do korzystania z nowych możliwości zatrudnienia. Programy szkoleniowe i kursy specjalistyczne mogą pomóc pracownikom w zdobyciu nowych umiejętności wymaganych na zmieniającym się rynku pracy.

W rezultacie, rozwój sztucznej inteligencji niesie ze sobą zarówno wyzwania, jak i możliwości dla rynków pracy. Wdrażanie odpowiednich strategii adaptacji społecznej jest kluczowe dla zapewnienia, że społeczeństwo może skorzystać z potencjału rozwoju SI, jednocześnie minimalizując negatywne skutki dla zatrudnienia i równości społecznej.

8. Regulacje i ramy prawne

Obecnie istnieje zróżnicowanie w zakresie regulacji dotyczących sztucznej inteligencji (SI) na poziomie międzynarodowym, krajowym oraz regionalnym. W różnych jurysdykcjach obowiązują przepisy dotyczące ochrony danych osobowych, odpowiedzialności za produkty oraz bezpieczeństwa technologicznego, które mają zastosowanie do dziedziny SI. Niemniej jednak, szybki postęp technologiczny i pojawianie się nowych wyzwań etycznych związanych z SI stawiają pod znakiem zapytania adekwatność istniejących regulacji. W odpowiedzi na te wyzwania, pojawiają się propozycje nowych regulacji i standardów etycznych, które mają na celu promowanie odpowiedzialnego i etycznego stosowania SI.

Jest to istotne zwłaszcza w kontekście rosnącej liczby zastosowań SI, które mogą mieć istotne konsekwencje dla jednostek i społeczeństw. Europejska Strategia Sztucznej Inteligencji oraz inicjatywy podejmowane na szczeblu krajowym i międzynarodowym wskazują na potrzebę stworzenia spójnych ram prawnych, które będą zarówno chronić prawa jednostek, jak i promować innowacje w dziedzinie SI. Propozycje regulacyjne skupiają się na różnych aspektach, takich jak zwiększenie przejrzystości działań algorytmów, zapewnienie uczciwego traktowania użytkowników oraz odpowiedzialności za skutki działań SI. Ponadto, standardy etyczne są coraz częściej przedmiotem dyskusji, co może prowadzić do powstania bardziej zharmonizowanych podejść do etycznego projektowania i wykorzystywania SI.

Wdrażanie nowych regulacji i standardów etycznych wymaga jednak uwzględnienia różnorodnych perspektyw i interesów, zarówno ze strony przemysłu, jak i społeczeństwa. Istotne jest, aby regulacje te były elastyczne i dostosowane do szybko zmieniającego się środowiska technologicznego, jednocześnie zachowując wysoki poziom ochrony praw jednostek i społeczeństwa. Dążenie do stworzenia spójnych i adekwatnych ram prawnych może przyczynić się do budowy zaufania społecznego wobec sztucznej inteligencji oraz promowania jej zrównoważonego i odpowiedzialnego rozwoju.

9. Edukacja i świadomość społeczna

Edukacja społeczna na temat sztucznej inteligencji odgrywa kluczową rolę w promowaniu świadomości społecznej na temat jej zastosowań, korzyści i zagrożeń. Wzrost świadomości społecznej może przyczynić się do lepszego zrozumienia potencjalnych skutków wprowadzenia SI w różnych dziedzinach życia oraz do akceptacji nowych technologii. Edukacja w zakresie SI powinna być

realizowana na różnych poziomach, począwszy od szkół podstawowych i średnich, gdzie uczniowie mogą zdobyć podstawową wiedzę na temat działania SI oraz sposobów jej wykorzystania.

Niebagatelne znaczenie ma również edukacja wśród pracowników i społeczności lokalnych. Programy szkoleniowe i warsztaty mogą pomóc pracownikom w zrozumieniu zmian, jakie niesie ze sobą wprowadzenie SI w miejscu pracy oraz w zdobyciu niezbędnych umiejętności do efektywnego korzystania z nowych technologii. Kształtowanie postaw społecznych sprzyjających etycznemu stosowaniu SI jest również istotnym aspektem edukacji społecznej. Promowanie wartości takich jak uczciwość, odpowiedzialność i szacunek dla godności ludzkiej może przyczynić się do wypracowania społeczeństwa, które podejmuje świadome i etyczne decyzje dotyczące wykorzystania SI. Wdrażanie programów edukacyjnych i promowanie świadomości społecznej na temat SI wymaga współpracy między sektorem publicznym, prywatnym i społeczeństwem obywatelskim. Istotne jest również uwzględnienie różnorodnych perspektyw i doświadczeń społecznych, aby programy edukacyjne były dostosowane do potrzeb i oczekiwań różnych grup społecznych.

W rezultacie, edukacja społeczna na temat SI jest kluczowym czynnikiem promującym zrównoważony i odpowiedzialny rozwój technologiczny. Działania w tym zakresie mogą przyczynić się do budowy społeczeństwa, które aktywnie uczestniczy w dyskusjach dotyczących sztucznej inteligencji i podejmuje świadome decyzje związane z jej wykorzystaniem.

10. Podsumowanie

Podczas gdy sztuczna inteligencja rozwija się w zastraszającym tempie, nie możemy ignorować towarzyszących jej wyzwań etycznych. Stajemy przed koniecznością rozważenia kwestii odpowiedzialności, bezpieczeństwa danych, eliminacji uprzedzeń oraz transparentności. Należy ustanowić zasady etyczne, które będą kierować naszą interakcją z sztuczną inteligencją. Te zasady muszą być elastyczne i gotowe dostosować się do dynamicznie zmieniającej się technologii. Ważne jest również, abyśmy opracowali ramy prawne, które będą chronić prawa jednostek, zapewniając jednocześnie, że postęp technologiczny nie jest ograniczany. Bezpieczeństwo danych to kolejny aspekt, który wymaga naszej uwagi. Wraz z gromadzeniem coraz większych ilości danych istnieje ryzyko ich nadużywania lub nieuprawnionego dostępu. Konieczne jest więc opracowanie środków ochrony danych, które będą skuteczne, ale jednocześnie nieograniczające innowacji. Sztuczna inteligencja może odzwierciedlać uprzedzenia obecne w danych, co może prowadzić do niesprawiedliwych decyzji. Wprowadzenie mechanizmów zapobiegających temu jest niezbędne, abyśmy mogli korzystać z sztucznej inteligencji w sposób sprawiedliwy i równy dla wszystkich. Wreszcie, ludzie muszą rozumieć, jak działa sztuczna inteligencja i jakie są jej potencjalne konsekwencje. Tylko wtedy będą mogli podejmować świadome decyzje dotyczące jej wykorzystania. Wnioskiem jest to, że rozwój sztucznej inteligencji musi iść w parze z rozwojem zasad etycznych i ram prawnych. Tylko wtedy będziemy mogli cieszyć się korzyściami płynącymi z tej technologii, nie narażając jednocześnie naszych wartości i praw.

Źródła internetowe:

1. https://pl.wikipedia.org/wiki/Sztuczna_inteligencja(dostęp: 24.04.2024).
2. https://www.gov.pl/web/ai/pierwsze-w-jezyku-polskim-popularnonaukowe-wprowadzenie-do-etyki-sztucznej-inteligencji(dostęp: 24.04.2024).
3. https://web.swps.pl/strefa-psyche/blog/relacje/22399-etyczna-sztuczna-inteligencja-czy-etyke-mozna-zaprogramowac?dt=1714922278674(dostęp: 24.04.2024).
4. https://www.gov.pl/web/ai/etyka-ai-okiem-ekspertow-dostrzegac-kontekst-myslec-krytycznie(dostęp: 24.04.2024).
5. https://studiamba.merito.pl/baza-wiedzy/sztuczna-inteligencja-w-zyciu-codziennym-praktyczne-zastosowania(dostęp: 24.04.2024).

Aleksandra Sawicka, Katarzyna Maternia, Magdalena Matuła, Aleksandra Rokita, Wiktor Kuczek
SKNI „Kod”

dr inż. Bartosz Trybus
Opiekun Koła Naukowego

Bezpieczeństwo sieci VLAN: Najlepsze praktyki i zagrożenia

Streszczenie

VLAN, czyli Virtual Local Area Network, to technologia sieciowa pozwalająca na podział jednej fizycznej sieci lokalnej na wiele logicznych sieci wirtualnych. Dzięki temu możliwe jest zwiększenie bezpieczeństwa i efektywności zarządzania siecią poprzez izolowanie grup użytkowników i zasobów sieciowych. VLAN-y są szczególnie przydatne w organizacjach, które potrzebują elastycznej i skalowalnej infrastruktury sieciowej. Popularność VLAN wynika z korzyści, takich jak uproszczona administracja, lepsza wydajność sieci oraz możliwość precyzyjnego zarządzania ruchem sieciowym między segmentami. W artykule omówione zostaną zagrożenia związane z VLAN, takie jak ataki VLAN hopping i MAC spoofing, oraz metody ich zapobiegania. Podkreślona zostanie również ważność bezpiecznej konfiguracji VLAN, w tym użycia silnych haseł, zasad najmniejszych uprawnień oraz list kontroli dostępu. Dodatkowo, omówione zostaną techniki monitorowania i reagowania na zmiany w konfiguracji sieci VLAN w celu zapewnienia jej ciągłego bezpieczeństwa.

Słowa kluczowe: sieci komputerowe, sieci VLAN, technologia, zagrożenia, ataki, bezpieczeństwo

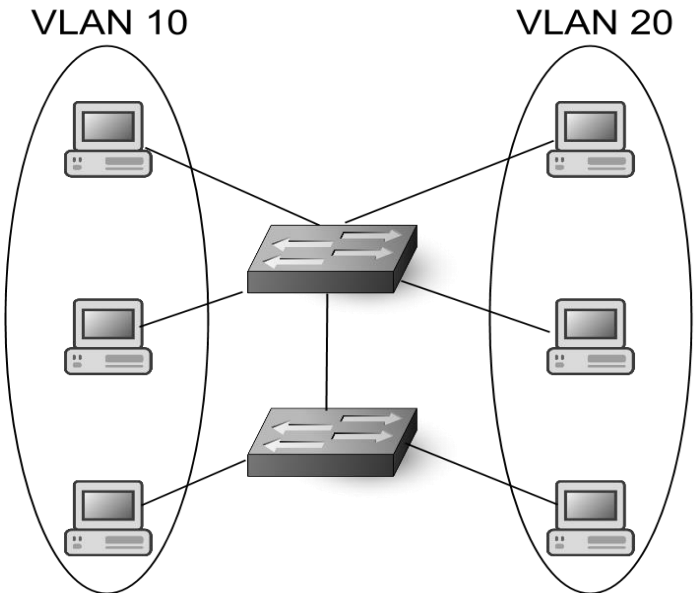
1. Wprowadzenie

VLAN, czyli Virtual Local Area Network, to rodzaj technologii sieciowej, która umożliwia podział jednej fizycznej sieci lokalnej na wiele logicznych sieci wirtualnych. Ta funkcjonalność opisana jest w standardzie 802.1Q i działa na warstwie drugiej modelu OSI, czyli warstwie łącza danych. W praktyce, konfiguracja i implementacja VLANów odbywają się głównie na przełącznikach sieciowych. Każda z tych sieci jest od siebie oddzielona, co oznacza, że urządzenia znajdujące się w różnych VLAN-ach nie mogą bezpośrednio ze sobą komunikować, chyba że pośredniczy w tym router. To sprawia, że VLAN-y są idealnym rozwiązaniem, gdy potrzebujemy izolacji między grupami urządzeń, które nie powinny się ze sobą komunikować. Innymi słowy, VLAN-y pozwalają na tworzenie logicznych granic w ramach jednej fizycznej sieci, co zwiększa bezpieczeństwo i efektywność zarządzania siecią.

VLAN-y są niezwykle popularne w sieciach komputerowych ze względu na wiele korzyści, jakie przynoszą. Dzięki nim możemy dostosować strukturę sieci do potrzeb organizacyjnych, niezależnie od fizycznej lokalizacji urządzeń, co znacznie ułatwia wprowadzanie zmian w sieci, bez konieczności zmiany samej infrastruktury fizycznej. Poprzez mniejszej liczbie urządzeń w jednej domenie rozgłoszeniowej, ilość komunikatów broadcast jest znacznie mniejsza, co przekłada się na zwiększenie ogólnej wydajności sieci. VLAN-y pozwalają na lepsze zarządzanie ruchem między segmentami sieci, np. tworząc osobne sieci tylko dla telefonii VoIP lub dodając oddzielne listy kontroli dostępu ACL dla

każdego VLAN-u. W mniejszych sieciach jest też łatwiej wykryć awarie i przeprowadzić diagnozę problemów. Bezpieczeństwo to kolejny ważny aspekt VLAN-ów. Poprzez VLAN-y możemy oddzielić ruch między różnymi grupami użytkowników, np. segregując ruch urządzeń IoT od reszty sieci w domu. Możemy również tworzyć specjalne sieci dla gości lub dla zarządzania urządzeniami sieciowymi, co dodatkowo zwiększa bezpieczeństwo.

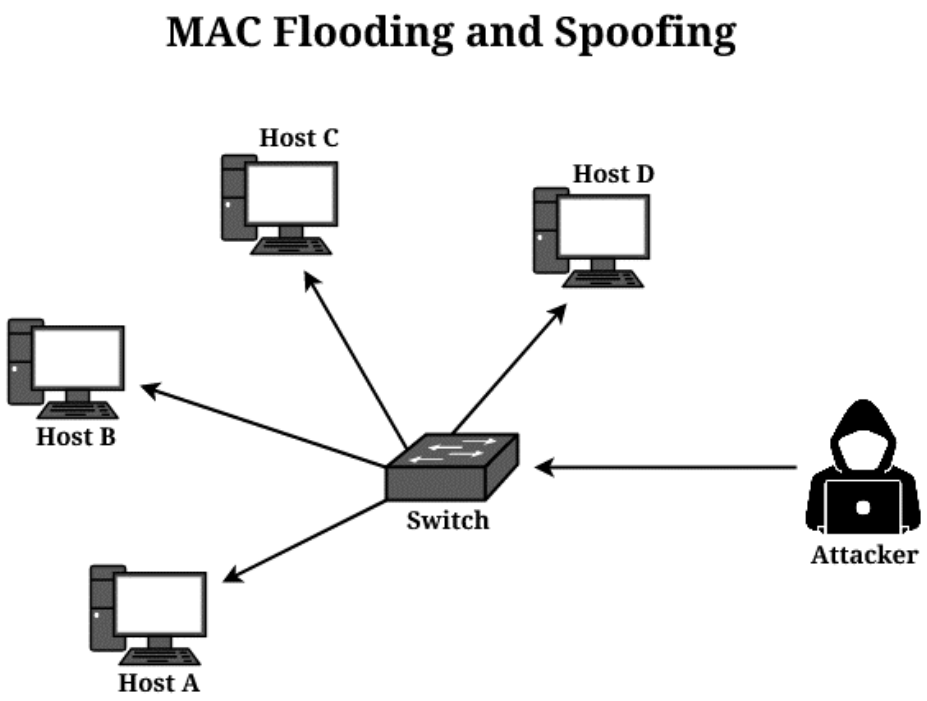
Rysunek poniżej pozwoli nam lepiej zobrazować, w jaki sposób działają sieci VLAN.



Rysunek 34 Zobrazowanie sieci VLAN,
Źródło: https://www.cs.put.poznan.pl/ddwornikowski/sieci/sieci2/vlans.html

2. Zagrożenia związane z VLAN

Ataki na sieci VLAN stanowią istotne zagrożenie dla bezpieczeństwa i integralności danych w dzisiejszych środowiskach informatycznych. Dwie z najczęstszych metod ataków to VLAN hopping oraz MAC spoofing. VLAN hopping wykorzystuje luki w konfiguracji przełącznika, aby atakujący mógł przenieść się między różnymi segmentami VLAN. Poprzez manipulację ruchem sieciowym, atakujący może uzyskać dostęp do zasobów w innych segmentach, gdzie normalnie nie miałby uprawnień. Jest to szczególnie niebezpieczne, gdyż pozwala na przechwycenie poufnych danych oraz zakłócenie normalnego funkcjonowania sieci. Z kolei ataki typu MAC spoofing polegają na fałszowaniu adresów MAC, aby atakujący mógł uzyskać nieautoryzowany dostęp do segmentów VLAN. Poprzez podszywanie się pod inne urządzenia sieciowe, haker może przekonać przełącznik, że jest częścią danego VLAN-u i uzyskać dostęp do zasobów sieciowych, do których normalnie nie miałby dostępu. Rysunek poniżej pokazuje nam jak wygląda jeden ze wspomnianych ataków, MAC spoofing.



Rysunek 2 Atak MAC spoofing; Źródło: https://www.geeksforgeeks.org/what-is-mac-spoofing-attack/

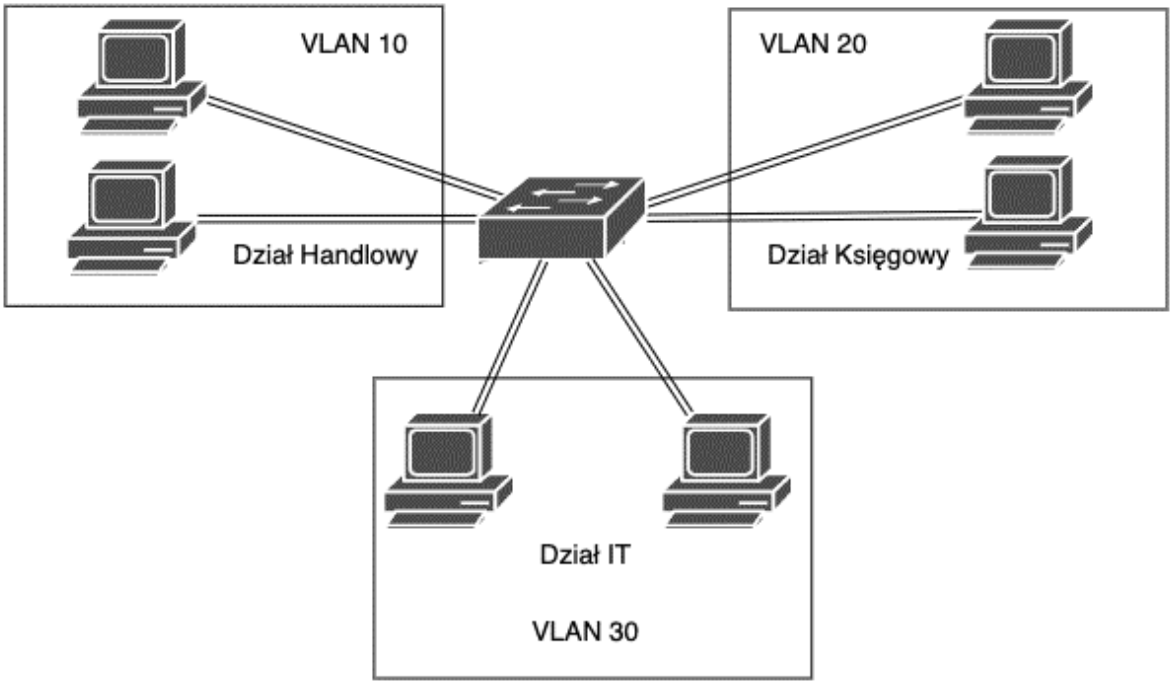
Ważne jest, aby organizacje stosowały kompleksowe strategie bezpieczeństwa, obejmujące monitorowanie ruchu sieciowego, aktualizacje oprogramowania oraz regularne przeglądy konfiguracji sieci VLAN. Tylko w ten sposób można efektywnie zabezpieczyć sieć przed atakami oraz minimalizować ryzyko poważnych konsekwencji dla firmy. Skutki finansowe, operacyjne i reputacyjne ataków na sieci VLAN mogą być katastrofalne dla firm. W przypadku utraty poufności danych, organizacje mogą być narażone na kary finansowe związane z naruszeniami przepisów dotyczących ochrony danych osobowych. Przerwy w działaniu systemów informatycznych mogą prowadzić do straty przychodów oraz utraty zaufania klientów. Dodatkowo, reputacja firmy może zostać poważnie nadszarpnięta, co może mieć długoterminowe konsekwencje dla jej pozycji na rynku.

3. Segmentacja sieci

Segmentacja sieci VLAN to kluczowy element budowy bezpiecznej i efektywnej infrastruktury sieciowej. Jej odpowiednie zaplanowanie i konfiguracja powinny być oparte na dokładnej analizie potrzeb biznesowych oraz wymagań bezpieczeństwa organizacji. Dzięki temu można zapewnić odpowiedni poziom izolacji między różnymi grupami użytkowników oraz zasobami sieciowymi. Przed przystąpieniem do segmentacji sieci VLAN, administratorzy powinni dokładnie zrozumieć strukturę organizacji, procesy biznesowe oraz specyficzne wymagania dotyczące dostępu do zasobów sieciowych.

Implementacja VLAN trunking stanowi kluczowy mechanizm umożliwiający komunikację między różnymi segmentami VLAN. Dzięki niemu można przesyłać ruch sieciowy za pomocą jednego fizycznego połączenia, co jest niezwykle istotne dla zapewnienia efektywnej komunikacji między

serwerami lub urządzeniami zlokalizowanymi w różnych segmentach VLAN. Trunking pozwala na wyznaczenie jednego lub więcej interfejsów przełącznika do przesyłania ramek z wielu VLAN-ów, co eliminuje konieczność stosowania oddzielnych fizycznych połączeń dla każdego segmentu VLAN. Rysunek niżej przedstawia zaimplementowanie przykładowego schematu VLAN trunking.



Rysunek 3 VLAN trunking Źródło: https://innasiec.pl/konfiguracja-vlan-cisco-switch/

Wykorzystanie VLAN routing to mechanizm umożliwiający komunikację między różnymi sieciami VLAN. Dzięki niemu ruch sieciowy może być przekazywany między interfejsami routera lub warstwy 3 przełącznika. To rozwiązanie umożliwia efektywne zarządzanie ruchem między segmentami VLAN oraz zapewnia płynną komunikację między nimi. VLAN routing pozwala na wyznaczenie ścieżek komunikacyjnych między VLAN-ami, co jest niezbędne dla zapewnienia elastyczności i efektywności w działaniu infrastruktury sieciowej. Dzięki temu możliwe jest skuteczne izolowanie i zabezpieczanie poszczególnych segmentów sieci, jednocześnie umożliwiając niezbędną komunikację między nimi. To rozwiązanie stanowi integralną część budowy skalowalnych i wydajnych sieci VLAN, które są w stanie sprostać rosnącym wymaganiom organizacji w zakresie komunikacji i bezpieczeństwa sieciowego.

4. Bezpieczna konfiguracja VLAN

W celu skutecznego zabezpieczenia sieci VLAN przed nieautoryzowanym dostępem oraz minimalizacji ryzyka ataków, istnieje kilka kluczowych praktyk, które warto uwzględnić. Pierwszym krokiem jest wybór silnych haseł dla VLAN oraz zabezpieczenie dostępu do panelu administracyjnego sieci VLAN oraz ich regularne zmienianie. Warto również ograniczyć dostęp do panelu administracyjnego jedynie dla uprawnionych użytkowników, stosując autoryzację wielopoziomową i monitorując logi dostępu. Kolejną ważną praktyką jest konfiguracja portów VLAN zgodnie z zasadami

najmniejszych uprawnień (least privilege). Każdy port VLAN powinien mieć skonfigurowane tylko te usługi i zasoby, które są niezbędne dla jego normalnego działania.

Ważnym elementem zabezpieczania sieci VLAN jest również ustawienie list kontroli dostępu (ACL) w celu ograniczenia dostępu do portów VLAN. ACL pozwalają na definiowanie reguł kontroli dostępu do ruchu sieciowego na poziomie portów VLAN. Poprzez ich skonfigurowanie, można kontrolować, które urządzenia mogą komunikować się między sobą, co znacząco zwiększa poziom bezpieczeństwa sieci. Ostatecznie, regularne monitorowanie i reagowanie na zmiany w konfiguracji VLAN są kluczowe dla zapobiegania atakom. Analiza zmian w konfiguracji VLAN może pomóc w wykrywaniu nieautoryzowanych zmian oraz umożliwia szybką reakcję na potencjalne zagrożenia. Praca nad zapewnieniem bezpieczeństwa sieci VLAN jest procesem ciągłym, który wymaga zaangażowania oraz monitorowania ze strony administratorów sieci.

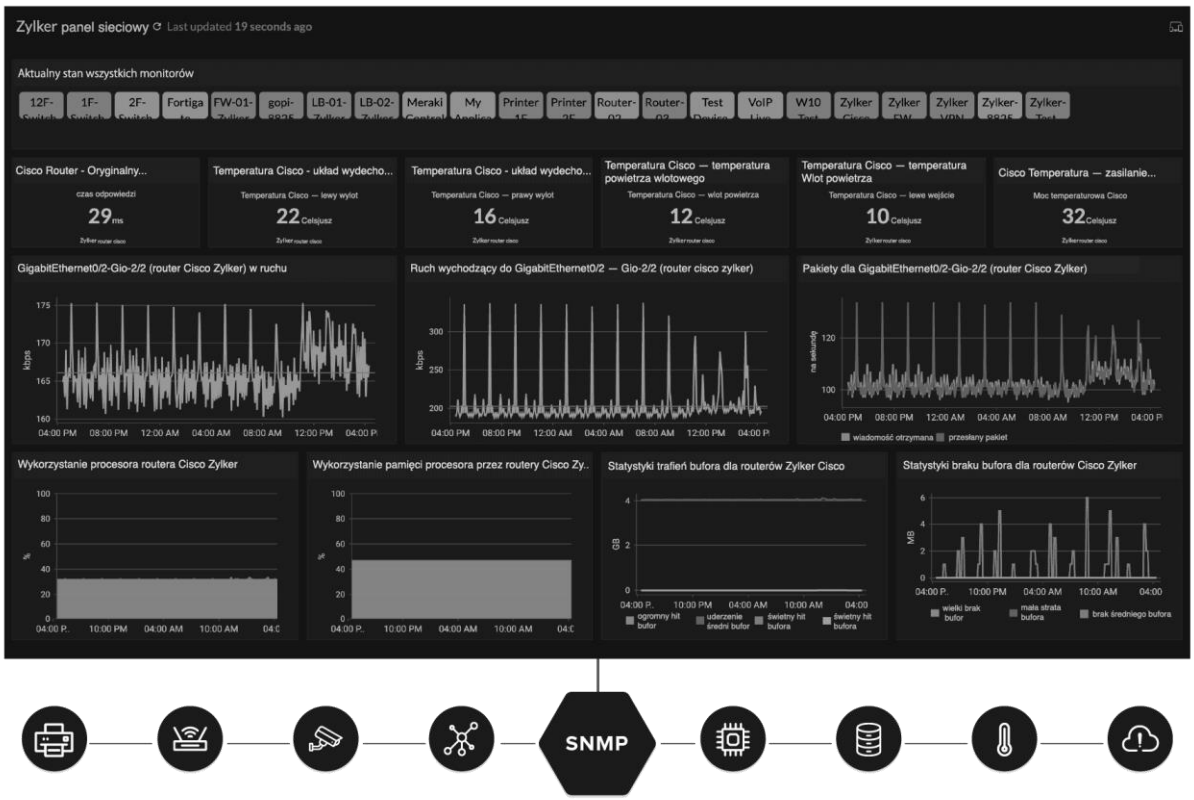
5. Monitorowanie ruchu w sieci VLAN

Tak jak wspomniane zostało powyżej, monitorowanie ruchu w sieci VLAN jest niezmiernie ważnym aspektem bezpieczeństwa na który musimy zwracać uwagę. Wybór odpowiednich narzędzi do monitorowania ruchu w sieci VLAN jest do tego kluczowe. Jednym z popularnych narzędzi są analizatory pakietów, które umożliwiają szczegółową analizę ruchu sieciowego na poziomie pakietów danych. Dzięki nim możliwe jest śledzenie przesyłanych danych oraz identyfikacja potencjalnych anomalii, jak również wykrywanie podejrzanych aktywności, które mogą wskazywać na próby ataku lub naruszenia zasad bezpieczeństwa. Kolejnym aspektem wartym uwagi jest konfiguracja monitorowania ruchu w czasie rzeczywistym. Monitorowanie w czasie rzeczywistym pozwala na bieżącą analizę i ocenę stanu sieci oraz szybką reakcję na potencjalne zagrożenia. To pozwala wykrywać nietypowe wzorce ruchu czy podejrzane aktywności, co umożliwia natychmiastowe podjęcie działań zapobiegawczych i zminimalizowanie potencjalnych szkód wynikających z ataków lub incydentów bezpieczeństwa. Rysunek poniżej przedstawia narzędzie Zylker do monitorowania ruchu w sieci.

Warto również pamiętać, że skuteczne monitorowanie ruchu w sieci VLAN wymaga nie tylko odpowiednich narzędzi, ale także właściwej konfiguracji i analizy ze strony personelu odpowiedzialnego za bezpieczeństwo sieci. Regularne szkolenia pracowników w zakresie wykorzystywania narzędzi monitorowania oraz interpretacji danych są niezbędne dla skutecznego wykrywania i reagowania na potencjalne zagrożenia. Dlatego też, oprócz wyboru odpowiednich narzędzi, ważne jest także inwestowanie w rozwój kompetencji personelu odpowiedzialnego za zarządzanie bezpieczeństwem sieci.

Regularna analiza logów pozwala na identyfikację nietypowych wzorców ruchu, które mogą wskazywać na ataki lub inne problemy związane z bezpieczeństwem. Poprzez nią możliwe jest wykrycie podejrzanych aktywności, takich jak nieudane próby logowania, podejrzane zapytania czy anomalia w wykorzystaniu zasobów sieciowych. Na podstawie tych informacji można podjąć odpowiednie działania zapobiegawcze lub reakcyjne w celu zabezpieczenia sieci przed atakami lub naruszeniami. Plan

reagowania, jako następny ważny krok do podjęcia, powinien zawierać szczegółowe procedury postępowania w przypadku wykrycia zagrożeń, w tym identyfikację i izolację źródła ataku, powiadomienie odpowiednich osób lub zespołów ds. bezpieczeństwa, oraz podjęcie działań naprawczych w celu przywrócenia integralności i bezpieczeństwa sieci. Ważne jest jego opracowanie oraz również regularne testowanie i aktualizowanie planu reagowania, aby zapewnić jego skuteczność w przypadku wystąpienia rzeczywistego incydentu. Dobrze opracowany plan reagowania pozwala na szybką identyfikację, analizę i reakcję na wszelkie zagrożenia w sieci VLAN, co minimalizuje ryzyko potencjalnych szkód wynikających z ataków lub incydentów bezpieczeństwa.



Rysunek 4: Monitorowanie ruchu w sieci, przykładowe narzędzie
Źródło: https://www.site24x7.com/pl/network-monitoring.html

6. Zarządzanie uprawnieniami

Zarządzanie uprawnieniami w sieci VLAN jest to kolejny punkt w zapewnieniu bezpieczeństwa oraz efektywnego zarządzania zasobami sieciowymi. Istnieje kilka praktyk, które warto uwzględnić, m.in. implementacja usług autoryzacji, autentykacji i rachunkowości (AAA) w sieci VLAN. Usługi AAA umożliwiają uwierzytelnianie użytkowników, autoryzację dostępu oraz monitorowanie ruchu sieciowego w sposób scentralizowany, przez co administratorzy mogą skutecznie zarządzać uprawnieniami użytkowników, zapewniając jednocześnie zgodność z politykami bezpieczeństwa. Konfiguracja VLAN-based access control list (VACL) stanowi kolejny istotny element pozwalając na definiowanie reguł kontroli dostępu na poziomie VLAN, co umożliwia precyzyjną kontrolę nad

dostępem do zasobów sieciowych w różnych segmentach VLAN. Dzięki temu możliwe jest ograniczenie dostępu do określonych zasobów lub usług tylko dla wybranych grup użytkowników, co zwiększa poziom bezpieczeństwa sieci. Trzeba pamiętać że zarządzanie uprawnieniami w sieci VLAN wymaga nie tylko odpowiedniej konfiguracji technicznej, ale także przestrzegania najlepszych praktyk oraz ciągłego monitorowania i aktualizowania polityk bezpieczeństwa. Ważne jest więc prowadzenie systematycznych audytów bezpieczeństwa oraz szkoleń dla personelu odpowiedzialnego za zarządzanie siecią.

Dzięki technologii takich jak IEEE 802.1X, możliwe jest autoryzowanie użytkowników VLAN na podstawie ich tożsamości, co umożliwia bardziej zaawansowane zarządzanie uprawnieniami. Mechanizm autoryzacji oparty na IEEE 802.1X wymaga od użytkownika podania odpowiednich danych uwierzytelniających, takich jak nazwa użytkownika i hasło, przed uzyskaniem dostępu do sieci. Pozytywna autoryzacja pozwala na przypisanie użytkownika do odpowiedniego segmentu VLAN, co zapewnia kontrolę nad dostępem do zasobów sieciowych oraz zwiększa bezpieczeństwo infrastruktury. Nadawanie uprawnień dostępu na podstawie ról użytkowników lub grup VLAN umożliwia zastosowanie zasad ‘najmniejszych uprawnień’, co oznacza, że każdy użytkownik ma dostęp tylko do tych zasobów sieciowych, które są mu niezbędne do wykonywania określonych zadań. Administrator może definiować różne role użytkowników oraz grupy VLAN i przypisywać im odpowiednie uprawnienia dostępu. W ten sposób możliwe jest ograniczenie dostępu tylko do niezbędnych zasobów sieciowych, co zwiększa bezpieczeństwo sieci oraz redukuje ryzyko naruszeń związanych z nadmiernymi uprawnieniami.

7. Regularne aktualizacje i audyty

Przeprowadzanie regularnych audytów bezpieczeństwa sieci VLAN jest kluczowe dla identyfikacji potencjalnych luk w zabezpieczeniach oraz oceny zgodności z najlepszymi praktykami bezpieczeństwa. Podczas nich analizowane są różne elementy infrastruktury sieciowej, w tym konfiguracja urządzeń, polityki bezpieczeństwa, dostępność aktualizacji oraz monitorowanie ruchu sieciowego. Audyty pozwalają na wykrycie słabych punktów w zabezpieczeniach oraz podjęcie odpowiednich działań naprawczych w celu zwiększenia odporności sieci na potencjalne zagrożenia. Producenci systemów ciągle wydają aktualizacje, które zawierają poprawki bezpieczeństwa oraz łatki usuwające znane podatności na ataki. Brak aktualizacji może prowadzić do wykorzystania zabezpieczeń przez potencjalnych atakujących, co zwiększa ryzyko naruszenia bezpieczeństwa sieci. Dzięki regularnej aktualizacji oprogramowania i firmware urządzeń sieciowych zapewniana jest ochrona przed znanymi lukami w zabezpieczeniach, minimalizacji ryzyka tych ataków oraz utrzymanie sieci VLAN w odpowiednio wysokim standardzie bezpieczeństwa.

Regularna ocena zgodności z najlepszymi praktykami branżowymi i standardami zabezpieczeń pozwala na identyfikację obszarów, które mogą wymagać ulepszeń lub dostosowań do aktualnych standardów bezpieczeństwa. Praktyki te obejmują zgodność z różnymi standardami, takimi jak ISO

27001, NIST czy PCI DSS, które określają wymagania dotyczące bezpieczeństwa informacji i sieci komputerowych. Poprzez regularne przeprowadzanie ocen zgodności, administratorzy mogą upewnić się, że sieć VLAN jest zabezpieczona zgodnie z aktualnymi standardami i najlepszymi praktykami branżowymi. Poprzez systematyczne śledzenie i analizę incydentów bezpieczeństwa, administratorzy mogą lepiej zrozumieć rodzaje zagrożeń, z jakimi muszą się zmagać, oraz identyfikować słabe punkty w obecnych zabezpieczeniach. Analiza incydentów pozwala na wyciągnięcie wniosków oraz wprowadzenie odpowiednich modyfikacji w strategiach zabezpieczeń, aby lepiej chronić sieć VLAN przed przyszłymi atakami. Dzięki temu możliwe jest ciągłe doskonalenie strategii bezpieczeństwa oraz adaptacja do zmieniającego się krajobrazu zagrożeń. Rysunek poniżej pokazuje jak wygląda przykładowy wspomniany wyżej audyt.

Rysunek 5:Jak wyglądają audyty
Źródło: https://racontrols.pl/bazawiedzy/audyt-sieciowy-inwestycja-w-bezpieczenstwo-przedsiebiorstwa

8. Szkolenie pracowników

Organizowanie regularnych szkoleń z zakresu bezpieczeństwa sieciowego dla pracowników jest obowiązkowym elementem w budowaniu świadomości i kompetencji w zakresie ochrony sieci VLAN. Podczas tych szkoleń pracownicy są zapoznawani z podstawowymi zagrożeniami związanymi z sieciami VLAN, takimi jak ataki typu VLAN hopping czy MAC spoofing, oraz z najlepszymi praktykami w zakresie korzystania z sieci. W ramach szkoleń omawiane są również procedury postępowania w przypadku wykrycia podejrzanej aktywności w sieci oraz sposoby, w jakie pracownicy mogą przyczynić się do zwiększenia bezpieczeństwa sieci poprzez odpowiedzialne korzystanie z zasobów sieciowych. Edukowanie pracowników na temat najnowszych zagrożeń związanych z sieciami VLAN pozwala na zwiększenie świadomości, zwiększenia gotowości do reakcji i zrozumienia

potencjalnych ryzyk związanych z sieciami VLAN. Ponadto, świadomość najnowszych zagrożeń umożliwia pracownikom podejmowanie odpowiednich działań prewencyjnych oraz wspieranie działań zespołu ds. bezpieczeństwa w zapobieganiu incydentom i zwiększaniu ogólnego poziomu bezpieczeństwa sieci VLAN. Wprowadzenie tych praktyk w ramach szkoleń pracowników umożliwia stworzenie świadomej i odpowiedzialnej społeczności w zakresie bezpieczeństwa sieci VLAN.

Poprzez edukację pracowników na temat zagrożeń związanych z niebezpiecznymi praktykami w sieciach VLAN, takimi jak udostępnianie poufnych informacji czy klikanie w podejrzane linki, można zmniejszyć ryzyko naruszeń bezpieczeństwa. Pracownicy powinni być świadomi konsekwencji niewłaściwego korzystania z zasobów sieciowych oraz zabezpieczeń, które należy zachować podczas pracy w różnych segmentach VLAN. Zachęcanie pracowników do zgłaszania podejrzanych aktywności w sieci VLAN oraz aktywnego monitorowania swojego otoczenia sieciowego pozwala na szybką reakcję zespołu ds. bezpieczeństwa oraz minimalizuje ryzyko naruszeń bezpieczeństwa.

9. Kontynuacja monitorowania i dostosowywania

Ustanowienie procesu ciągłego monitorowania ruchu w sieci VLAN jest niezbędne dla skutecznej ochrony przed zagrożeniami. Proces ten powinien być prowadzony w czasie rzeczywistym, aby umożliwić szybką reakcję na wszelkie zmiany w ruchu sieciowym i potencjalne incydenty bezpieczeństwa. Dzięki monitorowaniu ruchu w sieci VLAN na bieżąco, możliwe jest szybkie wykrywanie nietypowych wzorców ruchu, podejrzanych aktywności oraz prób nieautoryzowanego dostępu. W efekcie, zespół ds. bezpieczeństwa może podejmować natychmiastowe działania w celu zidentyfikowania i zneutralizowania potencjalnych zagrożeń, minimalizując tym samym ryzyko poważnych incydentów. Analiza trendów dotyczących bezpieczeństwa sieci VLAN stanowi kluczowy element skutecznej strategii ochrony. Poprzez śledzenie i analizowanie zmian w zagrożeniach oraz wzorcach ataków, zespół ds. bezpieczeństwa może zyskać wgląd w ewoluujące zagrożenia i dostosować zabezpieczenia w odpowiedzi na te zmiany. W ten sposób możliwe jest ciągłe doskonalenie strategii ochrony sieci VLAN, uwzględniając najnowsze trendów i wyzwania związane z bezpieczeństwem. Dostosowywanie zabezpieczeń na podstawie analizy trendów pozwala organizacji na proaktywne reagowanie na nowe i rozwijające się zagrożenia, co z kolei przyczynia się do zwiększenia ogólnego poziomu bezpieczeństwa sieci VLAN.

Prowadzenie regularnych przeglądów i aktualizacji polityk bezpieczeństwa sieci VLAN sprawia, że można ocenić ich skuteczność w kontekście zmieniających się wymagań biznesowych oraz nowych zagrożeń. Aktualizacje polityk są niezbędne do zapewnienia, że są one nadal adekwatne i skuteczne w obliczu dynamicznie zmieniającego się środowiska sieciowego. Utrzymywanie kontaktu z firmami i społecznościami branżowymi jest ważne dla śledzenia nowych trendów i rozwiązań w dziedzinie bezpieczeństwa sieci VLAN. Współpraca z ekspertami branżowymi oraz uczestnictwo w konferencjach, szkoleniach i forum internetowych pozwala na pozyskanie wiedzy na temat najnowszych zagrożeń, narzędzi i technik ataków, a także na wymianę doświadczeń i praktyk z innymi specjalistami ds.

bezpieczeństwa. Dzięki temu organizacja może pozostać na bieżąco z najnowszymi trendami i innowacjami w dziedzinie bezpieczeństwa sieci VLAN oraz dostosować swoje zabezpieczenia do zmieniających się potrzeb i wyzwań.

10. Podsumowanie

Bezpieczeństwo sieci VLAN jest kluczowym elementem zarządzania nowoczesnymi infrastrukturami sieciowymi. Poprawna segmentacja sieci, konfiguracja VLAN-ów oraz implementacja zaawansowanych mechanizmów bezpieczeństwa, takich jak VLAN trunking i VLAN routing, pozwalają na efektywne zarządzanie ruchem oraz izolację poszczególnych grup użytkowników. Jednakże, sieci VLAN są narażone na różne zagrożenia, w tym ataki typu VLAN hopping i MAC spoofing. Aby minimalizować ryzyko takich ataków, niezbędne jest stosowanie silnych haseł, konfiguracja portów zgodnie z zasadami najmniejszych uprawnień, oraz regularne monitorowanie i audyty bezpieczeństwa. Narzędzia do monitorowania ruchu, analiza logów zdarzeń oraz planowanie odpowiednich reakcji na zagrożenia stanowią fundament skutecznej strategii obronnej. Kontynuacja monitorowania i dostosowywania polityk bezpieczeństwa w odpowiedzi na nowe zagrożenia oraz regularne szkolenie pracowników w zakresie najnowszych zagrożeń i najlepszych praktyk są równie ważne. Współpraca z firmami i społecznościami branżowymi pozwala na bieżąco śledzić trendy i implementować najnowsze rozwiązania w dziedzinie bezpieczeństwa. Wdrożenie kompleksowych strategii bezpieczeństwa oraz ich regularne aktualizowanie i audytowanie jest niezbędne, aby zapewnić integralność i ochronę danych w sieciach VLAN. Tylko w ten sposób można efektywnie chronić sieci przed atakami, minimalizując jednocześnie ryzyko finansowe, operacyjne i reputacyjne związane z naruszeniami bezpieczeństwa.

Źródła internetowe:

https://pasja-informatyki.pl/sieci-komputerowe/vlan-wprowadzenie/(dostęp: 19.05.2024).
https://www.nastykusieci.pl/wprowadzenie-vlan/(dostęp: 19.05.2024).
https://itfocus.pl/dzial-it/sieci/vlan-dla-zaawansowanych/(dostęp: 19.05.2024).
https://www.computerworld.pl/news/Zagrozenie-bezpieczenstwa-sieci-w-warstwie-2,303061.html(dostęp: 19.05.2024).
https://networkexpert.pl/baza-wiedzy/vlany-polaczenia-typu-trunk-i-vtp/(dostęp: 19.05.2024).
https://sgit.pl/monitorowanie-ruchu-w-sieci-lan/(dostęp: 19.05.2024).
https://innasiec.pl/konfiguracja-vlan-cisco-switch/(dostęp: 19.05.2024).

Aleksandra Rokita, Katarzyna Maternia, Magdalena Matuła, Aleksandra Sawicka, Wiktor Kuczek
SKNI „Kod”

dr inż. Bartosz Trybus
Opiekun Koła Naukowego

Inteligentny dom - wykorzystanie technologii w zarządzaniu domem

Streszczenie

Inteligentny dom to zintegrowany system technologii umożliwiający zdalne sterowanie i automatyzację funkcji domowych, takich jak oświetlenie, ogrzewanie, klimatyzacja, monitoring i systemy alarmowe. Celem artykułu jest zbadanie roli sztucznej inteligencji i technologii w funkcjonowaniu inteligentnych domów oraz omówienie głównych elementów systemu, takich jak centralna jednostka sterująca (hub), urządzenia wykonawcze, aplikacje mobilne i asystenci głosowi. Artykuł skupia się na przedstawieniu głównych elementów inteligentnego domu, takich jak centralna jednostka sterująca, urządzenia wykonawcze, aplikacje mobilne oraz asystenci głosowi. Opisuje również rolę IoT w inteligentnych domach, w tym stosowane protokoły komunikacyjne jak MQTT i CoAP.
W artykule przewiduje się przyszłość inteligentnych domów z bardziej zaawansowanymi algorytmami AI, zwiększoną autonomią systemów oraz rozwijającymi się standardami komunikacyjnymi, co ma zapewnić lepszą interoperacyjność i wygodę użytkowników.

Słowa kluczowe: dom, inteligentny dom, AI, IoT, internet rzeczy.

1. Czym jest inteligentny dom?

Inteligentny dom to system zintegrowanych urządzeń i technologii, które umożliwiają zdalne sterowanie i automatyzację różnych funkcji w domu, takich jak oświetlenie, ogrzewanie, klimatyzacja, systemy alarmowe, monitoring, rolety, bramy garażowe i wiele innych. Celem inteligentnego domu jest zwiększenie bezpieczeństwa, oszczędności energii i efektywności zarządzania domem. Sztuczna inteligencja umożliwia inteligentnym systemom uczenie się zachowań domowników, co pozwala na zapewnienie użytkownikom większego komfortu przez dostosowanie się do ich preferencji. Dzięki wykorzystaniu sztucznej inteligencji możliwe jest analizowanie ogromnych ilości danych generowanych przez urządzenia domowe takie jak kamery, czujniki ruchu czy termometry.

Celem artykułu jest zbadanie roli sztucznej inteligencji i technologii w funkcjonowaniu inteligentnych domów oraz omówienie głównych elementów systemu, takich jak centralna jednostka sterująca (hub), urządzenia wykonawcze, aplikacje mobilne i asystenci głosowi. W artykule przedstawiono również rolę IoT w inteligentnych domach, w tym stosowane protokoły komunikacyjne jak MQTT i CoAP. Przewiduje się również przyszłość inteligentnych domów z bardziej zaawansowanymi algorytmami AI, zwiększoną autonomią systemów oraz rozwijającymi się standardami komunikacyjnymi, co ma zapewnić lepszą interoperacyjność i wygodę użytkowników.

2. Główne elementy inteligentnego domu

Centralna jednostka sterująca, znana również jako hub, pełni rolę „mózgu" systemu inteligentnego domu, integrując i kontrolując wszystkie podłączone urządzenia. Dzięki temu różne urządzenia mogą współpracować ze sobą, realizując złożone scenariusze automatyzacji. Huby mogą obsługiwać wiele różnych protokołów komunikacyjnych, takich jak Wi-Fi, Zigbee, Z-Wave czy Bluetooth. Pozwala to na integrację szerokiej gamy urządzeń od różnych producentów. Huby umożliwiają tworzenie scenariuszy automatyzacji, np. „wyjście z domu" – wyłączenie świateł, zamknięcie drzwi i obniżenie temperatury.

Urządzenia wykonawcze ułatwiają życie w inteligentnych domach. Przykładem urządzeń wykonawczych mogą być inteligentne gniazdka, które umożliwiają zdalne włączanie i wyłączanie podłączonych do nich urządzeń, monitorowanie zużycia energii oraz automatyzację ich pracy. Inteligentne żarówki pozwalają na zdalne sterowanie oświetleniem, dostosowywanie jasności i kolorów, a także automatyzację włączania/wyłączania w zależności od pory dnia lub obecności osób. Inteligentne termostaty umożliwiają zdalne zarządzanie temperaturą w domu, tworzenie harmonogramów ogrzewania/chłodzenia oraz optymalizację zużycia energii na podstawie preferencji użytkowników i danych z czujników. Czujniki ruchu wykrywają obecność osób w pomieszczeniu i mogą automatycznie włączać lub wyłączać oświetlenie, systemy alarmowe czy inne urządzenia. Czujniki dymu i temperatury monitorują środowisko w domu, wykrywając zagrożenia pożarowe i inne niebezpieczne sytuacje, a także mogą uruchamiać alarmy i powiadamiać użytkowników o niebezpieczeństwie. Kamery monitoringu umożliwiają zdalny podgląd wnętrza i otoczenia domu, wykrywanie ruchu i nagrywanie zdarzeń, a także zintegrowanie z systemem alarmowym.

Aplikacja mobilna pozwala użytkownikom na zdalne sterowanie wszystkimi urządzeniami w inteligentnym domu za pomocą smartfona lub tabletu, niezależnie od miejsca, w którym się

znajdują. Użytkownicy mogą na bieżąco monitorować stan urządzeń, zużycie energii, temperaturę w pomieszczeniach, bezpieczeństwo i inne parametry. Aplikacja może wysyłać powiadomienia push o ważnych zdarzeniach, takich jak wykrycie ruchu, alarm dymu, otwarcie drzwi itp. Umożliwia tworzenie i zarządzanie scenariuszami automatyzacji oraz harmonogramami działania urządzeń, co pozwala na optymalizację codziennych zadań i zwiększenie komfortu życia.

Integracja systemu inteligentnego domu z asystentami głosowymi, takimi jak Amazon Alexa, Google Assistant czy Apple Siri, wykorzystują AI do rozpoznawania i interpretowania poleceń głosowych, co pozwala na łatwe sterowanie urządzeniami w inteligentnym domu za pomocą mowy. Użytkownicy mogą np. włączyć światło, zmienić temperaturę, zablokować drzwi czy uruchomić odkurzacz, używając jedynie komend głosowych. Możliwość tworzenia specjalnych komend i scenariuszy głosowych, które wykonują zestaw zadań, np. komenda „dobranoc" może wyłączyć wszystkie światła, zamknąć drzwi i ustawić termostat na niższą temperaturę. AI może również dostarczać użytkownikom spersonalizowane informacje, przypomnienia i rekomendacje, bazując na analizie ich codziennych nawyków i preferencji oraz powiadomienia na temat stanu urządzeń.

3. IoT a inteligentny dom

Internet rzeczy (IoT) to koncepcja, w której różne urządzenia fizyczne są połączone z internetem, umożliwiając im analizę, wymianę danych i komunikację. Dzięki IoT możliwe jest tworzenie zaawansowanych systemów monitorowania i zarządzania, które znajdują zastosowanie w różnych dziedzinach życia. Urządzenia IoT muszą ze sobą współpracować, przez co wymagana jest zgodność z różnymi standardami (np. Wi-Fi, Bluetooth lub Zigbee umożliwiającymi bezproblemową integrację i komunikację między urządzeniami od różnych producentów) oraz protokołami. Protokoły komunikacyjne, np. MQTT (Message Queuing Telemetry Transport) lub CoAP (Constrained Application Protocol), są szeroko stosowane do przesyłania danych między urządzeniami.

MQTT działa w modelu Publish/Subscribe, gdzie urządzenia publikują informacje na określone tematy, a inne urządzenia subskrybują te tematy, aby otrzymać informacje. Centralnym elementem MQTT jest broker, który jest pośrednikiem w komunikacji między urządzeniami. Urządzenia publikujące informacje wysyłają je do brokera, który następnie przekazuje informacje do odpowiednich subskrybentów. MQTT jest zaprojektowany tak, aby

jak najbardziej zminimalizować zużycie zasobów, dzięki czemu dobrze się sprawdza w urządzeniach z ograniczoną mocą obliczeniową i przepustowością sieci. CoAP działa w modelu klient/serwer. Klient wysyła zapytanie (request) do serwera, który wysyła na nie odpowiedź (response). Protokół został zaprojektowany z myślą o urządzeniach z ograniczonymi zasobami, aby zminimalizować obciążenie sieci. CoAP korzysta z protokołu UDP, co pozwala na szybką transmisję danych z niskim opóźnieniem, jednak nie zapewnia mechanizmów kontroli błędów.

Ze względu na ilość danych generowanych przez systemy IoT do ich przetwarzania wymagana jest duża moc obliczeniowa i odpowiednie narzędzia analityczne. Inteligentny dom to specyficzne zastosowanie IoT, które koncentruje się na automatyzacji i usprawnieniu zarządzania urządzeniami domowymi. W inteligentnym domu połączenie różnych urządzeń w sieci umożliwia ich zdalne sterowanie i automatyzację. Jednym z aspektów zastosowania IoT w inteligentnych domach jest zdalne sterowanie urządzeniami. Dzięki temu użytkownicy mogą przez aplikacje na urządzenia mobilne kontrolować np. oświetlenie z dowolnego miejsca na świecie.

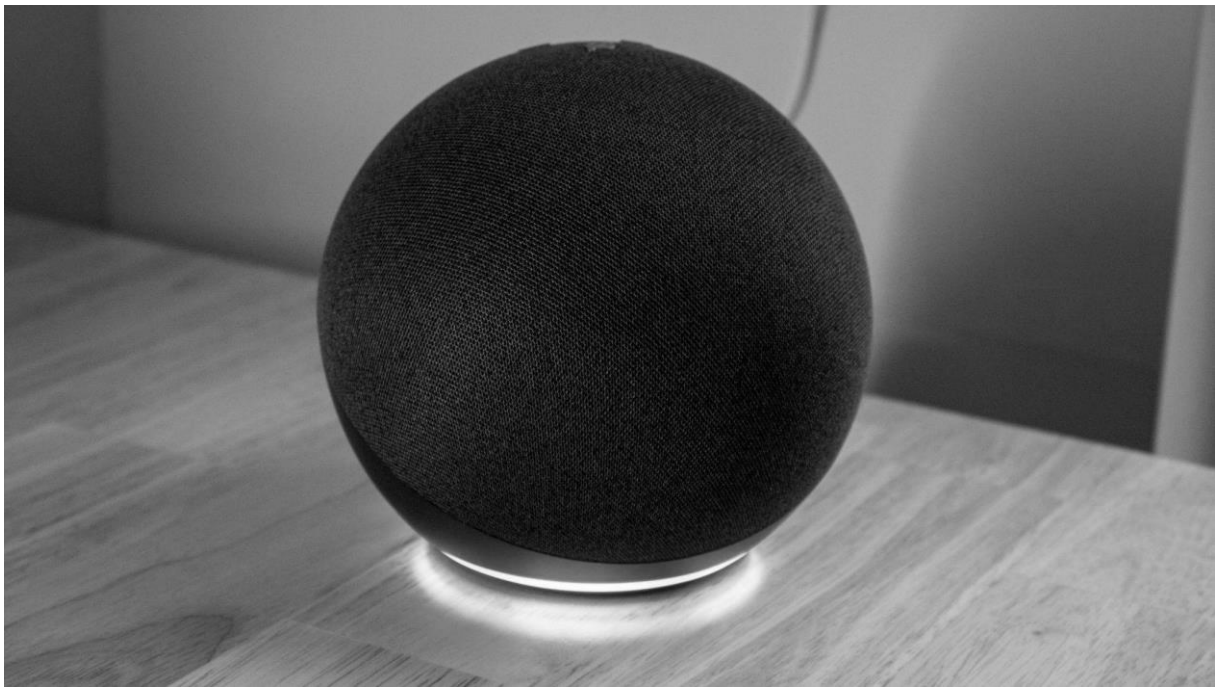
4. Alexa-inteligentny asystent

Alexa, inteligentny asystent stworzony przez Amazon, rewolucjonizuje sposób, w jaki zarządzamy naszymi domami i codziennymi czynnościami. Wyposażona w zaawansowane algorytmy uczenia maszynowego oraz technologię rozpoznawania mowy, Alexa potrafi zrozumieć i odpowiedzieć na różnorodne polecenia głosowe, dostosowując się do preferencji użytkowników. Dzięki temu, korzystanie z niej staje się niezwykle intuicyjne i komfortowe.

Jedną z największych zalet Alexy jest jej zdolność do integracji z szeroką gamą urządzeń smart home. Niezależnie od tego, czy chodzi o sterowanie oświetleniem, termostatem, systemem bezpieczeństwa, czy też sprzętem AGD, Alexa może stać się centralnym punktem zarządzania inteligentnym domem. Użytkownicy mogą w prosty sposób wydawać polecenia głosowe, takie jak "Alexa, włącz światła w salonie" lub "Alexa, ustaw temperaturę na 22 stopnie", a system automatycznie wykonuje te polecenia, zapewniając wygodę i oszczędność czasu.

Alexa nie tylko ułatwia zarządzanie domem, ale także wspiera w codziennych zadaniach. Może przypominać o ważnych spotkaniach, tworzyć listy zakupów, a nawet pomagać w przygotowywaniu posiłków, odczytując przepisy krok po kroku. Jej funkcjonalność można dodatkowo rozszerzać dzięki umiejętnościom (skills), które użytkownicy mogą pobierać i

instalować według własnych potrzeb. Dzięki temu Alexa może stać się jeszcze bardziej wszechstronna, oferując na przykład informacje o pogodzie, aktualnościach, czy odtwarzając ulubioną muzykę.



Rysunek 1 Alexa, źródło: https://edition.cnn.com/cnn-underscored/electronics/how-does-alexa-work-interview

5. Rola sztucznej inteligencji w inteligentnym domu

Sztuczna inteligencja odgrywa kluczową rolę w funkcjonowaniu inteligentnych domów, umożliwiając automatyzację i inteligentne zarządzanie różnymi systemami i urządzeniami domowymi. Oto główne obszary, w których AI przyczynia się do działania inteligentnego domu:

1. Automatyzacja i sterowanie urządzeniami. Algorytmy uczenia maszynowego analizują wzorce zachowań i preferencje użytkowników, aby automatycznie dostosowywać ustawienia urządzeń w inteligentnym domu. Na przykład, algorytmy mogą monitorować codzienne nawyki mieszkańców, takie jak godziny powrotu do domu czy pory snu, i na tej podstawie regulować oświetlenie, ogrzewanie i klimatyzację. Inteligentne termostaty mogą dostosowywać temperaturę w zależności od preferencji użytkowników oraz prognoz pogody, co optymalizuje zużycie energii. Algorytmy te mogą również zautomatyzować działanie systemów audio, wideo i urządzeń kuchennych, zwiększając komfort i efektywność energetyczną.

2. Optymalizacja zużycia energii. Inteligentne domy dążą do zmaksymalizowania efektywności energetycznej i redukcji kosztów zużycia energii. Kluczową rolę w systemach odgrywa sztuczna inteligencja, która monitoruje zużycie energii i optymalizuje pracę urządzeń, aby zmaksymalizować oszczędności i zminimalizować wykorzystanie energii. Systemy w inteligentnych domach wyposażone są w różne czujniki i urządzenia pomiarowe. Pozwalają one na zbieranie danych dotyczących zużycie energii przez różne urządzenia domowe i analizowanie ich w czasie rzeczywistym. Pozwala to zidentyfikować obszary o wysokim zużyciu energii i potencjalne źródła oszczędności. Na podstawie zebranych danych sztuczna inteligencja może dynamicznie dostosować pracę urządzeń domowych w celu maksymalizacji efektywności energetycznej, poprzez regulację oświetlenia w zależności od pory dnia.
3. Monitoring. Rozwiązania umożliwiające rozpoznawanie obiektów i twarzy, wykrywające ruch oraz anomalie znacząco zwiększają bezpieczeństwo użytkowników. Frigate jest systemem nagrywania wideo NVR (ang. Network Video Recorder to specjalny rodzaj rejestratora, który służy do nagrywania i przechowywania nagrań z kamer IP oraz umożliwia ich odtwarzanie i zarządzanie z poziomu sieci), którego zadaniem jest wykrywanie obiektów w czasie rzeczywistym przy użyciu sztucznej inteligencji. Tradycyjna detekcja ruchu w systemach monitorujących polega na analizowaniu różnic między kolejnymi klatkami wideo. Frigate natomiast wykorzystuje zaawansowane algorytmy uczenia maszynowego do wykrywania obiektów takich jak ludzie, samochody lub zwierzęta. Dzięki temu, że algorytmy są trenowane na ogromnych zbiorach danych, potrafią one rozróżniać istotne obiekty i ignorować te nieistotne. Frigate pozwala także na definiowanie obszarów wykluczenia, w których ruch będzie ignorowany. Analiza wideo odbywa się bez wysyłania strumieni wideo do chmury.

Rysunek 2 Obraz z monitoringu Frigate, źródło: https://frigate.video/

Aspekty, w których sztuczna inteligencja jest wykorzystywana w inteligentnych kamerach to m.in wykrycie wtargnięcia, wykrywanie przekraczania linii, wykrywanie skupisk ludzi czy dezaktywowanie alarmu wywołanego przez np. dzikie zwierzę.

4. Personalizacja i wygoda. Uczenie maszynowe umożliwia tworzenie spersonalizowanych scenariuszy i preferencji dla różnych aktywności i sytuacji w domu. Dzięki analizie danych o zachowaniach użytkowników, systemy te mogą automatycznie dostosowywać ustawienia, takie jak oświetlenie, ogrzewanie czy multimedia, aby najlepiej odpowiadały na potrzeby mieszkańców. Asystenci AI, wyposażeni w technologię rozpoznawania mowy, ułatwiają interakcję z inteligentnym domem, pozwalając na sterowanie urządzeniami za pomocą poleceń głosowych, co zwiększa wygodę i intuicyjność obsługi.

6. Wyzwania i ograniczenia

Bezpieczeństwo i prywatność jest jednym z najpoważniejszych wyzwań związanych z inteligentnymi domami. Urządzenia IoT zbierają ogromne ilości danych, które mogą być wrażliwe i osobiste. Hakerzy mogą próbować uzyskać nieautoryzowany dostęp do tych urządzeń, co może prowadzić do kradzieży danych, naruszeń prywatności, a nawet do fizycznego włamania do domu. Zapewnienie odpowiednich zabezpieczeń, takich jak szyfrowanie danych, silne hasła, regularne aktualizacje oprogramowania oraz firewall, jest

kluczowe. Równie ważna jest edukacja użytkowników na temat bezpieczeństwa i prywatności, aby świadomie chronili swoje dane i systemy. Integracja różnych urządzeń i systemów w inteligentnym domu może być wyzwaniem ze względu na różnorodność standardów i protokołów komunikacji. Urządzenia różnych producentów często nie są ze sobą kompatybilne, co utrudnia stworzenie spójnego i efektywnego systemu. Wprowadzenie uniwersalnych standardów i protokołów komunikacyjnych, takich jak Project CHIP, może pomóc w rozwiązaniu tego problemu, ale wymaga to szerokiej współpracy między producentami. Brak kompatybilności może również prowadzić do problemów z konserwacją i aktualizacją systemów, co dodatkowo komplikuje zarządzanie inteligentnym domem. Implementacja inteligentnych technologii w domu wiąże się z wysokimi kosztami. Urządzenia IoT, zaawansowane systemy bezpieczeństwa oraz profesjonalna instalacja i konfiguracja mogą być kosztowne. Dla wielu użytkowników bariera finansowa jest znacząca, co może hamować szerokie przyjęcie tych technologii. Chociaż długoterminowe oszczędności na kosztach energii i zwiększenie bezpieczeństwa mogą przynieść korzyści to jednak wysokie koszty początkowe pozostają istotnym ograniczeniem. Ponadto, koszty związane z konserwacją, aktualizacjami i potencjalnymi naprawami mogą również stanowić dodatkowe obciążenie finansowe dla użytkowników.

7. **Projekt chip**

Project CHIP (Connected Home over IP) to inicjatywa mająca na celu stworzenie otwartego, interoperacyjnego standardu komunikacyjnego dla urządzeń IoT używanych w inteligentnych domach. Projekt ten został zapoczątkowany przez organizację Connectivity Standards Alliance (CSA), wcześniej znanej jako Zigbee Alliance, i cieszy się wsparciem wielu czołowych producentów technologii, między innymi takich jak Apple, Google czy Amazon. Jednym z głównych celów Projektu CHIP jest kompatybilność, która umożliwia urządzeniom różnych producentów współprace ze sobą bez problemów związanych z różnymi standardami komunikacyjnymi. Dzięki temu użytkownicy mogą łatwiej integrować różnorodne urządzenia IoT w swoich inteligentnych domach. Projekt CHIP kładzie duży nacisk na zapewnienie wysokiego poziomu bezpieczeństwa. W standardzie uwzględnione są zaawansowane mechanizmy szyfrowania danych oraz autoryzacji, co ma chronić urządzenia i dane użytkowników przed potencjalnymi atakami hakerskimi.

Kolejnym celem projektu jest zapewnienie prostego i intuicyjnego interfejsu dla użytkowników. Standard ma umożliwiać łatwe konfigurowanie i zarządzanie urządzeniami IoT,

co jest kluczowe dla masowej adopcji technologii inteligentnych domów. Celem projektu jest umożliwienie działania na różnych platformach, i w różnych systemach operacyjnych takich jak Android i iOS, co ma zwiększyć, uniwersalność i dostępność dla szerokiego grona użytkowników.

Realizacja Project CHIP może znacząco przyczynić się do rozwiązania problemu braku kompatybilności między urządzeniami IoT, który obecnie jest jednym z głównych wyzwań w dziedzinie inteligentnych domów. Dzięki temu standardowi użytkownicy mogą spodziewać się łatwiejszej integracji urządzeń, lepszej ochrony prywatności oraz większego komfortu korzystania z nowoczesnych technologii w swoim domu.

8. Przyszłość AI w inteligentnych domach

Przyszłość AI w inteligentnych domach zapowiada się niezwykle obiecująco. Wiele nowych technologii i innowacji umożliwia nam zrewolucjonizowanie sposobu, w jaki zarządzamy naszymi domami. Przyszłe rozwiązania mogą obejmować bardziej zaawansowane algorytmy uczenia maszynowego, które będą w stanie jeszcze lepiej analizować dane i przewidywać potrzeby użytkowników. Rozwój technologii takich jak 5G zapewni szybsze i bardziej niezawodne połączenia, umożliwiając bardziej efektywną komunikację między urządzeniami w IoT. Ponadto, wprowadzenie kwantowej sztucznej inteligencji może otworzyć nowe możliwości w zakresie przetwarzania i analizy ogromnych ilości danych w czasie rzeczywistym.Przyszłość inteligentnych domów to także większa autonomia systemów. Dzięki zaawansowanym technologiom sztucznej inteligencji, urządzenia będą mogły podejmować decyzje bez konieczności interwencji użytkownika. Inteligentne systemy zarządzania energią będą w stanie optymalizować zużycie na podstawie prognoz pogody i zwyczajów domowników. Systemy bezpieczeństwa będą mogły samodzielnie monitorować i reagować na potencjalne zagrożenia, a inteligentni asystenci głosowe staną się bardziej interaktywni i intuicyjni w obsłudze. To wszystko sprawi, że domy będą bardziej samowystarczalne, a życie w nich będzie jeszcze wygodniejsze.Rozwój standardów i protokołów komunikacji będzie kluczowe dla poprawy interoperacyjności urządzeń w inteligentnych domach. Organizacje takie jak Zigbee Alliance, Z-Wave Alliance czy Thread Group już pracują nad ujednoliceniem standardów, co ułatwi integrację różnych urządzeń i systemów. Inicjatywy takie jak Project CHIP (Connected Home over IP) mają na celu stworzenie jednego, uniwersalnego standardu komunikacji dla urządzeń inteligentnego domu, co pozwoli na łatwiejsze i bardziej bezproblemowe połączenia między nimi. Dzięki temu użytkownicy będą mogli swobodnie

wybierać urządzenia różnych producentów, mając pewność, że będą one ze sobą współpracować. Rozwój technologii chmurowych również odegra znaczącą rolę, umożliwiając lepszą synchronizację i zarządzanie urządzeniami z dowolnego miejsca na świecie.

9. Podsumowanie

Inteligentny dom to zaawansowany system technologiczny, który umożliwia zdalne sterowanie i automatyzację funkcji domowych, takich jak oświetlenie, ogrzewanie i monitoring. Artykuł analizuje rolę sztucznej inteligencji w tych systemach, w tym centralną jednostkę sterującą (hub), urządzenia wykonawcze, aplikacje mobilne i asystentów głosowych. Omawia również znaczenie IoT oraz protokołów komunikacyjnych jak MQTT i CoAP. Przyszłość AI w inteligentnych domach zapowiada większą autonomię systemów, lepszą interoperacyjność i wygodę użytkowników.

Literatura internetowe

1. https://botland.com.pl/blog/inteligentny-dom-jakie-rozwiazania-wybrac-i-kiedy/ (dostęp: 14.06.2024)
2. https://idg-online.pl/blog/czym-rozni-sie-nvr-od-rejestratora-sieciowego (dostęp: 14.06.2024)
3. https://lenalighting.pl/?catid=86&id=1669&layout=blog&view=article (dostęp: 14.06.2024)
4. https://smartpanda.pl/przyszlosc-iot-projekt-chip/ (dostęp: 14.06.2024)

Krystian Pupiec
Studenckie Koło naukowe informatyków „KOD”

dr inż. Bartosz Trybus
Opiekun naukowy

Apache Kafka: Teoria, architektura i praktyczna implementacja

Artykuł przedstawia Apache Kafkę jako rozproszoną platformę przetwarzania strumieniowego, która zdobyła ogromną popularność dzięki swojej zdolności do niezawodnego przesyłania i przetwarzania dużych ilości danych w czasie rzeczywistym. Omawia teoretyczne podstawy Kafki, w tym jej architekturę oraz kluczowe pojęcia, a także rolę Zookeepera w zarządzaniu klastrem. Ponadto, prezentuje proces instalacji i krótkiej konfiguracji Kafki. W praktycznej części artykułu zawarte jest przykładowe zastosowanie Kafki, skupiające się na systemie monitorowania zachowania użytkowników na stronie internetowej. Demonstruje on pełny cykl przesyłania i odbierania danych, włączając w to tworzenie tematów, pisanie producenta wysyłającego dane do Kafki oraz konsumenta odbierającego i przetwarzającego te dane. Artykuł kładzie nacisk na potencjalne zastosowania tej technologii w dziedzinie analizy danych oraz wykorzystanie jej w celu poprawy efektywności i jakości działania systemów informatycznych.

Słowa kluczowe: Apache Kafka, przetwarzanie strumieniowe, platforma rozproszona, monitorowanie zachowań użytkowników.

1. Wprowadzenie

Apache Kafka to system wiadomości typu publikuj/subskrybuj, stworzony aby ułatwić rozwiązanie tego problemu. Często opisywany jest jako "rozdzielony dziennik transakcji" lub też jako "rozdzielona platforma przesyłania strumieniowego". Podobnie jak dziennik transakcji w systemie plików lub bazie danych, Kafka zapewnia trwały zapis wszystkich transakcji, dzięki czemu mogą one być odtwarzane, aby spójnie zbudować stan systemu. Dodatkowo, dane w Kafka są przechowywane trwale, w porządku, i mogą być odczytywane deterministycznie. Ponadto, dane mogą być dystrybuowane w systemie, aby zapewnić dodatkową ochronę przed awariami oraz znaczne możliwości skalowania wydajności.[1]

Kafka, w porównaniu z innymi brokerami wiadomości, mocno kładzie nacisk na wydajność oraz skalowalność. Dzięki możliwości konfiguracji wielu parametrów, Apache Kafka pozwala dostosować sposób działania do specyfiki danych, z którymi pracujemy, ich rozmiaru oraz częstotliwości odczytu. Dodatkowo, możemy określić czas przechowywania wysyłanych wiadomości, co odróżnia go od tradycyjnych brokerów wiadomości, gdzie dane zazwyczaj są

[1] N. Narkhede, G. Shapira, T. Palino, *Kafka: The Definitive Guide*, s. 4.

usuwane po odebraniu przez adresata. W przeciwieństwie do typowego modelu publish/subscribe, Apache Kafka nie wymaga z góry znanego adresata przy wysyłaniu wiadomości, co bardziej przypomina działanie baz danych NoSQL.[2]

Artykuł ma na celu zbadanie i zrozumienie funkcji oraz możliwości Apache Kafki jako platformy przetwarzania strumieniowego danych. Badanie obejmuje jego architekturę, możliwości konfiguracyjne oraz praktyczne zastosowania w rzeczywistych scenariuszach. Opiera się ono na analizie teoretycznej i praktycznej Apache Kafki. Teoretyczna analiza obejmuje studium dokumentacji, publikacji naukowych i innych materiałów źródłowych dotyczących Kafki. W części praktycznej, wykorzystana jest przykładowa implementacja systemu opartego na Kafce, który demonstruje jej możliwości w realnych zastosowaniach.

2. Historia Apache Kafka oraz rozwój projektu

Kafka została stworzona w LinkedIn, aby obsłużyć wewnętrzne wymagania dotyczące przetwarzania strumieniowego, których nie można było zaspokoić za pomocą tradycyjnych systemów kolejek wiadomości. Jej pierwsza wersja została wydana w styczniu 2011 roku. Kafka szybko zyskała popularność i od tego czasu stała się jednym z najpopularniejszych projektów Fundacji Apache. Obecnie projekt jest głównie utrzymywany przez firmę Confluent, z pomocą innych firm, takich jak na przykład IBM, Yelp, Netflix.[3]

Apache Kafka, w dzisiejszym świecie jak i w ubiegłych latach, stosowana była również na wielu innych platformach.

Oto kilka przykładów zastosowań Apache Kafka w różnych firmach (informacje gdzie używana była Kafka według książki na 2013 rok)[4]:

6. LinkedIn: Apache Kafka jest używany w LinkedIn do przesyłania danych dotyczących aktywności oraz metryk operacyjnych. Te dane napędzają różne produkty, takie jak strona główna LinkedIn oraz LinkedIn Today, a także systemy analizy offline, takie jak Hadoop.
7. DataSift: W DataSift Kafka jest wykorzystywany jako kolektor do monitorowania zdarzeń oraz śledzenia konsumpcji strumieni danych przez użytkowników w czasie rzeczywistym.
8. Twitter: Twitter używa Kafki jako części infrastruktury Storm do przetwarzania strumieniowego.

[2] https://bulldogjob.pl/readme/apache-kafka-opis-dzialania-i-zastosowania (dostęp: 10.06.2024).
[3] https://www.conduktor.io/kafka/what-is-apache-kafka/ (dostęp: 10.06.2024).
[4] N. Garg, *Apache Kafka*, Packt Publishing Ltd, October 2013, s.8.

9. Foursquare: Kafka napędza komunikację online do online oraz online do offline w Foursquare. Jest używany do integracji monitoringu i systemów produkcyjnych z infrastrukturami offline opartymi na Hadoop.
10. Square: Square wykorzystuje Kafkę jako autobus do przekazywania wszystkich zdarzeń systemowych przez różne centra danych Square. Obejmuje to metryki, dzienniki, niestandardowe zdarzenia itp. Po stronie odbiorcy dane są przekazywane do Splunka, Graphite'a lub podobnego systemu do alertów w czasie rzeczywistym.

Obecnie z Apache Kafki korzystają takie firmy jak Uber i Netflix. Uber używa Kafki do przetwarzania danych o lokalizacji swoich kierowców i pasażerów w czasie rzeczywistym, podczas gdy Netflix używa Kafki do monitorowania i analizowania aktywności użytkowników na swojej platformie.[5]

Kafka więc zarówno w przeszłości jak i obecnie jest powszechnie stosowana przez duże firmy jako rozwiązanie do przetwarzania strumieniowego danych oraz integracji systemów.

3. Podstawowe pojęcia związane z Apache Kafka

Aby dobrze zrozumieć pracę i sposób działania Apache Kafki, dobrym rozwiązaniem jest zaznajomienie się z podstawowymi pojęciami używanymi w działaniu i architekturze tego rozwiązania.

Główne pojęcia związane z Apache Kafka[6]:

1. Producent (ang. producer) - Producent w Kafka tworzy nowe wiadomości. W innych systemach publikacji/subskrypcji może być nazywany wydawcą lub pisarzem. Wiadomości są produkowane do określonego tematu (topic). Domyślnie producent nie martwi się, do której partycji trafi konkretna wiadomość i równomiernie rozdziela wiadomości między wszystkie partycje tematu. W niektórych przypadkach producent kieruje wiadomości do konkretnych partycji, używając klucza wiadomości i partycjonera, który generuje hash klucza i mapuje go na określoną partycję. To zapewnia, że wszystkie wiadomości z danym kluczem trafią do tej samej partycji. Producent może także używać niestandardowego partycjonera, który przestrzega innych reguł biznesowych dotyczących mapowania wiadomości na partycje.

[5] https://howtointerview.pl/definicje/apache-kafka-opis-dzialania-i-zastosowania/10101/#Przyklady_firm_korzystajacych_z_Apache_Kafka (dostęp: 14.06.2024).
[6] N. Narkhede, G. Shapira, T. Palino, *Kafka: The Definitive Guide*, s. 6-9.

2. Konsument (ang. consumer) - Konsument w Kafka odczytuje wiadomości. W innych systemach publikacji/subskrypcji może być nazywany subskrybentem lub czytelnikiem. Konsument subskrybuje jeden lub więcej tematów i odczytuje wiadomości w kolejności, w jakiej zostały wyprodukowane. Konsument śledzi, które wiadomości już przeczytał, zapamiętując offset wiadomości. Offset to metadane – wartość całkowita, która stale rośnie – którą Kafka dodaje do każdej wyprodukowanej wiadomości. Każda wiadomość w danej partycji ma unikalny offset. Przechowując offset ostatnio przeczytanej wiadomości dla każdej partycji, konsument może zatrzymać się i ponownie rozpocząć bez utraty swojej pozycji. Konsumenci działają jako część grupy konsumentów, która składa się z jednego lub więcej konsumentów współpracujących w celu konsumowania tematu. Grupa zapewnia, że każda partycja jest konsumowana tylko przez jednego członka.
3. Broker (ang. broker) - Broker w systemie Kafka to pojedynczy serwer, który odbiera wiadomości od producentów, przypisuje im offsety i zapisuje je na dysku. Broker obsługuje również konsumentów, odpowiadając na ich żądania pobrania wiadomości z partycji i dostarczając wiadomości, które zostały zapisane na dysku. Pojedynczy broker może obsługiwać tysiące partycji i miliony wiadomości na sekundę, w zależności od specyfikacji sprzętu i jego wydajności.
4. Klaster (ang. cluster) - Klaster w systemie Kafka to grupa brokerów współpracujących ze sobą. Jeden z brokerów w klastrze pełni rolę kontrolera klastra, który jest automatycznie wybierany spośród aktywnych członków klastra. Kontroler jest odpowiedzialny za operacje administracyjne, takie jak przypisywanie partycji do brokerów i monitorowanie awarii brokerów. Partycja jest zarządzana przez jednego brokera w klastrze, nazywanego liderem partycji. Partycja może być przypisana do wielu brokerów, co zapewnia replikację partycji i redundancję danych. Wszystkie operacje konsumentów i producentów na danej partycji muszą łączyć się z jej liderem. Klaster Kafka umożliwia trwałe przechowywanie wiadomości przez określony czas dzięki mechanizmowi retencji.
5. Temat (ang. topic) - Temat w Apache Kafka to logiczna kategoria lub strumień danych, do którego producenci mogą publikować dane, a konsumenci mogą subskrybować i odbierać te dane w miarę ich przybywania. Tematy są tworzone przez administratorów i mogą mieć wielu producentów oraz konsumentów. Mogą być partycjonowane, co pozwala na równoległe przetwarzanie wiadomości i zwiększa skalowalność. Kiedy wiadomość jest publikowana do tematu, jest dodawana na koniec logu tego tematu.

Partycje w tematach są podstawową jednostką równoległości, a każda partycja to sekwencja wiadomości przechowywana na jednym węźle brokera. Replikacje partycji zapewniają odporność na awarie i gwarantują, że dane nie zostaną utracone w przypadku awarii brokera.[7]

6. Partycja (ang. partition) - Partycja w Apache Kafka to część tematu, do której trafiają komunikaty. Gwarantuje ona zachowanie kolejności komunikatów wewnątrz niej, dlatego klucz wiadomości jest istotny przy ich przydzielaniu do partycji. Funkcja implementująca interfejs Partitioner określa, do której partycji trafi dany komunikat, najczęściej na podstawie hash(klucz)%liczba_partycji. Większa liczba partycji pozwala na zrównoleglenie przetwarzania, ponieważ wielu konsumentów może jednocześnie czytać z różnych partycji, zachowując kolejność przetwarzania wewnątrz każdej partycji. Partycje mogą być rozdzielone między wielu brokerów, co umożliwia przetwarzanie danych przez wiele maszyn, zwiększając tym samym szybkość systemu opartego na Kafka. Konsumenci są przypisywani do określonych partycji, a proces rebalancingu zapewnia, że nowi konsumenci mogą przejąć część ruchu, równoważąc obciążenie systemu. Jednak liczba konsumentów nie może przekraczać liczby partycji, aby zachować porządek przetwarzania komunikatów. Aby określić, ile partycji jest potrzebnych, należy znać oczekiwane obciążenie systemu oraz czas przetwarzania komunikatów.[8]
7. Apache Zookeeper – usługa, której Apache Kafka używa do utrzymywania i koordynowania brokerów.[9] Zookeeper odgrywa kluczową rolę w architekturze Kafka, zarządzając stanem klastra i umożliwiając jego prawidłowe funkcjonowanie. Kafka nie może działać bez Zookeepera. Jedną z jego funkcji jest wybór kontrolera, który jest jednym z brokerów odpowiedzialnych za zarządzanie partycjami, wybór liderów partycji, tworzenie tematów i replik. Zookeeper rejestruje również stan każdego brokera w klastrze Kafka, umożliwiając producentom i konsumentom uzyskanie aktualnych informacji o brokerach. Przechowuje także informacje o tematach, takie jak liczba partycji i specyficzne parametry konfiguracyjne. Ponadto Zookeeper zarządza limitami przepustowości dla klientów, określając maksymalną ilość danych, jakie mogą oni czytać i zapisywać. Wreszcie, Zookeeper przechowuje zasady autoryzacji w formie list kontroli dostępu (ACLs), które definiują role użytkowników oraz ich uprawnienia do

[7] T.P. Raptis, A. Passarella, *A Survey on Networked Data Streaming With Apache Kafka*, 2023
[8] https://softwareskill.pl/apache-kafka-ile-partycji (dostęp: 10.06.2024).
[9] https://www.ibm.com/docs/pl/oala/1.3.7?topic=components-apache-kafka-cluster (dostęp: 10.06.2024).

odczytu i zapisu na poszczególnych tematach. Dzięki tym funkcjom Zookeeper zapewnia, że klaster Kafka działa stabilnie i efektywnie, zarządzając jego stanem oraz konfiguracjami.[10]

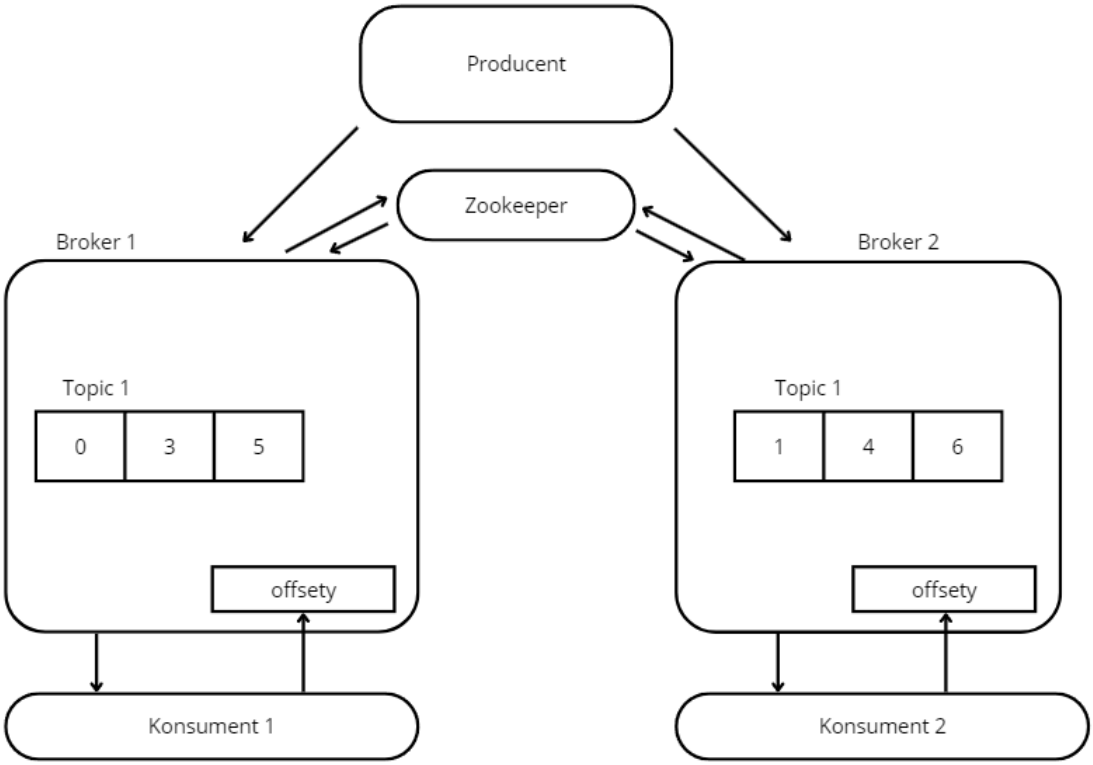
Warto znać powyższe pojęcia, ponieważ są one fundamentalne dla zrozumienia i efektywnego wykorzystania Apache Kafki w zarządzaniu i przetwarzaniu strumieni danych w czasie rzeczywistym.

4. Architektura Apache Kafka

Architektura Apache Kafki jest oparta na rozproszonym, partycjonowanym i replikowanym dzienniku zatwierdzeń. Kafka utrzymuje strumienie wiadomości w kategoriach zwanych tematami. Procesy, które publikują wiadomości do tematu Kafka, nazywane są producentami, a procesy subskrybujące tematy i przetwarzające strumień opublikowanych wiadomości nazywane są konsumentami. Kafka działa jako klaster, który składa się z jednego lub więcej serwerów, z których każdy nazywany jest brokerem. Producenci wysyłają wiadomości przez sieć do klastra Kafki, który następnie dostarcza je konsumentom. W przypadku każdego tematu klaster Kafka utrzymuje partycje w celu skalowania, równoległego przetwarzania i tolerancji na awarie. Każda partycja jest uporządkowaną, niezmienną sekwencją wiadomości, która jest ciągle dodawana do dziennika zatwierdzeń. Wiadomości w partycjach są przypisywane do unikalnego numeru zwanego offsetem. Offset jest kontrolowany przez konsumenta, który zazwyczaj przetwarza kolejną wiadomość na liście, chociaż może konsumować wiadomości w dowolnej kolejności, ponieważ klaster Kafka przechowuje wszystkie opublikowane wiadomości przez konfigurowalny okres. Konsumenci mogą przypisywać sobie nazwę grupy konsumentów, a każda wiadomość jest dostarczana do jednego konsumenta w każdej subskrybowanej grupie konsumentów. Jeśli wszyscy konsumenci mają różne grupy konsumentów, wiadomości są rozsyłane do każdego konsumenta. Kafka może być używana jako tradycyjny broker wiadomości z wysoką przepustowością, wbudowanym partycjonowaniem, replikacją i tolerancją na awarie, co czyni ją dobrym rozwiązaniem dla aplikacji przetwarzających duże ilości wiadomości. Może być również używana do śledzenia aktywności na stronach internetowych w czasie rzeczywistym, do agregacji logów oraz jako rozwiązanie do strumieniowego przetwarzania danych. Na Rysunek 35 została przedstawiona architektura Apache Kafki z producentem, Zookeeperem oraz dwoma brokerami. Każdy broker

[10] M. Kumar, Ch. Singh, *Building Data Streaming Applications with Apache Kafka*, Pact Publishing Ltd., 2017, s. 37-38.

zawiera partycje dla Topic 1, które są odczytywane przez dwóch konsumentów, każdy z własnymi offsetami.[11]



Rysunek 35. Architektura przetwarzania danych w Apache Kafka z Zookeeper i dwoma brokerami. Źródło: opracowanie własne na podstawie https://softwareskill.pl/apache-kafka-wprowadzenie .

Kafka jest systemem open source zaprojektowanym z kilkoma kluczowymi cechami. Po pierwsze, Kafka zapewnia trwałe wiadomości dzięki zastosowaniu struktur dyskowych O(1), które gwarantują stałą wydajność czasową nawet przy bardzo dużych wolumenach przechowywanych danych. To pozwala na utrzymanie dużych ilości danych bez utraty informacji. Po drugie, Kafka charakteryzuje się wysoką przepustowością, ponieważ jest zaprojektowana do pracy na standardowym sprzęcie i obsługiwania milionów wiadomości na sekundę. Dzięki temu, system jest w stanie sprostać wymaganiom dużych wolumenów danych. Trzecia cecha to rozproszenie. Kafka wspiera partycjonowanie wiadomości na serwery Kafki i dystrybucję konsumpcji na klaster maszyn konsumentów, przy zachowaniu porządku na poziomie partycji. To pozwala na efektywne zarządzanie dużymi zestawami danych. Kolejną istotną cechą jest wsparcie wielu klientów. Kafka wspiera łatwą integrację klientów z różnych platform, takich jak Java, .NET, PHP, Ruby i Python, co sprawia, że jest wszechstronna i elastyczna w różnych środowiskach. Ostatnią cechą jest przetwarzanie w czasie rzeczywistym. Wiadomości produkowane przez wątki producentów są natychmiast widoczne dla wątków

[11] https://softwareskill.pl/apache-kafka-wprowadzenie (dostęp: 12.06.2024).

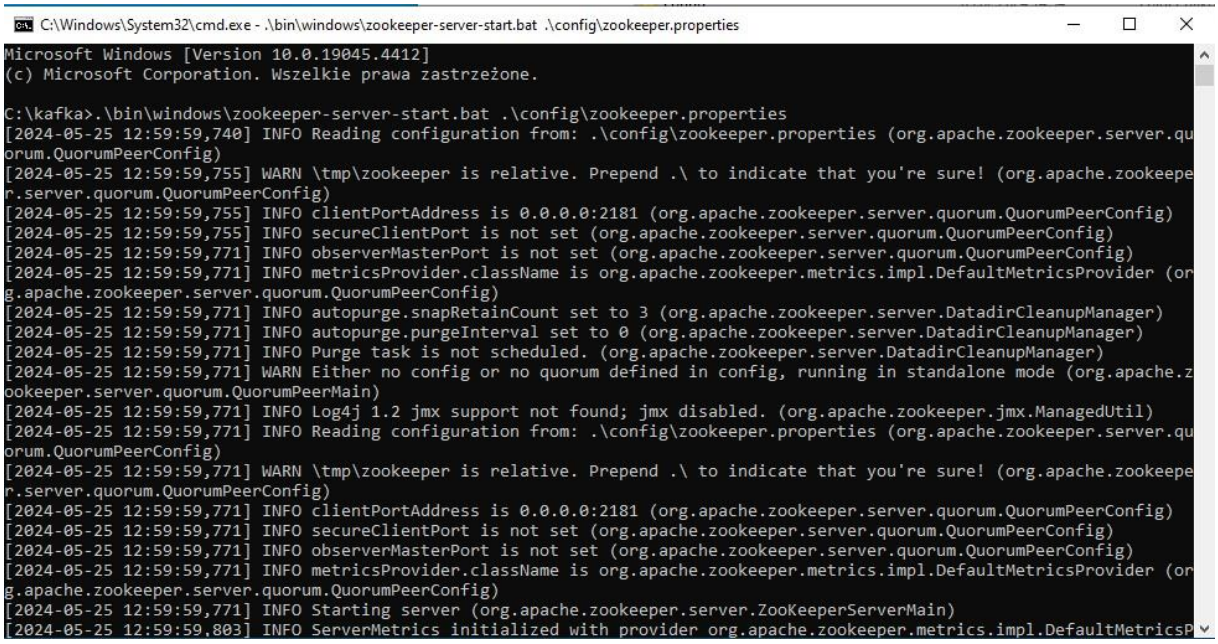
konsumentów, co jest krytyczne dla systemów opartych na zdarzeniach. Dzięki temu Kafka jest idealnym rozwiązaniem dla aplikacji wymagających natychmiastowego przetwarzania danych.[12]

5. Przykładowa implementacja

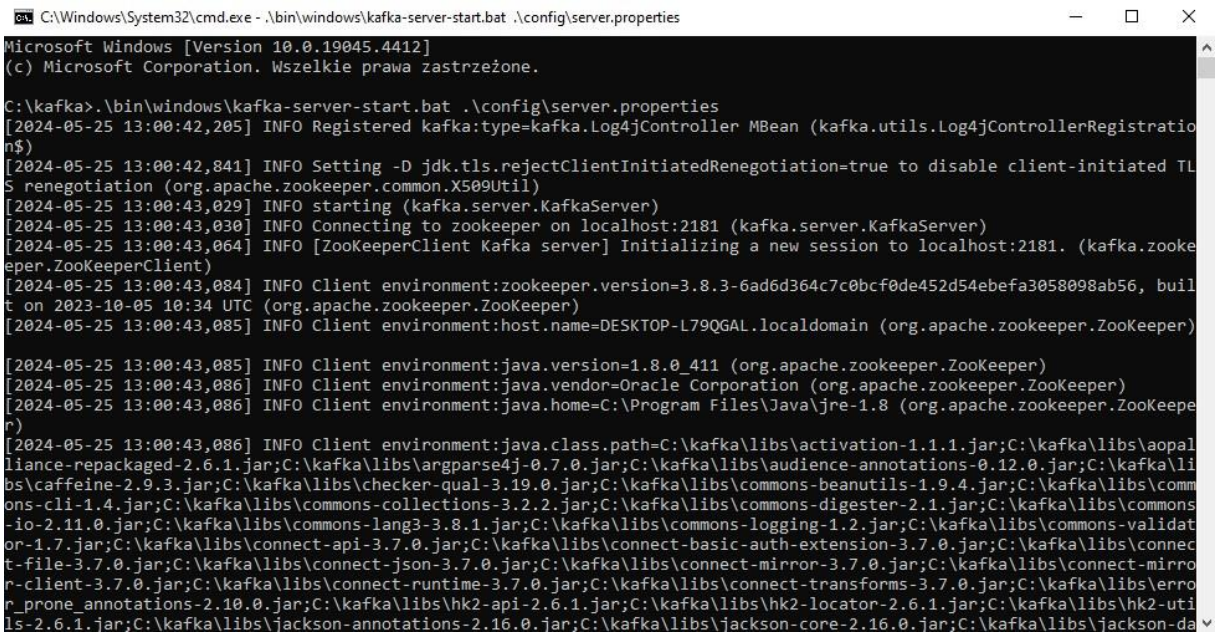
Implementacja rozwiązania będzie przestawiona w systemie Windows. Przed pierwszym użyciem Apache Kafki należy ją pobrać i zainstalować. Apache Kafkę można pobrać z oficjalnej strony *https://kafka.apache.org/downloads*. Kolejnym krokiem jest rozpakowanie pobranego pliku, który jest archiwum, oraz przeniesienie go do dowolnego folderu. Następnie można zmienić konfigurację Zookeepera oraz serwera, tak aby logi oraz snapshoty był zapisywane we wskazanej lokalizacji. Aby to wykonać należy zmienić ścieżkę w parametrze *log.dirs* w pliku *server.properties* oraz ścieżkę w parametrze *dataDir* w pliku *zookeeper.properties*. Oba te pliki znajdują się w folderze *config*. W pliku *server.properties* można również zmienić wartość parametru *log.retention.hours*, który określa minimalny wieku pliku dziennika w godzinach, aby był uznany za kandydata do usunięcia ze względu na wiek. W przypadku tego zastosowania Parametr *log.retention.hours* można zmienić na wartość ′−1′, aby dane nigdy nie były usuwane automatycznie ze względu na wiek.

Po pobraniu, zainstalowaniu oraz ewentualnej wstępnej konfiguracji, należy uruchomić serwer Zookeeper oraz serwer Kafki. Zookeeper został uruchomiony poleceniem *.\bin\windows\zookeeper-ssrver-start.bat .\config\zookeeper.properties*, tak jak na Rysunek 36 natomiast drugi serwer poleceniem *.\bin\windows\kafka-server-start.bat .\cconfig\server.properties*, tak jak na Rysunek 37.[13]

[12] K. M. M. Thein, *Apache Kafka: Next Generation Distributed Messaging System*, 2014, s. 2-3.
[13] https://kafka.apache.org/quickstart (dostęp: 12.06.2024).



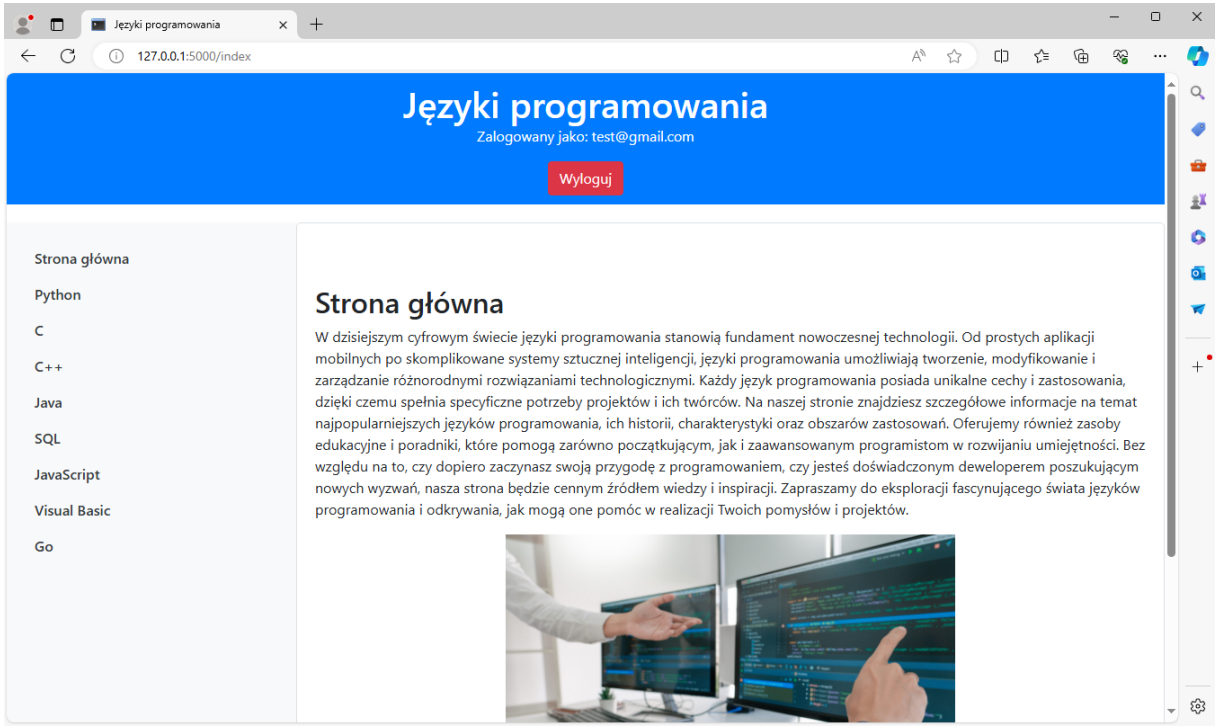
Rysunek 36. Uruchomienie Zookeepera
Źródło: opracowanie własne.



Rysunek 37. Uruchomienie serwera Kafki
Źródło: opracowanie własne.

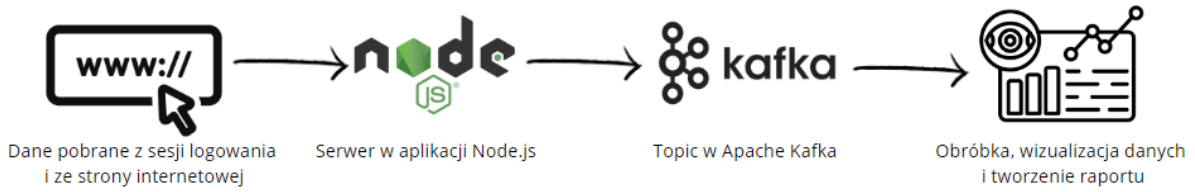
Aby pokazać zastosowanie Apache Kafki do monitorowania zachowania użytkowników na stronie internetowej, została wykorzystana prosta strona internetowa, której główny interfejs przedstawia Rysunek 38, stworzona za pomocą frameworka Flask. Łączy się ona z bazą danych w języku MySQL, tak aby rejestracja i logowanie na stronie były możliwe. Użytkownicy mogą się zarejestrować, podając swoje imię, email, wiek, miasto i hasło, które są zapisywane w bazie

danych. Po zalogowaniu się, dane użytkownika są przechowywane w sesji, umożliwiając dostęp do strony głównej, która wyświetla informacje o użytkowniku. Stworzony został również serwer w aplikacji Node.js, który przekazuje dane ze strony internetowej do Kafki. Konsument, który został napisany w języku Python, odczytuje wiadomości z topiców i je przetwarza.



Rysunek 38. Interfejs graficzny prostej strony internetowej
Źródło: opracowanie własne.

Sposobem w jaki będzie wykorzystana analiza zachowań użytkowników na stronie internetowej jest monitorowanie poszczególnych kliknięć użytkowników. Po kliknięciu w dane łącze na stronie internetowej, informacja o tym zdarzeniu wraz z danymi aktualnie zalogowanego użytkownika trafia do serwera w aplikacji Node.js, z którego później, zebrane dane trafiają do topicu w Apache Kafka. Następnie co określony czas, automatycznie uruchamiany jest program w języku Python, który odpowiednio odczytuje, przekształca dane i tworzy z nich wizualizacje. Pozwala to na sprawdzenie, które łącze jest najczęściej odwiedzane oraz dodatkowo, dla każdego łącza, tworzone są wykresy z informacją o użytkownikach w nie klikających, a konkretniej, informacje o wieku, płci oraz mieście zamieszkania. Może to być wykorzystane do analizy, jaka grupa wiekowa lub płeć interesuje się danym tematem lub z jakiej części świata lub kraju, ludzie najczęściej odwiedzają zakładki o danym temacie. Struktura tego rozwiązania została przedstawiona na Rysunek 39.



Rysunek 39. Struktura implementacji rozwiązania
Źródło: opracowanie własne.

Aplikacja Node.js, zbudowana przy użyciu Express.js i kafkajs, służy do wysyłania danych do Apache Kafka. Aplikacja uruchamia serwer, który nasłuchuje na wskazanym porcie. Podczas uruchamiania serwera inicjalizowane są połączenia z adminem i producentem Kafka. Gdy serwer otrzymuje zapytanie GET na odpowiedniej ścieżce, wyciąga dane z zapytania, takie jak wiek, miasto, płeć, temat i identyfikator. Aplikacja tworzy odpowiednie tematy w Kafka, jeśli jeszcze nie istnieją. Następnie dane są wysyłane jako wiadomości do odpowiednich tematów: wiek do tematu *{topic}_age*, miasto do *{topic}_city*, płeć do *{topic}_gender*, a informacje o kliknięciach do *{topic}_clicks*. W przypadku powodzenia, serwer odpowiada komunikatem potwierdzającym wysłanie wiadomości, a w przypadku błędu, odpowiedzią jest komunikat o błędzie. Kod serwera w aplikacji Node.js we fragmentach przedstawiono na Rysunek 40, Rysunek 41, Rysunek 42, Rysunek 43.

```
// Importowanie wymaganych modułów

// Express.js - framework do budowy aplikacji sieciowych
const express = require('express');
// kafkajs - biblioteka do komunikacji z Apache Kafka
const { Kafka } = require('kafkajs');
// axios - biblioteka do wykonywania żądań HTTP
const axios = require('axios');

// Inicjalizacja aplikacji Express.js
const app = express();

// Inicjalizacja klienta Kafka

// Identyfikator klienta Kafka
const kafka = new Kafka({
  clientId: 'my-app',
  // Adresy brokerów Kafka
  brokers: ['localhost:9092']
});

// Inicjalizacja admina Kafka oraz producenta Kafka

// Admin - do operacji administracyjnych na Kafka
const admin = kafka.admin();
// Producent - do wysyłania wiadomości do Kafka
const producer = kafka.producer();
```

Rysunek 40. Serwer w aplikacji Node.js – część 1
Źródło: opracowanie własne.

```
// Funkcja sprawdzająca i ewentualnie tworząca temat w Kafka
const createTopicIfNotExists = async (topic) => {
  // Nawiązanie połączenia z adminem Kafka
  await admin.connect();
  // Pobranie listy tematów z Kafka
  const topics = await admin.listTopics();
  // Sprawdzenie, czy temat już istnieje
  if (!topics.includes(topic)) {
    // Tworzenie tematu, jeśli nie istnieje
    await admin.createTopics({
      topics: [{ topic }]
    });
  }
};

// Ustawienie nagłówków CORS
app.use((req, res, next) => {
  res.header("Access-Control-Allow-Origin", "*");
  res.header("Access-Control-Allow-Headers", "Origin, X-Requested-With, Content-Type, Accept");
  next();
});

// Połączenie z Kafka i uruchomienie serwera
(async () => {
  try {
    // Nawiązanie połączenia z producentem Kafka
    await producer.connect();
    // Nawiązanie połączenia z adminem Kafka
    await admin.connect();
    // Uruchomienie serwera Express.js
    await app.listen(3000, () => console.log('Serwer działa na porcie 3000'));
  } catch (error) {
    // Obsługa błędów inicjalizacji serwera
    console.error('Błąd podczas inicjalizacji serwera:', error);
  }
})();
```

Rysunek 41. Serwer w aplikacji Node.js - część 2
Źródło: opracowanie własne.

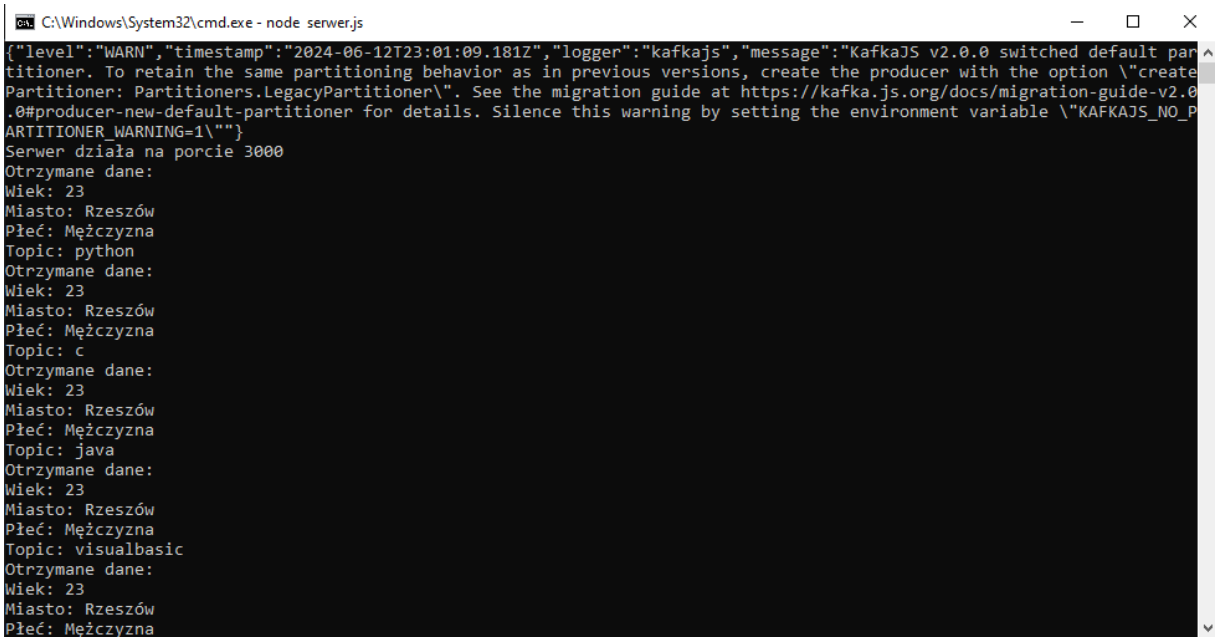
```
66    // Obsługa zapytania GET
67    app.get('/sendToKafka', async (req, res) => {
68      // Pobranie danych z zapytania GET
69      const { age, city, gender, topic, id } = req.query;
70
71      // Sprawdzenie, czy ID to #home
72      if (topic === 'home') {
73        res.header("Access-Control-Allow-Origin", "*");
74        // Pominięcie tworzenia tematu, jeśli ID=#home
75        return res.send('Skipping topic creation for ID=#home');
76      }
77      else{
78        console.log('Otrzymane dane:');
79        console.log('Wiek:', age);
80        console.log('Miasto:', city);
81        console.log('Płeć:', gender);
82        console.log('Topic:', topic);
83        // Nazwa tematu dla wieku
84        const ageTopic = `${topic}_age`;
85        // Nazwa tematu dla miasta
86        const cityTopic = `${topic}_city`;
87        // Nazwa tematu dla płci
88        const genderTopic = `${topic}_gender`;
89        // Nazwa tematu dla kliknięć
90        const clicksTopic = `${topic}_clicks`;
91
92        try {
93          // Sprawdzenie i ewentualne tworzenie tematów w Kafka
94          await createTopicIfNotExists(ageTopic);
95          await createTopicIfNotExists(cityTopic);
96          await createTopicIfNotExists(genderTopic);
97          await createTopicIfNotExists(clicksTopic);
98
99          // Wysyłanie wiadomości do odpowiednich tematów Kafka
100         await producer.send({
101           topic: ageTopic,
102           messages: [{ value: age.toString() }],
103         });
104
105         await producer.send({
106           topic: cityTopic,
107           messages: [{ value: city.toString() }],
108         });
109
110         await producer.send({
111           topic: genderTopic,
112           messages: [{ value: gender.toString() }],
113         });
114
115         await producer.send({
116           topic: clicksTopic,
117           messages: [{ value: '1' }],
118         });
119
```

Rysunek 42. Serwer w aplikacji Node.js - część 3
Źródło: opracowanie własne.

```
120         // Dodanie nagłówka Access-Control-Allow-Origin
121         res.header("Access-Control-Allow-Origin", "*");
122         // Odpowiedź na zapytanie
123         res.send('Wiadomość wysłana do Kafka');
124       } catch (error) {
125         // Obsługa błędów wysyłania do Kafka
126         console.error('Błąd podczas wysyłania wiadomości do Kafka:', error);
127         // Wysłanie odpowiedzi błędu
128         res.status(500).send('Wystąpił błąd podczas wysyłania wiadomości do Kafka');
129       }
130       }
131
132   });
```

Rysunek 43. Serwer w aplikacji Node.js - część 4
Źródło: opracowanie własne.

Rysunek 44 przedstawia komunikaty podczas pracy serwera w aplikacji Node.js.



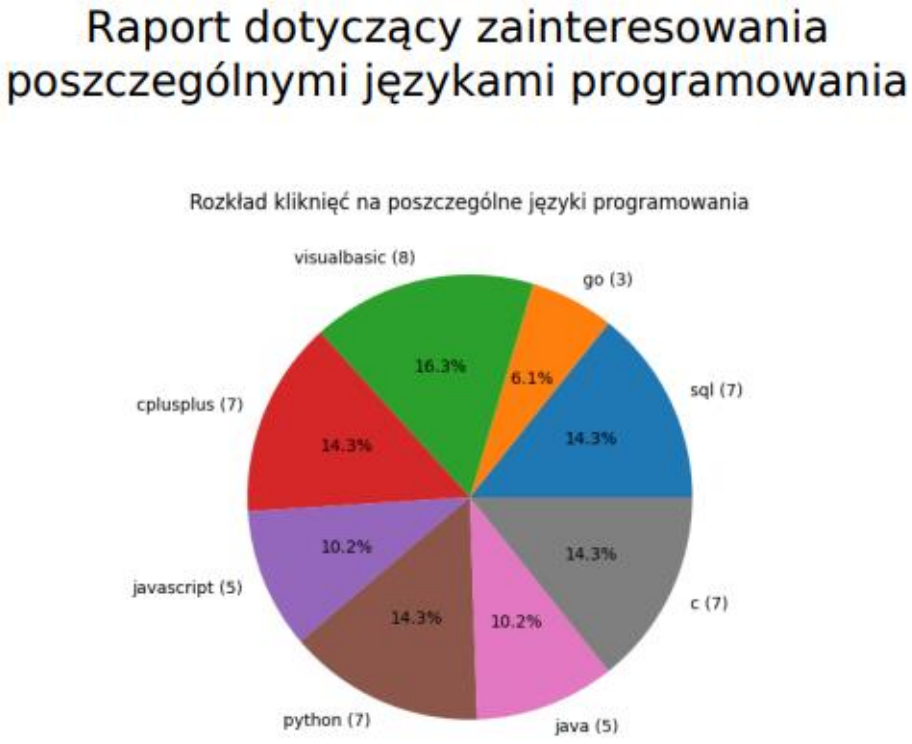
Rysunek 44. Komunikaty podczas pracy serwera w aplikacji Node.js
Źródło: opracowanie własne

Kolejnym elementem implementacji jest kod tworzący raport na podstawie kliknięć użytkowników. Kod konfiguruje serwery Kafki i inicjalizuje klienta administracyjnego do pobierania tematów. Główna funkcja odczytuje wiadomości z wszystkich tematów, subskrybuje je i przetwarza odebrane wiadomości. Dodatkowo, kod agreguje dane z wiadomości, co pozwala na analizę i dalsze przetwarzanie tych informacji. Po odczytaniu i agregacji danych, wyniki są zapisane w pliku PDF, aby przedstawić wizualizacje zgromadzonych danych. Na Rysunek 45 jest przedstawiony fragment tego kodu dotyczący tylko Apache Kafka, inne elementy kodu obsługujące tworzenie wykresów i raportu nie zostały przedstawione.

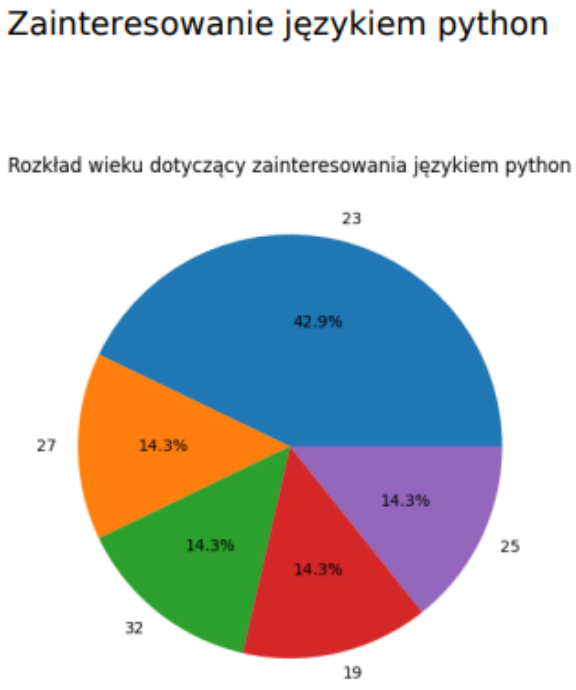
```python
from kafka import KafkaConsumer, KafkaAdminClient
# Konfiguracja Kafka
bootstrap_servers = ['localhost:9092']
# Funkcja do pobierania wszystkich tematów z Kafka
def get_all_topics():
    try:
        # Inicjalizacja klienta administracyjnego Kafka
        admin_client = KafkaAdminClient(bootstrap_servers=bootstrap_servers)
        # Pobranie listy tematów
        topics = admin_client.list_topics()
        return topics
    except Exception as e:
        print("Błąd podczas pobierania tematów Kafka:", e)
        return []

# Funkcja do odczytywania wiadomości z Kafka
def read_kafka_messages():
    all_data = {}
    try:
        # Pobranie wszystkich tematów
        topics = get_all_topics()
        if not topics:
            print("Brak dostępnych tematów.")
            return all_data
        # Inicjalizacja konsumenta Kafka
        consumer = KafkaConsumer(bootstrap_servers=bootstrap_servers, auto_offset_reset='earliest', enable_auto_commit=False)
        # Subskrypcja wszystkich tematów
        consumer.subscribe(topics)
        print("Subskrybowane tematy:", topics)

        # Ustawienie czasu początkowego i timeoutu
        start_time = time.time()
        timeout = 10  # Czas w sekundach do zakończenia oczekiwania na wiadomości

        # Pętla do odczytywania wiadomości z Kafka
        while time.time() - start_time < timeout:
            # Polling wiadomości
            messages = consumer.poll(timeout_ms=1000)
            if not messages:
                print("Brak nowych wiadomości, czekam...")
                continue

            # Przetwarzanie odebranych wiadomości
            for topic_partition, records in messages.items():
                topic = topic_partition.topic
                for record in records:
                    value = record.value.decode('utf-8')
                    if topic not in all_data:
                        all_data[topic] = []
                    all_data[topic].append(value)
                    print(f'Odebrano wiadomość z tematu {topic}: {value}')

        # Zamknięcie konsumenta
        consumer.close()
    except Exception as e:
        print("Błąd podczas odczytywania wiadomości z Kafka:", e)

    return all_data
```

Rysunek 45. Kod klienta odczytującego wiadomości z Kafka – fragment dotyczący Apache Kafka
Źródło:: opracowanie własne.

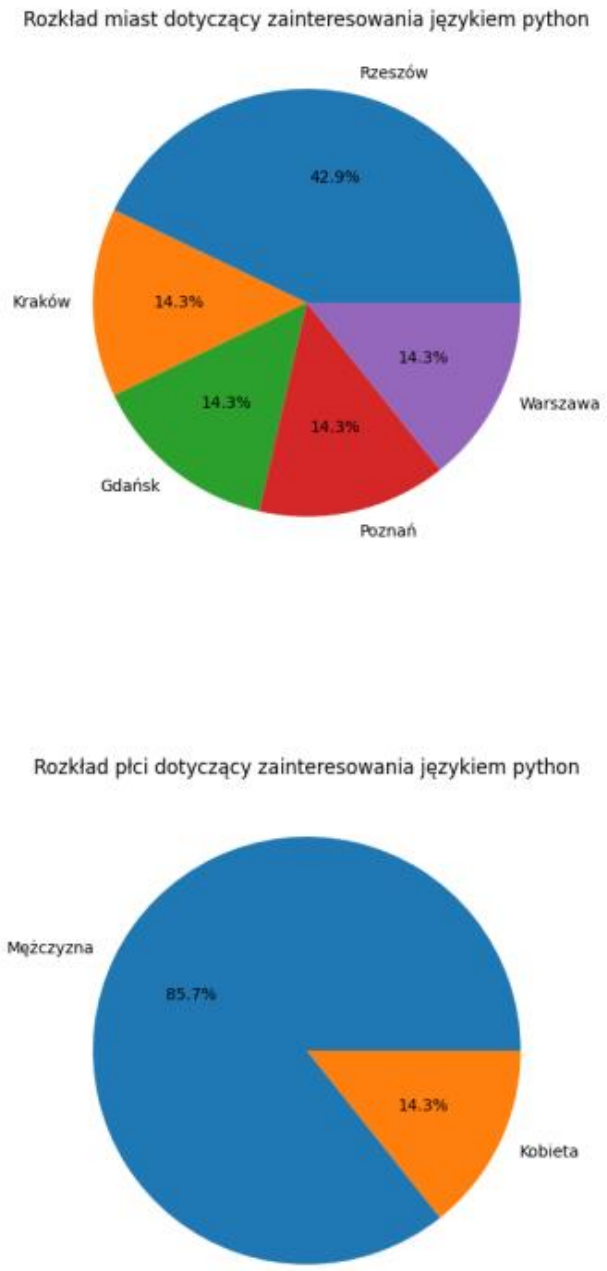
Na Rysunek 46 widoczny jest wykres przedstawiający ilość kliknięć w poszczególne zakładki na stronie internetowej. Na Rysunek 47 oraz Rysunek 48 przedstawione są wykresy dla przykładowej wartości, które prezentują odpowiednio rozkład wieku dotyczący zainteresowania konkretnym tematem oraz rozkład miast i płci dotyczącej zainteresowania konkretnym tematem.



Rysunek 46. Rozkład kliknięć w poszczególne zakładki - ogólne zainteresowanie
Źródło: opracowanie własne.



Rysunek 47. Rozkład wieku dotyczący zainteresowania danym tematem
Źródło: opracowanie własne.



Rysunek 48. Rozkład miast oraz rozkład płci dotyczący zainteresowania danym tematem
Źródło: opracowanie własne.

Przykład implementacji monitorowania zachowania użytkowników na stronie internetowej przy użyciu Apache Kafka ilustruje, jak wykorzystać tę platformę do zbierania, przetwarzania i analizowania strumieni danych. Kafka umożliwia efektywne przesyłanie informacji o kliknięciach oraz danych demograficznych użytkowników do różnych tematów, co pozwala na ich dalsze przetwarzanie i analizę. Dzięki temu rozwiązaniu możliwe jest generowanie w czasie rzeczywistym wizualizacji i raportów, co wspiera decyzje biznesowe dotyczące optymalizacji treści oraz personalizacji doświadczeń użytkowników na stronie internetowej.

6. Podsumowanie

Artykuł omawia Apache Kafkę jako rozproszoną platformę przetwarzania strumieniowego, która zyskała popularność dzięki swojej zdolności do niezawodnego przesyłania i przetwarzania dużych ilości danych w czasie rzeczywistym. Dane w Kafka są przechowywane trwale, w określonym porządku, i mogą być odczytywane deterministycznie, co gwarantuje wysoką wydajność i skalowalność. Kafka wyróżnia się możliwością konfiguracji wielu parametrów, co pozwala na dostosowanie jej działania do specyficznych wymagań dotyczących rozmiaru danych i częstotliwości odczytu. Istotną cechą Kafki jest możliwość określenia czasu przechowywania wysyłanych wiadomości, co różni ją od tradycyjnych brokerów wiadomości, gdzie dane są zazwyczaj usuwane po odebraniu przez adresata. Architektura Kafki opiera się na rozproszonym, partycjonowanym i replikowanym dzienniku zatwierdzeń, co pozwala na skalowanie, równoległe przetwarzanie i tolerancję na awarie. W artykule przedstawiono także proces instalacji i konfiguracji Kafki, oraz przykładowe zastosowanie w systemie monitorowania zachowań użytkowników na stronie internetowej. Kafka została wykorzystana do przesyłania danych dotyczących kliknięć użytkowników, które są następnie analizowane i wizualizowane w celu lepszego zrozumienia interakcji użytkowników z witryną.

Bardzo dużą zaletą Kafki jest możliwość obsługi milionów wiadomości na sekundę, co czyni ją idealnym rozwiązaniem dla aplikacji wymagających natychmiastowego przetwarzania dużych wolumenów danych. Dzięki wysokiej przepustowości, trwałości danych, możliwości skalowania oraz wsparciu dla wielu klientów, Apache Kafka jest wszechstronnym narzędziem, które znajduje zastosowanie w wielu branżach i scenariuszach przetwarzania danych w czasie rzeczywistym.

Literatura

1. Garg N., *Apache Kafka*, Packt Publishing Ltd., October 2013.
2. Kumar M., Singh Ch., *Building Data Streaming Applications with Apache Kafka*, Pact Publishing Ltd., 2017.
3. Narkhede N., Shapira G., Palino T., *Kafka: The Definitive Guide*, O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.
4. Raptis T. P., Passarella A., *A Survey on Networked Data Streaming With Apache Kafka*, 2023.
5. Thein K. M. M., *Apache Kafka: Next Generation Distributed Messaging System*, 2014.

Źródła internetowe

1. https://bulldogjob.pl/readme/apache-kafka-opis-dzialania-i-zastosowania (dostęp: 10.06.2024).
2. https://howtointerview.pl/definicje/apache-kafka-opis-dzialania-i-zastosowania/10101/#Przyklady_firm_korzystajacych_z_Apache_Kafka (dostęp: 14.06.2024).
3. https://kafka.apache.org/quickstart (dostęp: 12.06.2024).
4. https://softwareskill.pl/apache-kafka-ile-partycji (dostęp: 10.06.2024).
5. https://softwareskill.pl/apache-kafka-wprowadzenie (dostęp: 12.06.2024).
6. https://www.conduktor.io/kafka/what-is-apache-kafka/ (dostęp: 10.06.2024).
7. https://www.ibm.com/docs/pl/oala/1.3.7?topic=components-apache-kafka-cluster (dostęp: 10.06.2024).

Krystian Pupiec
Studenckie Koło naukowe informatyków „KOD”

Dr inż. Bartosz Trybus
Opiekun naukowy

Wykorzystanie biblioteki Selenium w Pythonie do web scrapingu

Artykuł przedstawia techniki web scrapingu z wykorzystaniem biblioteki Selenium, koncentrując się na jej roli, metodach i praktycznych zastosowaniach. Web scraping to proces automatycznego pobierania danych z witryn internetowych, istotny dla analiz danych i aplikacji komercyjnych. Selenium to narzędzie do automatyzacji przeglądania stron, umożliwia scrapowanie dynamicznych treści generowanych przez JavaScript oraz obsługę AJAX. Artykuł omawia również działanie, architekturę oraz instalację i konfigurację Selenium. Przedstawia techniki scrapingu dynamicznych treści i zarządzania opóźnieniami ładowania stron. W artykule zaprezentowano również praktyczne zastosowanie użycia biblioteki Selenium. Analizowane są także wyzwania i ograniczenia web scrapingu przy pomocy Selenium. Artykuł łączy teoretyczne wyjaśnienia z praktycznymi przykładami wykorzystania biblioteki Selenium w web scrapingu.

Słowa kluczowe: web scraping, Selenium, automatyzacja przeglądania stron, dynamiczne treści.

1. Wprowadzenie

Automatyczne gromadzenie danych z Internetu, znane jako web scraping, nie jest nową praktyką – wcześniej określano je mianem screen scrapingu, data miningu czy web harvestingu. Obecnie termin web scraping jest najbardziej powszechny. Potocznie można to przetłumaczyć jako skrobanie sieci. Teoretycznie, web scraping to proces zbierania danych z internetu bez użycia interfejsu API lub przeglądarki internetowej obsługiwanej przez człowieka. Najczęściej polega to na napisaniu programu, który wysyła zapytania do serwera, pobiera dane (zwykle w formie HTML i innych plików tworzących strony internetowe), a następnie analizuje te dane, aby wyciągnąć potrzebne informacje. W praktyce, web scraping obejmuje szeroki zakres technik i technologii programistycznych, takich jak analiza danych, przetwarzanie języka naturalnego i bezpieczeństwo informacji. Umożliwia to dostęp do ogromnej ilości danych w sposób, który nie jest możliwy przy użyciu przeglądarki. Przeglądarki są użyteczne do wykonywania JavaScriptu, wyświetlania obrazów i układania obiektów w bardziej czytelny sposób, ale narzędzia do automatycznego zbierania danych potrafią szybko zbierać i

przetwarzać duże ilości danych. Mogą one analizować dane z tysięcy, a nawet milionów stron jednocześnie, co jest niemożliwe dla tradycyjnych wyszukiwarek internetowych[1].

Techniki web scrapingu obejmują szeroki wachlarz metod i narzędzi, które umożliwiają wydobywanie danych z witryn internetowych. Wybór techniki zależy od charakterystyki strony, rodzaju danych, które chcemy pobrać, oraz preferencji osoby przeprowadzającej scraping. Jedną z popularnych metod jest użycie bibliotek programistycznych takich jak Beautiful Soup, lxml lub Scrapy. Narzędzia te oferują zaawansowane funkcje do analizy struktury HTML i ekstrakcji danych z poszczególnych elementów strony. Mogą być używane w różnych językach programowania, w tym w Pythonie, a także w innych, które obsługują parsowanie HTML. Inną metodą jest bezpośrednie pobieranie danych za pomocą protokołu HTTP z API lub stron internetowych. W tym przypadku skrypt lub program wysyła zapytania HTTP do określonych adresów URL i otrzymuje odpowiedzi, które są następnie przetwarzane w celu wydobycia potrzebnych informacji. Scraping stron dynamicznych, wykorzystujących technologie takie jak JavaScript, wymaga bardziej zaawansowanych technik. W takich sytuacjach często korzysta się z narzędzi takich jak Selenium, Puppeteer czy Splash. Pozwalają one na sterowanie przeglądarką, symulowanie interakcji użytkownika oraz pobieranie danych z dynamicznie generowanych treści. Bez względu na wybraną metodę, kluczowe jest zrozumienie struktury strony internetowej, wybranie odpowiednich selektorów (CSS lub XPath) do wskazania interesujących nas elementów oraz umiejętność przetwarzania i zapisywania pobranych danych w odpowiednim formacie[2].

Celem artykułu jest przedstawienie różnych technik web scrapingu z wykorzystaniem biblioteki Selenium, w tym omówienie ich praktycznych zastosowań, wyzwań i ograniczeń. Obiektem badań są techniki web scrapingu, a w szczególności wykorzystanie Selenium do automatyzacji zbierania danych z dynamicznych stron internetowych. Metody badawcze obejmują analizę literatury i dokumentacji technicznej dotyczącej web scrapingu i Selenium, zastosowanie techniki scrapingu z użyciem Selenium w praktycznych przykładach oraz analizę wyzwań i aspektów prawnych.

[1] R. Mitchell, *Web Scraping with Python: collecting more data from the modern web,* April 2018: Second Edition, O'Reilly Media, s. 1-2.
[2] https://boringowl.io/blog/web-scraping-co-to-jest-i-jak-dziala (dostęp: 29.06.2024).

2. Historia Selenium

Selenium zostało opracowane w 2004 roku przez Jasona Hugginsa, który jest twórcą i inżynierem tego narzędzia. Huggins rozpoczął pracę nad Selenium, gdy zauważył, że ręczne testowanie aplikacji webowych jest czasochłonne i można by ten proces zautomatyzować. Zaczął pracować z JavaScriptem i stworzył bibliotekę do tworzenia i używania skryptów napisanych w Javie, która mogła definiować interfejs z witryną i automatycznie testować różne programy. W ten sposób powstała biblioteka Selenium Core, która zawierała wszystkie funkcje Selenium Remote Control (RC) oraz Selenium IDE. Selenium było dużym projektem komputerowym, ale miało swoje wady. Zabezpieczenia Selenium nie były wystarczająco dokładne, co sprawiało, że wykonanie niektórych zadań było trudne. Aplikacje webowe stawały się coraz bardziej zaawansowane i skomplikowane, co czyniło używanie Selenium wyjątkowo trudnym. Simon Stewart, deweloper Google, zaczął pracować nad problemami Selenium i stworzył nowy komponent, który nazwał WebDriver. W 2008 roku WebDriver stał się kluczowym elementem rozwiązania problemów związanych z ograniczeniami Selenium. WebDriver umożliwiał bezpośrednią komunikację z przeglądarką internetową, wykorzystując lokalne metody i działając w ramach systemu operacyjnego. To podejście znacząco poprawiło wydajność i funkcjonalność Selenium, czyniąc je bardziej skutecznym narzędziem do testowania aplikacji webowych[3].

3. Podstawy działania Selenium

Selenium to framework do testowania aplikacji internetowych, który automatyzuje działania przeglądarki i może być używany zarówno do prostych, jak i złożonych zadań związanych ze scrapingiem. Selenium udostępnia przeglądarkę jako interfejs lub narzędzie automatyzacji. Z jego pomocą można ładować, testować, a nawet scrapować dynamiczne treści internetowe, które korzystają z JavaScriptu, ciasteczek, skryptów i innych technologii.

Dzięki Selenium możliwe jest wykonywanie poniższych zadań bez bezpośredniego udziału człowieka:

- przeglądanie stron,
- klikanie w linki,
- zapisywanie zrzutów ekranu,
- pobieranie obrazów,

[3] S. Nyamathulla, Dr. P. Ratnababu, N. Sultana Shaik, Lakshmi. N Bhagya, *A Review on Selenium Web Driver with Python*, s. 16761.

- wypełnianie szablonów formularzy HTML i wiele innych.

Selenium jest oprogramowaniem open source i jest dostępne na różnych platformach. Do testowania można używać różnych przeglądarek internetowych, korzystając z bibliotek dostępnych dla języków programowania takich jak Java i Python. Biblioteki te służą do tworzenia skryptów, które współpracują z Selenium, aby realizować automatyzację działań przeglądarki.

Chociaż korzystanie z Selenium w testowaniu aplikacji ma wiele zalet, jeśli chodzi o takie działania jak przeszukiwanie i scrapowanie stron, ma również swoje ograniczenia. Choć jest rozbudowane i efektywne, działa powoli i zużywa dużą ilość pamięci[4].

Selenium składa się z czterech głównych narzędzi, z których każde ma swoje unikalne zastosowanie. Pierwszym z nich jest Selenium IDE. Jest to narzędzie do nagrywania i odtwarzania skryptów, które obsługuje przeglądarki Firefox i Chrome. Jest idealne dla początkujących, ponieważ umożliwia tworzenie skryptów bez konieczności pisania kodu. Kolejnym jest Selenium RC (Remote Control). Znane jako Selenium 1.0, łączyło serwer Selenium i klienta, co umożliwiało automatyzację dowolnej przeglądarki na dowolnym systemie operacyjnym. Choć nie jest już wspierane, odegrało kluczową rolę we wczesnym rozwoju Selenium. Następnym narzędziem jest SeleniumWebDriver. Znane jako Selenium 2.0, a obecnie dostępne w wersji 3.0. Używa WebDriver API, co pozwala każdej przeglądarce na automatyzację poprzez swoje API. Jest bardziej zaawansowane i elastyczne niż Selenium RC, dlatego jest obecnie standardem w automatyzacji testów przeglądarek. Ostatnim jest Selenium Grid. Umożliwia ono równoczesne wykonywanie wielu testów, wykorzystując serwer Selenium RC w trybach hub i node. Pozwala to na uruchamianie testów na wielu maszynach jednocześnie, co oszczędza czas i koszty[5].

Architektura Selenium w Pythonie składa się z wielu komponentów współpracujących ze sobą w celu automatyzacji interakcji z przeglądarkami internetowymi. Poniżej znajduje się szczegółowy opis architektury Selenium z Pythonem[6]:

8. Selenium WebDriver – kluczowy element, który odpowiada za interakcję z przeglądarkami internetowymi. WebDriver udostępnia zestaw API (Interfejsów Programowania Aplikacji), które umożliwiają komunikację między skryptami Python a

[4] A. Chapagain, *Hands-On Web Scraping with Python*, 2019 Packt Publishing, s. 450-455.
[5] P. R Sharma, *Selenium with Python – A Beginner's Guide: Get started with Selenium using Python as a Programming Language*, BPB Publications, s. 16-17.
[6] https://medium.com/@sureshkannan.gss/describe-the-python-selenium-architecture-in-detail-fd8c431aedca (dostęp: 29.06.2024).

przeglądarką. WebDriver działa jako most pomiędzy frameworkiem testowym (Python) a przeglądarką, umożliwiając automatyzację działań w przeglądarce.

9. Skrypt Python – definiuje sekwencję działań do wykonania w przeglądarce, takich jak otwieranie strony, klikanie w elementy, wypełnianie formularzy i ekstrakcja danych.
10. Sterowniki przeglądarki - każda przeglądarka (np. Chrome, Firefox, Edge) wymaga specyficznego sterownika do nawiązania połączenia z Selenium WebDriver. Sterowniki te działają jako pośrednicy, przekształcając polecenia WebDriver na specyficzne dla przeglądarki akcje. Na przykład, ChromeDriver jest używany z przeglądarką Chrome, a GeckoDriver z Firefoxem.
11. Protokół JSON Wire - to RESTful web service protocol, który umożliwia komunikację między Selenium WebDriver a sterownikiem przeglądarki. Definiuje on standardowy sposób wymiany danych i poleceń, co pozwala na kompatybilność między różnymi przeglądarkami.
12. Przeglądarka internetowa - to aplikacja, która jest automatyzowana. Selenium obsługuje różne przeglądarki, a skrypty automatyzujące definiują interakcje z przeglądarką, takie jak otwieranie URL-ów, interakcja z elementami i nawigacja między stronami.

Uproszczony schemat można przedstawić w sposób następujący. Skrypt Pythona wysyła polecenia do Selenium WebDriver. Następnie WebDriver komunikuje się ze sterownikiem przeglądarki przy użyciu protokołu JSON Wire. Sterownik przeglądarki przekształca polecenia WebDriver na akcje wykonywane przez przeglądarkę. Przeglądarka wykonuje te akcje i zwraca wyniki do WebDriver. Na koniec skrypt Pythona odbiera wyniki i kontynuuje przepływ automatyzacji.

Instalacja i konfiguracja Selenium w Pythonie obejmuje kilka prostych kroków. Najpierw trzeba zainstalować Python na komputerze, a następnie za pomocą menedżera pakietów pip zainstalować bibliotekę Selenium, wpisując w terminalu polecenie pip install selenium. Kolejnym krokiem jest pobranie odpowiedniego sterownika przeglądarki, na przykład ChromeDriver dla przeglądarki Chrome. Po pobraniu sterownika, umieszcza się go w łatwo dostępnym katalogu. Następnie w skrypcie Python konfiguruje się Selenium do pracy z pobranym sterownikiem, tworząc instancję przeglądarki i otwierając stronę internetową. Przykładowo, dla ChromeDriver tworzy się obiekt webdriver. Chrome z podaniem ścieżki do sterownika. Po uruchomieniu skryptu przeglądarka powinna się otworzyć, załadować wskazaną stronę i zamknąć. Dodatkowo można skonfigurować Selenium do pracy w trybie bez interfejsu graficznego lub innych specyficznych ustawień, w zależności od potrzeb automatyzacji.

4. Techniki scrapingu z Selenium

Podstawą skutecznego scrapowania stron internetowych za pomocą Selenium jest umiejętne określenie selektorów HTML, które pozwalają na identyfikację i interakcję z konkretnymi elementami na stronie. Dzięki różnorodnym akcjom dostępnym w kodzie Pythona z wykorzystaniem Selenium, jak klikanie, przewijanie strony czy wypełnianie formularzy, możliwe jest symulowanie działań użytkownika na stronie internetowej. Te działania nie tylko umożliwiają pobieranie danych, ale także pozwalają na testowanie interakcji użytkownika w różnych scenariuszach, co jest kluczowe dla automatyzacji testów aplikacji webowych oraz efektywnego zbierania danych z dynamicznych stron internetowych.

Współczesne strony internetowe często używają technologii AJAX do dynamicznego ładowania zawartości bez konieczności przeładowania całej strony. AJAX jest to metoda wykorzystywana do tworzenia szybkich i dynamicznych stron internetowych, które mogą aktualizować części strony bez konieczności przeładowywania całej zawartości. Technologia ta jest oparta na połączeniu JavaScriptu i XML.

W Selenium WebDriver obsługa AJAX polega na umiejętnym zarządzaniu czasem oczekiwania na załadowanie się dynamicznie generowanych treści na stronie internetowej. Selenium umożliwia zastosowanie różnych metod oczekiwania, takich jak:

1. Implicit Wait - czekanie na elementy przez cały czas trwania sesji przeglądarki.
2. Explicit Wait - wstrzymywanie wykonania testu do momentu spełnienia określonego warunku lub upływu określonego maksymalnego czasu.
3. WebDriverWait - czekanie na określony warunek przy użyciu klas ExpectedCondition w połączeniu z WebDriverWait.
4. Fluent Wait - czekanie z określonym czasem oczekiwania i interwałem sprawdzania warunku.

Obsługa AJAX w Selenium Webdriver jest kluczowa ze względu na dynamiczną naturę nowoczesnych aplikacji internetowych, które często używają technologii AJAX do szybkiego i płynnego aktualizowania treści bez konieczności przeładowywania strony. Automatyzacja testów aplikacji opartych na AJAX wymaga precyzyjnego zarządzania czasem oczekiwania, aby zapewnić stabilność i niezawodność testów.[7]

[7] https://www.guru99.com/handling-ajax-call-selenium-webdriver.html (dostęp: 30.06.2024).

5. Wyzwania i ograniczenia web scrapingu z Selenium

Selenium, jako narzędzie do automatyzacji, posiada kilka istotnych wyzwań i ograniczeń, które warto rozważyć przed jego użyciem. Poniżej wymieniono najważniejsze ograniczenia Selenium oraz sposoby ich łagodzenia:

5. Wolna wydajność - Selenium może działać wolno z powodu dużej ilości pamięci potrzebnej do uruchomienia przeglądarki, czasu renderowania, przeciążenia zasobami oraz nieprawidłowego wyboru selektorów. Można poprawić wydajność poprzez blokowanie zasobów graficznych, optymalizację opóźnień drivera, korzystanie z trybu headless oraz wybór optymalnych selektorów.
6. Problemy z kompatybilnością WebDrivera i przeglądarek - aktualizacje WebDrivera w Selenium nie są automatyczne, co może prowadzić do niezgodności wersji między WebDriverem a lokalną przeglądarką. Aby uniknąć tego problemu, należy regularnie aktualizować WebDriver wraz z przeglądarką.
7. Blokowanie podczas scrapingu - Selenium może zostać zablokowane przez mechanizmy anty-botów, takie jak HeadlessChrome w nagłówku User Agent. Można temu zaradzić zmieniając User Agent lub używając proxy. Alternatywnie, istnieją biblioteki wspierające omijanie botów, jak Undetected ChromeDriver i Selenium Stealth.
8. Koszty skalowalności - złożoność projektów automatyzacji w Selenium może wymagać zakupu maszyn wirtualnych, ustawienia lokalnych Gridów oraz monitorowania wydajności, co generuje dodatkowe koszty. Konteryzacja testów za pomocą Docker'a i korzystanie z chmury może zmniejszyć te koszty.
9. Interakcje z elementami strony przed ich załadowaniem - Selenium czasem prowadzi interakcję z elementami strony przed ich pojawieniem się, co może prowadzić do błędów braku elementu. Można temu zaradzić stosując oczekiwanie na załadowanie elementów przed ich interakcją.
10. Trudności w obsłudze błędów - nieprzewidziane błędy, takie jak zmiany dynamicznych atrybutów, ukryte elementy czy wolne ładowanie strony, są wyzwaniem w Selenium. Skuteczne logowanie, użycie explicit waits oraz odpowiednie monitorowanie obiektów mogą pomóc w obsłudze błędów.
11. Brak wbudowanego rozwiązania do rozwiązywania CAPTCHA - Selenium nie posiada wbudowanego sposobu na automatyzację rozwiązywania CAPTCHA. Można próbować opóźniać działanie skryptu, aby uniknąć pojawienia się go lub rozważyć płatne rozwiązania CAPTCHA.

12. Problemy z pop-upami - Selenium ma wbudowane rozwiązania do obsługi okienek przeglądarki, ale nie radzi sobie z dialogami systemowymi, jak okienka lokalizacji plików. Można rozważyć użycie zewnętrznych narzędzi, takich jak AutoIt czy Robot Class, do automatyzacji takich pop-upów.
13. Limity w testowaniu mobilnym - Selenium wymaga użycia zewnętrznych frameworków, jak Appium czy Selendroid, do testowania mobilnego. Dla testowania mobilnego w chmurze można rozważyć rozwiązania takie jak BrowserStack.
14. Wymagania względem konserwacji rosną szybko - kod Selenium jest bardziej rozbudowany wraz z postępem projektu, co utrudnia utrzymanie struktury kodu. Dokumentowanie testów i izolowanie selektorów elementów mogą pomóc w utrzymaniu klarowności kodu.

Pomimo tych ograniczeń, Selenium pozostaje popularnym narzędziem do automatyzacji testów i web scrapingu, oferując szerokie wsparcie dla różnych języków programowania i przeglądarek[8].

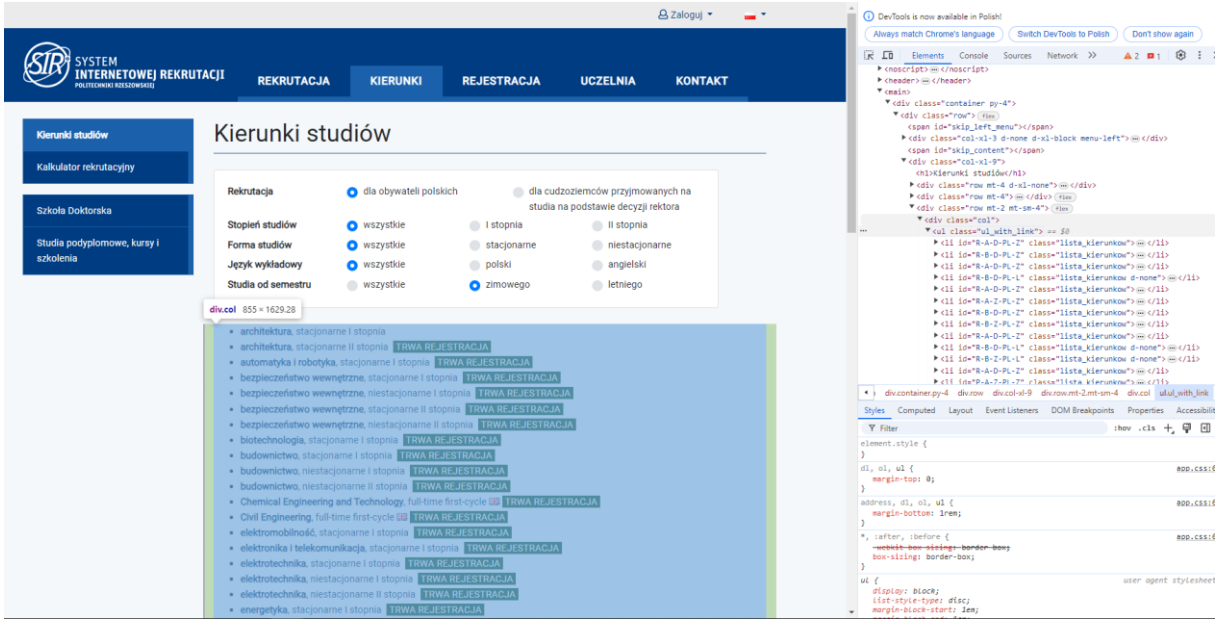
6. Praktyczne zastosowanie Selenium w web scrapingu

W tym punkcie artykułu zostanie przedstawione praktyczne zastosowanie biblioteki Selenium w Pythonie w web scrapingu. Dane zostaną pobrane ze strony *https://rekrutacja.prz.edu.pl/kierunki* widocznej na rysunku 1. Do wykonania tej czynności będzie wykorzystana omawiana w artykule biblioteka Selenium. Pobranymi danymi będą kierunki studiów – ich nazwy oraz typ.

Kod wykonujący scrapowanie danych został napisany w języku Python. Kod rozpoczyna się od importowania niezbędnych bibliotek i modułów: *webdriver* i *Service* z *Selenium*, *By*, *WebDriverWait* oraz *expected_conditions* z *selenium.webdriver*, *pandas* jako *pd* oraz *ChromeDriverManager* z *webdriver_manager*. Następnie tworzony jest obiekt Service używając *ChromeDriverManager().install()*, który pobiera i konfiguruje odpowiednią wersję *ChromeDrivera* dla zainstalowanej wersji przeglądarki *Google Chrome* na komputerze użytkownika. Kolejnym krokiem jest inicjalizacja przeglądarki Chrome poprzez *webdriver.Chrome(service=service)*, gdzie przekazujemy wcześniej utworzony obiekt *Service*. Skrypt otwiera stronę internetową pod wskazanym adresem URL *'https://rekrutacja.prz.edu.pl/kierunki'* przy użyciu *driver.get(url)*. Aby upewnić się, że strona została w pełni załadowana, używamy *WebDriverWait* z warunkiem

[8] https://www.zenrows.com/blog/selenium-limitations#selenium-overview (dostęp: 30.06.2024).

EC.presence_of_element_located((By.CLASS_NAME, 'ul_with_link')), co oznacza, że skrypt będzie czekał maksymalnie 10 sekund na pojawienie się elementu o klasie *ul_with_link*. Jeśli element nie zostanie znaleziony w tym czasie, program wyświetli komunikat o problemie i zakończy działanie przeglądarki. Po poprawnym załadowaniu strony następuje pobranie wszystkich elementów z klasą *'lista_kierunkow'* za pomocą *driver.find_elements(By.CLASS_NAME, 'lista_kierunkow')*. Następnie tworzone są dwie puste listy: *nazwa_kierunku* i *typ_kierunku*, które posłużą do przechowywania nazw i typów kierunków studiów. Iteracja przez wszystkie znalezione elementy umożliwia wyodrębnienie nazwy i typu każdego kierunku poprzez odszukanie elementu a wewnątrz elementu *lista_kierunkow* oraz podzielenie tekstu na nazwę i typ przy użyciu *split(', ')*. Odpowiednio oczyszczone nazwy i typy dodawane są do odpowiednich list. Po zakończeniu iteracji przeglądarka jest zamykana za pomocą *driver.quit()*. Następnie tworzony jest obiekt DataFrame z pobranymi danymi, gdzie *nazwa_kierunku* i *typ_kierunku* są przypisane do odpowiednich kolumn *'Nazwa kierunku'* i *'Typ kierunku'*. Ostatecznie dane są zapisywane do pliku Excel o nazwie *kierunki_studiow.xlsx* za pomocą *df.to_excel('kierunki_studiow.xlsx', index=False)*.



Rysunek 49. Strona internetowa, z której zbierane są dane
Źródło: opracowanie własne.

Cały opisany kod przedstawia rysunek 2, natomiast wyniki programu - rysunek 3.

scraper.py > ...

```python
# Import potrzebnych bibliotek
from selenium import webdriver
from selenium.webdriver.chrome.service import Service
from selenium.webdriver.common.by import By
from selenium.webdriver.support.ui import WebDriverWait
from selenium.webdriver.support import expected_conditions as EC
import pandas as pd
from webdriver_manager.chrome import ChromeDriverManager

# Inicjalizacja przeglądarki z użyciem ChromeDriverManager i Service
service = Service(ChromeDriverManager().install())  # Inicjalizacja usługi ChromeDriverManager
driver = webdriver.Chrome(service=service)  # Inicjalizacja przeglądarki Chrome z użyciem utworzonej usługi
# Otwieranie strony internetowej
url = 'https://rekrutacja.prz.edu.pl/kierunki'
driver.get(url)
# Czekanie, aż elementy będą dostępne
try:
    WebDriverWait(driver, 10).until(EC.presence_of_element_located((By.CLASS_NAME, 'ul_with_link')))
    print("Strona załadowana pomyślnie")
except Exception as e:
    print(f"Problem z załadowaniem strony: {e}")
    # Zamknięcie przeglądarki w przypadku problemu z ładowaniem strony
    driver.quit()
    # Wyjście z programu
    exit()

# Pobranie wszystkich elementów z klasą 'lista_kierunkow'
elements = driver.find_elements(By.CLASS_NAME, 'lista_kierunkow')
# Listy do przechowywania danych
nazwa_kierunku = []
typ_kierunku = []

# Iteracja po wszystkich elementach
for element in elements:
    try:
        a_tag = element.find_element(By.TAG_NAME, 'a')
        # Pobieranie tekstu z elementu 'a' i podzielenie na nazwę kierunku i typ
        kierunek_text = a_tag.text
        if ', ' in kierunek_text:
            nazwa, typ = kierunek_text.split(', ')
            # Dodanie oczyszczonej nazwy kierunku do listy
            nazwa_kierunku.append(nazwa.strip())
            # Dodanie oczyszczonego typu kierunku do listy
            typ_kierunku.append(typ.strip())
    except Exception as e:
        # Wyświetlenie komunikatu o problemie z elementem
        print(f"Problem z elementem: {e}")

# Zakończenie działania przeglądarki
driver.quit()

# Tworzenie DataFrame z pobranymi danymi
df = pd.DataFrame({
    'Nazwa kierunku': nazwa_kierunku,
    'Typ kierunku': typ_kierunku
})

# Zapis do pliku Excel
df.to_excel('kierunki_studiow.xlsx', index=False)  # Zapis danych do pliku Excel bez indeksów
```

Rysunek 50. Kod źródłowy utworzonego programu do scrapowania danych
Źródło: opracowanie własne.

	A	B
1	**Nazwa kierunku**	**Typ kierunku**
2	architektura	stacjonarne I stopnia
3	architektura	stacjonarne II stopnia
4	automatyka i robotyka	stacjonarne I stopnia
5	bezpieczeństwo wewnętrzne	stacjonarne I stopnia
6	bezpieczeństwo wewnętrzne	niestacjonarne I stopnia
7	bezpieczeństwo wewnętrzne	stacjonarne II stopnia
8	bezpieczeństwo wewnętrzne	niestacjonarne II stopnia
9	biotechnologia	stacjonarne I stopnia
10	budownictwo	stacjonarne I stopnia
11	budownictwo	niestacjonarne I stopnia
12	budownictwo	niestacjonarne II stopnia
13	Chemical Engineering and Technology	full-time first-cycle
14	Civil Engineering	full-time first-cycle
15	elektromobilność	stacjonarne I stopnia
16	elektronika i telekomunikacja	stacjonarne I stopnia
17	elektrotechnika	stacjonarne I stopnia
18	elektrotechnika	niestacjonarne I stopnia
19	elektrotechnika	niestacjonarne II stopnia
20	energetyka	stacjonarne I stopnia
21	finanse i rachunkowość	stacjonarne I stopnia
22	finanse i rachunkowość	niestacjonarne I stopnia

Rysunek 51. Wyniki utworzonego programu do zbierania danych
Źródło: opracowanie własne.

Można zauważyć, że kod wykorzystujący Selenium do web scrapingu nie tylko umożliwia efektywne pozyskiwanie danych z internetu, ale także zapewnia stabilność, elastyczność i łatwość w zarządzaniu automatyzacją procesów związanych z analizą treści online.

7. Podsumowanie

Artykuł przedstawia techniki web scrapingu z wykorzystaniem biblioteki Selenium, koncentrując się na jej roli, metodach i praktycznych zastosowaniach. Web scraping, czyli automatyczne pobieranie danych z witryn internetowych, jest istotny dla analiz danych i aplikacji komercyjnych. Selenium, jako narzędzie do automatyzacji przeglądania stron, umożliwia scrapowanie dynamicznych treści generowanych przez JavaScript oraz obsługę AJAX. Artykuł omawia historię Selenium, jego architekturę oraz działanie.

Przedstawiono techniki scrapingu dynamicznych treści i zarządzania opóźnieniami ładowania stron. Szczegółowo opisano również praktyczne zastosowanie Selenium w web scrapingu, w tym przykład pobierania danych z konkretnej strony internetowej. Analizowane są także wyzwania i ograniczenia idące za używaniem biblioteki Selenium do web scrapingu. Artykuł łączy teoretyczne wyjaśnienia z praktycznymi przykładami, co czyni go przydatnym źródłem wiedzy dla osób zainteresowanych web scrapingiem z wykorzystaniem Selenium.

Wnioski z artykułu podkreślają, że Selenium jest potężnym narzędziem do web scrapingu, szczególnie przydatnym do scrapowania dynamicznych stron internetowych. Pomimo pewnych wyzwań, takich jak wolna wydajność czy problemy z kompatybilnością, Selenium oferuje dużą elastyczność i możliwość automatyzacji skomplikowanych interakcji z przeglądarką. Kluczowe jest między innymi zrozumienie struktury strony internetowej oraz odpowiednie zarządzanie czasem oczekiwania na załadowanie elementów. Pomimo ograniczeń, Selenium pozostaje popularnym narzędziem, wspierającym rozwój zaawansowanych technik web scrapingu.

Literatura

1. Chapagain A., *Hands-On Web Scraping with Python*, 2019 Packt Publishing.
2. Mitchell R., *Web Scraping with Python: collecting more data from the modern web*, April 2018: Second Edition, O'Reilly Media.
3. Nyamathulla S., Ratnababu Dr. P., Sultana Shaik N., Bhagya L. N., *A Review on Selenium Web Driver with Python*.
4. Sharma P. R, *Selenium with Python – A Beginner's Guide: Get started with Selenium using Python as a Programming Language*, BPB Publications.

Źródła internetowe

1. https://boringowl.io/blog/web-scraping-co-to-jest-i-jak-dziala (dostęp: 29.06.2024).
2. https://medium.com/@sureshkannan.gss/describe-the-python-selenium-architecture-in-detail-fd8c431aedca (dostęp: 29.06.2024).
3. https://www.guru99.com/handling-ajax-call-selenium-webdriver.html (dostęp: 30.06.2024).
4. https://www.zenrows.com/blog/selenium-limitations#selenium-overview (dostęp: 30.06.2024).

Katarzyna Maternia, Magdalena Matuła, Aleksandra Sawicka, Aleksandra Rokita, Wiktor Kuczek
SKNI „Kod”

dr inż. Bartosz Trybus
Opiekun Koła Naukowego

Rozwój metodologii DevOps i jej wpływ na szybkie wdrażanie oprogramowania i automatyzację procesów

Streszczenie

Artykuł omawia rozwój metodologii DevOps i jej znaczący wpływ na szybkie wdrażanie oprogramowania oraz automatyzację procesów. DevOps łączy zespoły programistyczne i operacyjne, co prowadzi do skrócenia czasu wdrażania, zwiększenia częstotliwości wydań oraz poprawy jakości oprogramowania. Kultura DevOps, charakteryzująca się współpracą, przejrzystością i odpowiedzialnością, promuje kompleksowe podejście do zarządzania całym cyklem życia oprogramowania. W artykule omówiono również narzędzia wspierające DevOps oraz wyzwania związane z jego wdrożeniem, a także przedstawiono krok po kroku, jak implementować DevOps w organizacji.

Słowa kluczowe: DevOps, automatyzacja, ciągłe dostarczanie, szybkie wdrażanie oprogramowania, kultura DevOps, współpraca zespołów, narzędzia DevOps.

1. Wprowadzenie

Artykuł przedstawia kulturę oraz praktyki związane z ciągłym dostarczaniem oprogramowania przed DevOps. W dzisiejszym dynamicznym świecie cyfrowym, tempo rozwoju technologicznego stawia przed organizacjami coraz większe wyzwania w zakresie dostarczania oprogramowania. Wraz z coraz większą konkurencją na rynku i coraz wyższymi oczekiwaniami klientów, organizacje są zmuszone znaleźć sposoby, aby szybko i efektywnie wdrażać innowacyjne rozwiązania, zachowując jednocześnie wysoką jakość produktów.

2. Zalety DevOps

Łączenie zespołów programistycznych i operacyjnych prowadzi do znacznego skrócenia czasu wdrażania i zwiększenia jego częstotliwości oraz poprawy jakości oprogramowania. Wdrażanie DevOps niesie ze sobą duże wartości, dzięki min respondenci twierdzą, że DevOps ma pozytywny wpływ na ich organizację. Zespoły, które praktykują DevOps, zauważa się że szybciej pracują, sprawniej reagują na incydenty oraz usprawniają współpracę i komunikację między zespołami.

Budowanie kultury opartej na współdzieleniu odpowiedzialności, przejrzystości oraz szybkim udzielaniu informacji zwrotnych jest kluczowe dla sprawnie funkcjonującego zespołu DevOps. Zespoły pracujące w izolacji często nie są w stanie myśleć systemowo, co jest niezbędne w kulturze DevOps. Myślenie systemowe polega na zrozumieniu, że działania pojedynczych pracowników wpływają nie tylko na ich własny zespół, ale również na wszystkie inne zespoły zaangażowane w proces wydawania oprogramowania. Brak widoczności i wspólnych celów prowadzi do braku planowania zależności, różnic w priorytetach, szukania winnych oraz przekonania, że "to nie nasz problem", co ostatecznie spowalnia pracę i obniża jej jakość. DevOps to zmiana podejścia do procesu programowania, która zachęca do całościowego myślenia i przełamuje bariery między działami programistycznymi i operacyjnymi.

Szybkość działania jest kluczowa. Zespoły pracujące zgodnie z zasadami DevOps udostępniają produkty częściej, przy jednoczesnym utrzymaniu wysokiej jakości i stabilności. Brak zautomatyzowanych cykli testów i recenzji opóźnia wdrażanie wydań do produkcji, a długi czas reakcji na incydenty spowalnia pracę i obniża morale zespołu. Różnorodne narzędzia i procesy zwiększają koszty operacyjne, wymuszają częste zmiany kontekstu i zmniejszają efektywność. Jednakże, dzięki narzędziom wspierającym automatyzację oraz nowe procesy, zespoły mogą zwiększyć produktywność i częstotliwość wydań, jednocześnie minimalizując problemy.

Zespół, który ma najkrótszą pętlę informacji zwrotnej, osiąga największe sukcesy. Pełna przejrzystość i sprawna komunikacja pozwalają zespołom DevOps zminimalizować przestoje i szybciej usuwać problemy. Jeśli krytyczne problemy nie będą rozwiązywane szybko, satysfakcja klientów spadnie. Brak otwartej komunikacji powoduje, że kluczowe zgłoszenia mogą zostać przeoczone, co prowadzi do wzrostu napięcia i frustracji wśród członków zespołu. Dzięki otwartej komunikacji zespoły programistyczne i operacyjne mogą wspólnie rozwiązywać problemy, szybciej reagować na incydenty.

3. Kultura DevOps

Kultura DevOps wprowadza głębszą współpracę i wspólną odpowiedzialność między programistami a pracownikami operacyjnymi, koncentrując się na dostarczaniu lepszych produktów dla klientów. Multidyscyplinarne zespoły biorą odpowiedzialność za cały cykl życia produktu, od koncepcji po wdrożenie i utrzymanie. Dzięki temu programiści mają lepsze zrozumienie potrzeb użytkowników, a zespoły operacyjne mogą skuteczniej integrować wymagania konserwacyjne i klienta, co prowadzi do wyższej jakości produktu.

Kluczowym elementem kultury DevOps jest zwiększona przejrzystość, komunikacja i współpraca między zespołami, które wcześniej pracowały w izolacji. Wymaga to istotnych zmian kulturowych, takich jak ciągłe uczenie się i doskonalenie, autonomia zespołów, szybkie informacje zwrotne oraz wysoki poziom empatii i zaufania. DevOps promuje podejście „odpowiadasz za to, co tworzysz", co oznacza, że programiści nie tylko tworzą kod, ale również biorą udział w jego testowaniu, wdrażaniu i obsłudze.

Autonomiczne zespoły są fundamentalnym aspektem DevOps. Mają one możliwość podejmowania decyzji i wprowadzania zmian bez długotrwałego procesu zatwierdzania, co buduje zaufanie i tworzy środowisko sprzyjające innowacjom. Procesy i narzędzia, które wspierają szybkie podejmowanie decyzji, są kluczowe dla skuteczności zespołów DevOps.

Typowy przepływ pracy w DevOps obejmuje zautomatyzowane procesy, które ułatwiają wdrażanie zmian. Na przykład programista wprowadza zmiany w kodzie, które są automatycznie kompilowane i wdrażane w środowisku testowym, a narzędzie do śledzenia zgłoszeń jest aktualizowane bez potrzeby ręcznej interwencji. Zespoły DevOps eliminują bariery, takie jak długotrwałe procesy zatwierdzania drobnych zmian w infrastrukturze, co pozwala na szybsze wdrażanie i reagowanie na potrzeby użytkowników.

Szybka informacja zwrotna jest również istotna w kulturze DevOps. W tradycyjnych modelach zespoły programistyczne i operacyjne są odizolowane, co opóźnia przepływ informacji zwrotnych na temat wydajności i stabilności oprogramowania. W DevOps programiści otrzymują natychmiastową informację zwrotną, co pozwala na szybsze iteracje i ulepszanie kodu. Narzędzia do ciągłej integracji automatyzują procesy kompilacji i testowania, dostarczając programistom natychmiastowych informacji na temat jakości ich kodu.

Automatyzacja jest kluczowym elementem kultury DevOps, umożliwiając płynną współpracę i optymalizację zasobów. Integracja procesów między zespołami programistycznymi a IT przyspiesza kompilację, testowanie i wydawanie oprogramowania, co sprawia, że procesy są bardziej niezawodne i efektywne.

3.1 Zalety kultury DevOps

Jednym z najważniejszych atutów kultury DevOps są usprawnione, częste i wysokiej jakości wydania oprogramowania. To nie tylko zwiększa wydajność firmy, ale również zadowolenie pracowników.

Kultura DevOps promuje wysoki poziom zaufania i współpracy. Skutkuje to podejmowaniem lepszych decyzji i wyższym poziomem satysfakcji z pracy. Implementacja kultury DevOps jest kluczowa dla tworzenia wysoko wydajnych organizacji inżynierskich, które nie tylko osiągają lepsze wyniki biznesowe, ale również dbają o zadowolenie

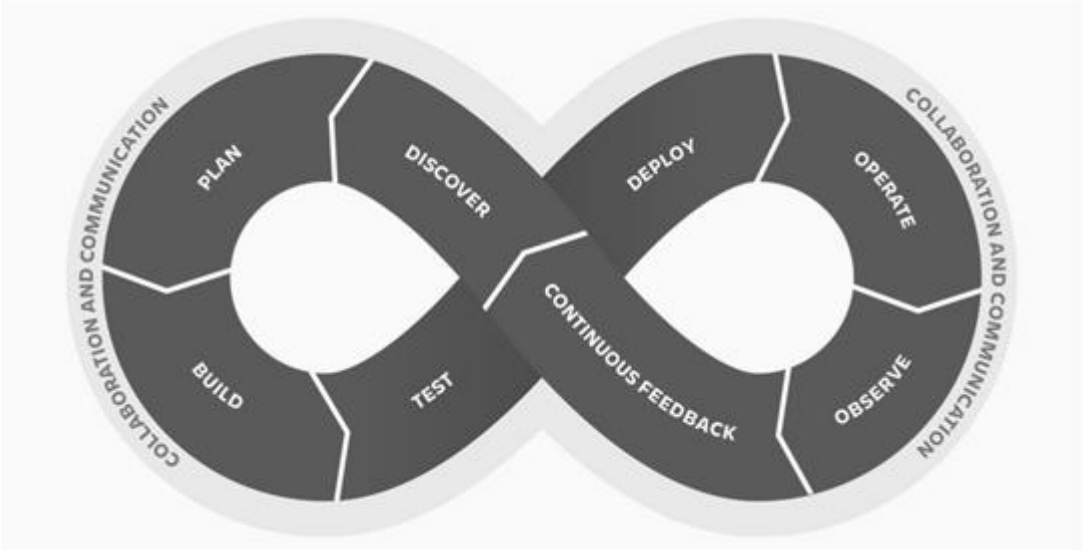
pracowników. Inżynierowie czerpią satysfakcję z częstego i bezproblemowego wdrażania stabilnego, wydajnego oprogramowania, co przekłada się na zadowolenie użytkowników i sukces biznesowy.

3.2 Wyzwania związane z wdrożeniem kultury DevOps

Wprowadzenie pełnej kultury DevOps wymaga znaczących zmian w sposobie działania zespołów i organizacji. Kluczowe jest wsparcie kierownictwa najwyższego szczebla, choć często inicjatywa zaczyna się oddolnie, od małych zespołów, które pokazują skuteczność DevOps.

Wysoki poziom autonomii i zaufania, niezbędny w kulturze DevOps, może być trudny do utrzymania w przypadku konfliktów między zespołami. Im bardziej złożona struktura organizacyjna, tym trudniej wdrożyć współpracę. Ważne jest, aby korzyści płynące ze zmian były wyraźnie komunikowane i zrozumiane przez wszystkich, aby zwiększyć akceptację i gotowość do wdrożenia DevOps.

4. Jak stosować DevOps w 8 krokach



Rysunek 52 8 kroków do DevOps; Źródło: https://www.atlassian.com/pl/devops/what-is-devops/how-to-start-devops

Krok 1 — Wybór komponentu

Pierwszym krokiem jest rozpoczęcie od małej skali, wybierając komponent obecny w fazie produkcji. Idealnie byłoby, gdyby miał on prostą bazę kodu z minimalnymi zależnościami i infrastrukturą. Taki komponent stanie się poligonem doświadczalnym, na którym zespół będzie mógł ćwiczyć wdrażanie DevOps.

Krok 2 — Zastosowanie metodologii Agile, takiej jak Scrum

DevOps często współpracuje z metodologią Agile, taką jak Scrum. Nie trzeba jednak wdrażać wszystkich praktyk związanych z Scrumem. Wystarczy skupić się na trzech kluczowych elementach: backlogu, sprintach i planowaniu sprintu. W zespole DevOps można nadawać i dodawać piorytety pracy w backlogu Scrum oraz wprowadzać podzbiór tej pracy do sprintu, czyli określonego czasu jaki jest przeznaczony na wykonanie danego zadania. Takie planowanie sprintu polega na decydowaniu jakie zadania przechodzą z backlogu zaległości na następny sprint.

Krok 3 — Wykorzystanie kontroli źródła opartego na GIT

Kontrola wersji jest kluczową praktyką DevOps, umożliwiającą lepszą współpracę i skrócenie cykli wydawania. Narzędzia takie jak Bitbucket ułatwiają udostępnianie, scalanie i tworzenie kopii zapasowych oprogramowania.

Krok 4 — Integracja kontroli źródła ze śledzeniem pracy

Zintegruj narzędzie do kontroli źródła z narzędziem do śledzenia pracy, aby zobaczyć wszystko, co jest związane z danym projektem w jednym miejscu. Można to osiągnąć poprzez dodanie identyfikatora zgłoszenia do komunikatów i nazw gałęzi pracy związanych ze zgłoszeniem.

Krok 5 — Napisanie testów

Pipeline'y CI/CD wymagają testów, aby sprawdzić, czy kod działa poprawnie w różnych środowiskach. Zaczynając od testów jednostkowych, można stopniowo zwiększać pokrycie kodu testami jednostkowymi, integracyjnymi i systemowymi.

Krok 6 — Tworzenie procesu CI/CD w celu wdrożenia komponentu

Stwórz pipeline CI/CD do wdrażania infrastruktury i kodu. Upewnij się, że proces ten jest powtarzalny i odporny na błędy, umożliwiając szybkie wycofanie w razie potrzeby.

Krok 7 — Dodawanie monitorowania, alarmów i oprzyrządowania

Monitoruj zachowanie aplikacji w każdym środowisku, reagując na wykryte problemy i wprowadzając poprawki. Optymalizuj wydajność systemu, dostosowując te działania do najbardziej krytycznych komponentów.

Krok 8 — Użycie flag funkcji do wdrażania "testów kanarka"

Zabezpiecz nowe funkcje za pomocą flag funkcji, umożliwiając kontrolowany dostęp do nich przez wybrane grupy użytkowników. Monitoruj zachowanie tych funkcji i reaguj na sygnały problemów przed przeniesieniem ich do kolejnych środowisk.

Początkowo wdrożenie DevOps dla przeniesienia komponentu do produkcji może wydawać się wymagające, ale w dłuższej perspektywie przynosi znaczne korzyści. Wdrożenie drugiego komponentu powinno być prostsze, ponieważ narzędzia są już dostępne, technologie są znane,

a zespół jest przeszkolony w pracy zgodnie z DevOps. Proces użyty przy pierwszym komponencie można wykorzystać i dostosować do potrzeb drugiego komponentu.

5. DevOps a DevSecOps

DevSecOps to rozszerzenie filozofii DevOps, dlatego ważne jest zrozumienie ich wspólnych cech. Obie te filozofie, DevOps i DevSecOps, odnoszą się do podejścia do rozwoju oprogramowania, a nie do konkretnych narzędzi. Podobnie jak instalacja systemu śledzenia problemów nie oznacza automatycznie „robienia DevOps", tak samo instalacja narzędzi bezpieczeństwa nie oznacza „robienia DevSecOps".

DevOps i DevSecOps skupiają się na współpracy, automatyzacji i aktywnym monitorowaniu aplikacji. Kluczową rolę odgrywa możliwość przechwytywania danych aplikacji w czasie rzeczywistym, co pozwala na ciągłą analizę i poszukiwanie sposobów na poprawę wydajności oraz wprowadzanie ulepszeń.

Obie te filozofie kładą nacisk na współpracę i eliminację silosów organizacyjnych. DevOps ma na celu zniwelowanie barier między zespołami deweloperskimi a operacyjnymi, co prowadzi do szybszego i bardziej niezawodnego wydawania oprogramowania. DevSecOps idzie krok dalej i stara się zapewnić udział operacji bezpieczeństwa, co pozwala na wydawanie oprogramowania szybciej, z lepszą jakością i większym bezpieczeństwem.

„Robienie" DevSecOps prawidłowo oznacza, że aplikacje są chronione przed zagrożeniami jeszcze przed wdrożeniem. Praktyka ta, znana jako „przesunięcie w lewo", polega na integracji zabezpieczeń na wczesnym etapie projektu, zanim kod zostanie napisany, zamiast dodawania zabezpieczeń na późniejszych etapach.

W środowisku DevSecOps programiści tworzą kod z uwzględnieniem bezpieczeństwa, co nie jest priorytetem samego DevOps. Wprowadzając praktyki takie jak analiza kodu, ocena zagrożeń i testowanie podatności na wczesnym etapie, DevSecOps zapewnia, że kod jest bezpieczny od samego początku. Oprócz zwiększenia bezpieczeństwa, DevSecOps poprawia produktywność. Wykrywanie i naprawianie problemów bezpieczeństwa na wczesnym etapie jest znacznie mniej czasochłonne i kosztowne niż refaktoryzacja kodu na późniejszych etapach cyklu życia oprogramowania.

6. Bezpieczeństwo DevOps

Mimo licznych korzyści płynących z DevSecOps, organizacje mogą napotykać trudności z jego właściwym wdrożeniem. Jednym z głównych problemów jest nadmierny nacisk na narzędzia kosztem procesów. DevOps i DevSecOps to filozofie, a nie konkretne

oprogramowanie. Kolejną barierą jest opór kulturowy ze strony programistów, którzy mogą być przyzwyczajeni do tradycyjnych metod pracy i obawiają się, że konieczność dbania o bezpieczeństwo spowolni produkcję. Zespoły bezpieczeństwa również mogą mieć trudności z przystosowaniem się do szybkiego tempa pracy DevOps, ponieważ często polegają na ręcznych procesach, podczas gdy DevOps automatyzuje jak najwięcej działań.

Innym wyzwaniem jest niewłaściwe zarządzanie wpisami tajnymi w złożonych i połączonych środowiskach DevOps, gdzie setki grup bezpieczeństwa i tysiące instancji serwerów wykorzystują różne poświadczenia i klucze. Jeden błąd w konfiguracji może prowadzić do ujawnienia tych wpisów i poważnych cyberataków. Problemem jest także niewystarczające zarządzanie dostępem uprzywilejowanym. Aby przyspieszyć produkcję, zespoły DevOps często przyznają praktycznie nieograniczony dostęp do kont uprzywilejowanych, co stanowi duże zagrożenie bezpieczeństwa i komplikacje przy audytach zgodności. Narzędzia DevOps mogą mieć poziomy dostępu znacznie przewyższające ich potrzeby, co dodatkowo zwiększa ryzyko.

7. Narzędzia DevOps

DevOps to podejście, które integruje zespoły programistyczne i operacyjne, przyspieszając proces dostarczania oprogramowania i poprawiając jego jakość. Otwarty łańcuch narzędzi DevOps pozwala na dostosowanie rozwiązań do specyficznych potrzeb organizacji. W każdej fazie cyklu życia DevOps można wykorzystać odpowiednie narzędzia, które usprawniają współpracę, automatyzację i monitorowanie.

Faza odkrywania polega na badaniu i definiowaniu zakresu projektu. Narzędzia takie jak Jira Product Discovery pozwalają organizować informacje i ustalać priorytety działań, natomiast Mural i Miro umożliwiają prowadzenie asynchronicznej burzy mózgów, gdzie każdy członek zespołu może udostępniać i komentować pomysły, strategie oraz dokumentację. Planowanie wspomagane jest przez Jira Software, które pomaga w zarządzaniu projektami, śledzeniu postępów i koordynowaniu działań zespołowych.

Faza kompilowania opiera się na narzędziach takich jak Kubernetes i Docker, które umożliwiają aprowizację środowisk programistycznych identycznych z produkcyjnymi. To eliminuje problemy związane z różnicami w środowiskach, a Ansible, Chef, Puppet i Terraform pozwalają na szybką i spójną aprowizację infrastruktury jako kodu.

W fazie kontroli źródła i kodowania opartego na współpracy kluczowe są narzędzia takie jak Bitbucket, GitHub i GitLab. Ułatwiają one przechowywanie kodu i współpracę poprzez pull requesty, co znacząco poprawia jakość oprogramowania. Ciągłe dostarczanie realizowane jest

przy użyciu narzędzi takich jak Jenkins, AWS, CircleCI i SonarSource. Automatyzują one testy i wdrażanie kodu, przyspieszając procesy programistyczne i umożliwiając szybsze udostępnianie użytkownikom nowych funkcji. Testowanie w DevOps obejmuje automatyczne testowanie, które pozwala na przeprowadzanie testów na wczesnym etapie i z dużą częstotliwością. To z kolei poprawia jakość oprogramowania i zmniejsza ryzyko.

W fazie wdrażania narzędzia takie jak Bitbucket i Zephyr umożliwiają automatyzację procesów wdrożeniowych, standaryzację środowisk oraz eliminację różnic między nimi. Obsługa jest wspomagana przez narzędzia takie jak Jira Service Management, Opsgenie i Statuspage, które pozwalają na śledzenie incydentów, zmian i problemów. Umożliwia to efektywną współpracę między zespołami programistycznymi i operacyjnymi.

Monitorowanie aplikacji i serwerów zapewniają narzędzia takie jak AppDynamics, Datadog, Slack, Splunk, New Relic, Opsgenie, Pingdom, Nagios, Dynatrace, Hosted Graphite i Sumo Logic. Automatyzują one zbieranie danych, dostarczając bieżący wgląd w kondycję systemów. Ciągłe informacje zwrotne od użytkowników są gromadzone dzięki narzędziom takim jak GetFeedback, Slack, Jira Service Management i Pendo. Umożliwiają one integrację z platformami do przeprowadzania ankiet oraz mediami społecznościowymi, co zapewnia dostęp do opinii użytkowników w czasie rzeczywistym.

Dzięki otwartemu łańcuchowi narzędzi DevOps, organizacje mogą przyspieszyć prace, skracając czas wprowadzania produktów na rynek, jednocześnie dostosowując narzędzia do swoich unikalnych potrzeb.

8. Podsumowanie

Artykuł podkreśla, że DevOps jest nie tylko metodologią, ale również fundamentalną zmianą kulturową w sposobie pracy zespołów IT. Poprzez promowanie współpracy, przejrzystości i wspólnej odpowiedzialności, DevOps pozwala organizacjom na szybsze i bardziej efektywne dostarczanie wysokiej jakości oprogramowania. Wdrożenie DevOps przynosi korzyści w postaci zwiększonej produktywności, lepszej jakości produktów i wyższej satysfakcji klientów. Niemniej jednak, aby w pełni wykorzystać potencjał DevOps, organizacje muszą przejść przez szereg kroków wdrożeniowych i stawić czoła różnym wyzwaniom, w tym kulturowym i technicznym.

Źródła internetowe:

1. https://learn.microsoft.com/pl-pl/training/modules/introduction-to-devops/1-introduction (dostęp: 29.05.2024)

2. https://www.atlassian.com/pl/devops/what-is-devops/benefits-of-devops (dostęp: 29.05.2024)
3. https://www.atlassian.com/pl/devops/what-is-devops/devops-culture(dostęp: 29.05.2024)
4. https://www.atlassian.com/pl/devops/what-is-devops/how-to-start-devops(dostęp: 29.05.2024)
5. https://www.keepersecurity.com/pl_PL/resources/glossary/what-is-devops-security/(dostęp: 29.05.2024)
6. https://www.atlassian.com/pl/devops/devops-tools(dostęp: 29.05.2024)

Katarzyna Maternia, Magdalena Matuła, Aleksandra Sawicka, Aleksandra Rokita, Łukasz Książek
SKNI „Kod”

dr inż. Bartosz Trybus
Opiekun Koła Naukowego

Rola sztucznej inteligencji w przemyśle motoryzacyjnym: Od autonomicznych pojazdów do inteligentnych systemów nawigacyjnych

Streszczenie

Stale rozwijająca się sztuczna inteligencja stanowi coraz większą cześć naszego codziennego życia, nawet jeśli nie zawsze zdajemy sobie sprawę z tego, że mamy z nią do czynienia. Sztuczną inteligencję wykorzystuje się również w motoryzacji i nie tylko w tych najnowszych modelach. Rozwiązania oparte na sztucznej inteligencji między innymi zwiększają bezpieczeństwo

na drodze. Tak jak każda inna dziedzina rozwijającego się przemysłu, również motoryzacja ciągle ewoluuje. Pojawiają się nowe technologie, które są wykorzystywanie przez inżynierów, dzięki sztucznej inteligencji.

Artykuł omawia rosnącą rolę sztucznej inteligencji w przemyśle motoryzacyjnym, począwszy od autonomicznych pojazdów aż po inteligentne systemy nawigacyjne. Opisuje różne aspekty wykorzystania AI w samochodach, włączając w to rozwój pojazdów autonomicznych, funkcje ADAS i DMS oraz zaawansowane systemy nawigacyjne.

Słowa kluczowe: sztuczna inteligencja, przemysł motoryzacyjny, autonomiczne pojazdy.

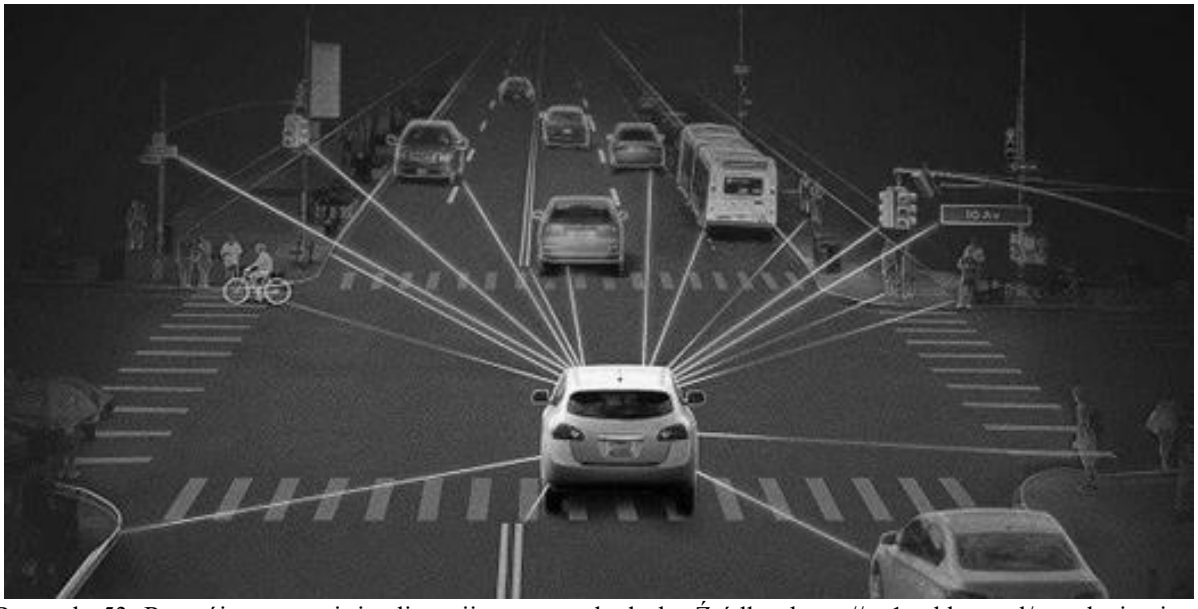
1. Wprowadzenie

Algorytmy uczą się rozpoznawania obrazu i głosu, dlatego samochody są w stanie samodzielnie wykrywać zagrożenia i podejmować decyzje jak dobieranie odpowiednich parametrów jazdy, ostrzegając kierowcę a nawet ingerować w jazdę. Miliony ludzi na całym świecie na co dzień, używają narzędzi AI (ang. – artificial intelligence), takich jak generatory obrazów, programy analityczne, inteligentna nawigacja czy chatboty. W ciągu ostatnich lat sztuczna inteligencja ma dużo większy rozwój w dziedzinach technologii. Bywają tacy co są przerażeni przyszłością AI, jednak dla wielu ludzi oraz rozwijających się stanowisk pracy, stanowi ona inny wymiar cywilizacyjny. Ta technologia ma ogromny wpływ na rozwój i produkcje całej branży motoryzacyjnej.

W chwili obecnej wartość rynku przeznaczona na sztuczną inteligencje wynosi 1/5 na technologie związane z motoryzacją. Sfera samochodów przechodzi zmianę, która jest

integracją sztucznej inteligencji. Pojazdy takie wyposażone w sztuczną inteligencję nie są jak dawniej ograniczone do sfery science fiction, tylko z każdym dniem stają się coraz bardziej naszą przyszłością, w której samochody komunikują się, przewidują i podejmują decyzje w czasie rzeczywistym. Ta ewolucja pojazdów autonomicznych oznacza na nowo zdefiniowaną mobilność, oferując większy, niespotykany dotąd poziom bezpieczeństwa, wydajności a nawet doświadczenia kierowców. Z pewnością jest to rewolucja oparta na sztucznej inteligencji, która czeka na horyzoncie w świecie motoryzacji.

Rysunek poniżej pozwoli nam lepiej zrozumieć, wyobrazić w jak sposób sztuczna inteligencja postrzega otoczenie pojazdu, takie jak obiekty, które mogą stanowić także przeszkody.



Rysunek 53 Rozwój sztucznej inteligencji w samochodach, Źródło: https://no1-reklama.pl/rewolucja-ai-w-samochodach/

2. Samochody autonomiczne

Samochody autonomiczne to tak zwane samochody bezzałogowe są to pojazdy, które poruszają się bez udziału człowieka. To komputer steruje maszyną dlatego takie pojazdy potrafią być nawet bez osoby kierującej. Pojazdy takie potrafią same omijać przeszkody czy przemierzać wyznaczoną trasą.

Obecnie, zazwyczaj auta autonomiczne opierają się na zmodyfikowanych modelach jakie już są dostępne na rynku. Takie samochody są produkowane przez duże koncerty samochodowe jak np. Mercedes-Benz, Nissan, Audi czy Volvo jak i przez przedsiębiorstwa, które specjalizują się w nowoczesnych technologiach przede wszystkim informatycznych, takich jak Nvidia czy Google.

Takie samochody mają odpowiedni stopień autonomiczny. Wyróżniamy samochody, które poruszają się samodzielnie ale kierowca musi czuwać i cały czas mieć kontrolę nad pojazdem. Są także takie pojazdy, w której kierowca nie musi ciągle śledzić jaka jest sytuacja na drodze oraz samochody, które są całkowicie zautomatyzowane i nie potrzebują kontroli człowieka.

Pojazdy, które są autonomiczne pozwalają na jazdę bez kierowcy, co sprawia, że samochód jeździ w dobrą stronę, nie zderza się z innymi obiektami, auta poruszają się we właściwym kierunku i omijają przeszkody jakie napotkają, dzięki takim technologią jak:

- radar - to urządzenie, które wyszukuje obiekty za pomocą fal radiowych,
- urządzenie LIDAR - działa podobnie do radaru, lecz zamiast fal wykorzystuje światło lasera,
- urządzenie GPS - to system nawigacji satelitarnej, który dostarcza informacje o położeniu na podstawie wysyłanych na orbitę sygnałów radiowych,
- widzenie komputerowe takie jak rozpoznawanie obrazu - polega na przetwarzaniu obrazu przez maszynę w opis cyfrowy w celu dalszego wykorzystania.

Rozwój technologii pojazdów autonomicznych przyspieszył dzięki wyścigom DARPA Grand Challenge, których organizatorem zawodów jest agencja rządu USA. Ma ona duży nacisk na rozwój technologii wojskowej. Również ta agencja sponsoruje projekty badawcze, które mogą mieć powiązanie z obroną Stanów Zjednoczonych. Samochody autonomiczne wymagają wprowadzenia rozwiązań AI w rzeczywistym świecie. Pojawiły się już pojazdy, które autonomicznie przewożą pasażerów po drogach.

3. Sztuczna inteligencja zamiast kierowcy

Idąc tropem sztucznej inteligencji w samochodach autonomicznych, przez wielu są uważane za przyszłość motoryzacji. Takie samochody poruszające się bez udziału kierowcy są już od dawna testowane przez duże koncerny motoryzacyjne i można się spodziewać, że kiedyś staną się naszą codziennością. Oparte są one na sztucznej inteligencji, która w dużej mierze wykorzystuje algorytmy uczenia maszynowego i szereg użytych w pojeździe technologii jest w stanie: analizować otoczenie, zaplanować optymalną trasę, rozróżnić znaki drogowe czy reagować na zmiany warunków.

To wszystko jest robione po to aby było jak najbardziej efektywne, a zarazem bezpieczne dostarczenie użytkownika do celu. W pełni autonomiczne pojazdy, choć są już produkowane i wprowadzane do użytku jako autonomiczne taksówki w USA czy Chinach, pozostają jeszcze przyszłością. Obserwując rozwój i stałe udoskonalanie technologii w tym zakresie, trzeba

jednak pamiętać, że nie może ona zastąpić rozsądku kierowcy, przynajmniej na razie. Udział ludzi ma nadal większe znaczenie w kwestii prowadzenia pojazdów i być może nic nigdy nie będzie w stanie w pełni zająć miejsca człowieka.

4. Samochodowe technologie oparte na sztucznej inteligencji

Na co dzień widzimy w naszych samochodach, że są powiązane z działaniem komputera i sztucznej inteligencji. Można więc zauważyć, że sztuczna inteligencja ma istotny wpływ na bezpieczeństwo na drogach. Używa się sztucznej inteligencji do nawigacji samochodowej w celu optymalizacji trasy pod kątem przebiegu, czasu zużycia paliwa i wpływu na środowisko na podstawie bieżącej analizy danych o ruchu drogowym. Sztuczna inteligencja w motoryzacji pozwala na zarządzanie ruchem poprzez monitorowanie jego przepływu, dostosowywanie sygnałów drogowych czy generowanie alternatywnych tras. Algorytmy sztucznej inteligencji zwiększają bezpieczeństwo za pomocą zaawansowanych systemów wspomagania kierowcy.

Zaawansowane systemy wspomagania kierowcy ADAS możemy je spotkać na co dzień. Te systemy używają sztuczną inteligencję, nie są bynajmniej domeną wyłącznie samochodów najnowszych generacji. Asystent hamowania, który wspomaga reakcję kierowców, gdy potrzebne jest gwałtowne i intensywne hamowanie. Inteligentny asystent prędkości, który kontroluje prędkość pojazdu i może dostosować ją do zagrożeń na drodze. Elektroniczny system stabilizacji toru jazdy, który zapobiega utracie przyczepności podczas pokonywania zakrętu. System adaptacyjnego tempomatu, który pozwala zachować odpowiedni odstęp między pojazdami poruszającymi się po tym samym pasie. TSR - system rozpoznawania znaków drogowych, który informuje kierowcę o oznaczeniach na drodze. Drive Alert – system wykrywania zmęczenia kierowcy, który analizuje zachowanie kierowcy, ostrzegając o wykrytych objawach zmęczenia. TPMS – system kontroli ciśnienia w oponach, który informuje kierowcę o zbliżaniu się lub przekroczeniu znaków poziomych na jezdni.

Sztuczna inteligencja w motoryzacji nie jest tylko wykorzystywana w systemach wspomagania kierowcy. Korzysta się z niej również podczas produkcji samochodów. Służy do tworzenia symulacji testujących reakcje samochodu na różne, nawet ekstremalne sytuacje. Jest też wykorzystywana do optymalizowania projektów części mechanicznych, dzięki temu koszty są obniżone.

5. Siły napędowe sztucznej inteligencji w samochodach

Ulepszone mechanizmy bezpieczeństwa są jednym z głównych czynników sprzyjających sztuczną inteligencję w samochodach. Pojazdy mogą przewidywać zagrożenia, zareagować na dynamiczne warunki drogowe a nawet zapobiegać wypadkom, dzięki zdolności sztucznej inteligencji do szybkiego przetwarzania ogromnej ilości danych. Występują już zaawansowane systemy wspomagania kierowcy ADAS (Advanced Driver Assistance Systems) wykorzystując sztuczną inteligencję, aby działały takie mechanizmy w pojazdach jak wspomaganie utrzymania pasa ruchu, unikanie kolizji czy automatyczne hamowanie awaryjne.

Sztuczna inteligencja zmienia doświadczenie z jazdy takich jak przyzwyczajenie się do pewnych nawyków i preferencji kierowcy. Od personalizacji ustawień ulubionej rozrywki i otoczenia aż do przewidywanych momentów, kiedy pojazd będzie wymagał konserwacji. Sztuczna inteligencja zapewnia płynną i spersonalizowaną podróż.

Dzięki sztucznej inteligencji pojazdy mogą komunikować się między sobą a także z systemami zarządzania ruchem, aby zapewnić płynniejszy przepływ ruchu, są to korzyści dla zwiększenia ogólnej wydajność dróg. Dodatkowo, sztuczna inteligencja może również być wykorzystana do optymalizacji wyznaczonych tras podróży, sugerując alternatywne trasy w przypadku wystąpienia wypadków, remontów drogowych czy korków. Korzystają z takich rozwiązań kierowcy mogą wybierać bardziej efektywne i bezpieczne trasy, co prowadzi do oszczędzenia paliwa i czasu kierowcy.

Rysunek 54 Kierunek w kierunku przyszłości opartej na sztucznej inteligencji, źródło: https://no1-reklama.pl/rewolucja-ai-w-samochodach/

6. Funkcje o wartości dodanej – sztuczna inteligencja poza jazdą

Systemy oparte na sztucznej inteligencji również dostarczają wiadomości, rozrywkę, wyznaczone trasy w oparciu o preferencje i nawyki kierowcy. Systemy takie poprawiają jakoś wrażeń z podróży, co sprawia, że jazda staje się przyjemniejsza i dużo bardziej świadoma. W dużej mierze dzięki nowym możliwością sztucznej inteligencji możliwe jest zdiagnozowanie potencjalnych problemów mechanicznych, z dużym wyprzedzeniem, co pozwala na zmniejszanie ryzyka nagłych awarii, które powodują wypadki.

Komunikacja V2X (pojazd-wszystko) oparta na sztucznej inteligencji pozwala na interakcję ze wszystkim w otoczeniu, mogą to był przedmioty takie jak sygnalizacja świetlna jak i również poruszający się ludzie na pasach, zapewniając optymalność wykorzystania dróg i bezpieczeństwa.

7. ADAS i DMS

Sztuczna inteligencja jest kluczowa w rozwoju i funkcjonalności technologii ADAS i DMS. Są to technologie, które mogą poprawić bezpieczeństwo i efektywność pojazdów autonomicznych. ADAS jest to Advanced Driver Assistance Systems, czyli w skrócie są to układy stanowiące aktywne systemy bezpieczeństwa. Technologia DMS, czyli Driver Monitoring System wykorzystuje kombinację czujników do monitorowania zachowania i uwagi kierowcy. To algorytmy sztucznej inteligencji pozwalają tym systemom na przetwarzanie dużej ilości danych, by podejmować inteligentne decyzje i odpowiednią adaptację do dynamicznych warunków jazdy. Sztuczna inteligencja wykorzystuje te technologie w rozpoznawaniu obiektów.

Algorytmy AI analizują dane z czujników kamer w czasie rzeczywistym w celu rozpoznania obiektów. Dotyczą identyfikacji pojazdów, pieszych, znaków drogowych oraz oznakowania drogowego, przez co sztuczna inteligencja umożliwia tym systemom reagowanie w odpowiednim momencie.

Algorytmy sztucznej inteligencji analizują dane dotyczące zachowania kierowcy, które są gromadzone przez system DMS, w celu wykrywania oznak senności, rozproszeń czy nagłego braku uwagi podczas jazdy. Ten algorytm jest w stanie pomóc określić, kiedy wymagane jest zaangażowanie kierowcy, ewentualnie również może wysłać ostrzeżenie, aby przyciągnąć uwagę kierowcy z powrotem na drogę.

Techniki uczenia maszynowego pozwalają systemom ADAS i DMS uczyć się na podstawie doświadczenia z przeszłości i poprawnego działania. Patrząc na przykładowe wzorce w danych, te systemy mogą dostosować i optymalizować swoje funkcje na podstawie jazdy.

Technologie ADAS i DMS zmieniają przemysł motoryzacyjny i definiują, w jaki sposób prowadzimy pojazdy. Wykorzystując te technologie i ich zaawansowane czujniki, sztuczną inteligencję i systemy monitorowania kierowcy, poprawiają dużo bardziej bezpieczeństwo i komfort jazdy, zapobiegają wypadkom, monitorując otoczenie i asystując jednocześnie kierowcy.

8. Cele, korzyści i wady wynikające z pojazdów autonomicznych

Korzyści wynikające z pojazdów autonomicznych to na pewno zmniejszenie liczby kolizji, co oznacza mniej wypadków i zwiększa przepustowość, czyli zmniejsza zatory na drogach i ulicach. Za tym idzie skrócenie czasu podróży, łatwiejsze parkowanie czy pokonywanie dłuższych tras bez konieczności zatrzymywania się co sprawia, że kierowca nie musiałby odpoczywać. Dzięki temu podróże stają się mniej stresujące i bez ograniczeń takich jak obciążenie kierowcy od czynności związanych z prowadzeniem pojazdu, zmniejsza zapotrzebowanie na ciągłą kontrole dróg przez policję i inne służby. Zmniejszenie kradzieży samochodów a także mniejsze koszty zatrudnienia dla firm transportowych.

Sztuczna inteligencja w pojazdach ma też swoje wady. Nie jest to do końca udoskonalone, więc wadą jest kwestia związana z bezpieczeństwem nieprzewidywalności zachowanych sytuacji. Pomimo zaawansowanych algorytmów sztucznej inteligencji, pojazdy autonomiczne mogą nie do końca i nie zawsze radzić sobie lepiej niż człowiek w sytuacjach, których wcześniej nie doświadczyły. Warunki pogodowe bywają różne a wręcz czasami ekstremalne, przez co te systemy mogą nie do końca dobrze działać. Niesprzyjające warunki dla samochodów autonomicznych są też awarie systemów czy niechciana obecność zwierząt na drodze. Dla pojazdów autonomicznych utrudnienie będzie, kiedy na drogach pojawi się całkowite zaśnieżenie, nieczytelne znaki czy niewidoczne linie.

W algorytmach sztucznej inteligencji może wystąpić ryzyko błędu technicznego w pojazdach autonomicznych, jakie wynika z technologii, która może okazać się zawodna. Nie jest do końca pewne, jak zachowa się sieć połączonych samochodów w przypadku błędu technicznego. Bywają też ryzyka ataków hakerskich. Jest to ważne zagrożenie związane z pojazdami autonomicznymi. Mimo tego, że takie samochody są projektowane, żeby były jak najbardziej bezpieczne, istnieje szansa, że ktoś może złamać zabezpieczenia i przejąć kontrolę

nad pojazdem. Takie cyberataki mogą mieć poważne konsekwencje dla bezpieczeństwa pasażerów jak i dla infrastruktury drogowej.

9. Jazda na fali innowacji

Rozpoznawanie głosu jest ekscytującym zastosowaniem sztucznej inteligencji w samochodach. To innowacyjne rozwiązanie pozwala kierowcom sterować równymi funkcjami w pojeździe, takimi jak nawigacja, klimatyzacja czy odtwarzanie muzyki, za pomocą zwykłych, prostych poleceń głosowych. Taka funkcja nie tylko zapewnia większy komfort podczas jazdy ale również znacznie zwiększa bezpieczeństwo, które umożliwia kierowcom skupienie się na drodze, bez konieczności przedstawiania rąk czy wzroku na ekran czy guziki w samochodzie.

Istotnym zastosowaniem sztucznej inteligencji w samochodach jest analiza danych. Pojazdy wyposażone w algorytmy sztucznej inteligencji mogą zbierać, analizować i przetwarzać duże ilości danych z różnych źródeł, takich jak radary, kamery czy czujniki. Dzięki nim, samochody mogą reagować na zmienne warunki drogowe, unikać kolizji, czy nawet przewidywać zagrożenia i potencjalne awarie. Stosowanie analizy danych również pozwala na doskonalenie systemów autonomicznych, które kiedyś w przyszłości mogą całkowicie wyeliminować potrzebę człowieka jako kierowcy.

Sztuczna inteligencja w samochodach autonomicznych jest zaledwie jak wierzchołek góry lodowej, jeśli chodzi o innowację całej branży motoryzacyjnej. Dzięki ciągłemu rozwojowi tej technologii, możemy spodziewać się coraz większych osiągnięć w najbliższych latach. Warto jest obserwować na bieżąco jakie niesie ze sobą sztuczna inteligencja w samochodach.

10. Obniżenie kosztów montażu i kontroli w produkcjach zastosowania sztucznej inteligencji

Sztuczna inteligencja znajduje szerokie zastosowanie w branży motoryzacyjnej. Dzięki zastosowaniu systemów sztucznej inteligencji, firmy mogą dużo bardziej zwiększyć bezpieczeństwo swoich pojazdów oraz zwiększyć wydajność w procesie produkcyjnym, dzięki temu pozwala na obniżenie kosztów i strat oraz lepsze wykorzystanie dostępnych surowców. Dodatkowo, dzięki zastosowaniu sztucznej inteligencji w branży motoryzacyjnej, możliwe jest tworzenie nowoczesnych, innowacyjnych rozwiązań technologicznych, które zwiększają komfort jazdy kierowcy a także pasażerów oraz zmniejszają emisję szkodliwych substancji do atmosfery.

11. Przyszłość sztucznej inteligencji w samochodach

W branży motoryzacyjnej sztuczna inteligencja wciąż się rozwija, a jej przyszłość wygląda obiecująco. Przewiduje się, że coraz bardziej zaawansowane stają się samochody autonomiczne, dzięki stale udoskonalanym algorytmom i rozwijającym się technologią. Sztuczna inteligencja będzie odgrywać ważną rolę w rozwoju ekologicznym pojazdów, dzięki optymalizacji zużycia energii i redukcji emisji C02. Istnieje również chęć wykorzystania sztucznej inteligencji w rozwoju systemów interakcji człowiek-maszyna, takich tak inteligentny asystent podróży czy rozszerzona rzeczywistość.

Sztuczna inteligencja ma ogromny wpływ na przemysł motoryzacyjny. Wykorzystanie jej w systemach bezpieczeństwa samochodowego, pojazdach autonomicznych, personalizacji doświadczenia kierowcy i innych przyszłych perspektyw otwiera nowsze możliwości dla rozwoju inteligentnych pojazdów. Wraz z rozwojem tych zalet pojawiają się również wyzwania, takie jak kwestie bezpieczeństwa, akceptacja społeczeństwa czy prywatność danych. Takie wprowadzenie do samochodów sztucznej inteligencji wymaga odpowiednich regulacji i dbałości o równowagę między czynnikiem ludzkim a technologią.

12. Podsumowanie

Artykuł podkreśla, że sztuczna inteligencja odgrywa coraz większą rolę w przemyśle motoryzacyjnym, przynosząc za sobą zarówno nowe możliwości, jak i wyzwania. Wskazuje na konieczność współpracy między sektorem przemysłowym a konsumentami w celu zapewnienia bezpiecznej i efektywnej przyszłości samochodów autonomicznych.

W miarę upływu czasu wkraczania sztucznej inteligencji w przemyśle motoryzacyjnym jest, że przyszłość pojazdów będzie bardziej inteligentniejsza, wydajniejsza i bezpieczniejsza. Ważne jest a nawet konieczne, żeby cały przemysł związany ze sztuczną inteligencją został zjednoczony przez konsumentów i decydentów aby zapewnić dobry i bezpieczny start w przyszłość mobilności.

Wdrożenie w przyszłość sztucznej inteligencji to nie jest tylko postęp technologiczny, ale również chodzi o stworzenie przyszłości, której nasze podróże będą czymś więcej niż tylko zwykłą jazdą.

Źródła internetowe:

1. https://elektromobilni.pl/sztuczna-inteligencja-w-motoryzacji-czyli-przyszlosc-ktora-jest/ (dostęp: 24.04.2024).
2. https://mubi.pl/poradniki/samochod-autonomiczny/ (dostęp: 24.04.2024).

3. https://mubi.pl/poradniki/sztuczna-inteligencja-w-samochodach/ (dostęp: 24.04.2024).
4. https://no1-reklama.pl/rewolucja-ai-w-samochodach/ https://www.addsecure.pl/blog/wszystko-co-powinienes-wiedziec-o-technologii-adas-i-dms/ (dostęp: 24.04.2024).
5. https://www.addsecure.pl/blog/wszystko-co-powinienes-wiedziec-o-technologii-adas-i-dms/ (dostęp: 24.04.2024).
6. https://www.rental-planet.pl/wykorzystanie-sztucznej-inteligencji-w-samochodach-systemy-nawigacyjne-rozpoznawanie-glosu-analiza-danych/ (dostęp: 24.04.2024).
7. https://www.researchgate.net/profile/Svitlana-Ivanova/publication/332333681_Strategia_poiska_proektnoj_idei/links/5fc6678fa6fdcc92169cc7fa/Strategia-poiska-proektnoj-idei.pdf#page=185 (dostęp: 24.04.2024).
8. https://www.smartney.pl/blog/lifestyle/sztuczna-inteligencja-w-samochodach/(dostęp: 24.04.2024).

Jakub Jucha, Adam Krawczyk, Sebastian Cwynar, Hubert Kraus, Maciej Karczmarz

dr. inż. Bartosz TRYBUS
Opiekun naukowy

Rozpoznawanie cyfr zbioru danych MNIST za pomocą sieci głębokiej

Streszczenie

Artykuł omawia zastosowanie głębokiej sieci neuronowej do rozpoznawania zbioru danych MNIST, który jest standardowym zestawem benchmarkowym używanym do testowania algorytmów rozpoznawania pisma ręcznego. Szczególną uwagę poświęcono sieciom konwulacyjnym (CNN), które są szczególnie skuteczne w analizie obrazów dzięki zdolności do automatycznego wykrywania cech istotnych w różnych regionach obrazu. W artykule przedstawiono strukturę typowej sieci konwulacyjnej, w tym warstwy konwulacyjne, poolingowe oraz w pełni połączone. Ponadto, omówione są wyzwania związane z trenowaniem głębokiej sieci, takie jak problem zanikającego gradientu, który utrudnia efektywne uczenie się bardzo głębokich sieci. Przedstawiono techniki zapobiegające temu problemowi, w tym normalizację batchową (batch normalization) oraz zastosowanie funkcji aktywacji takiej jak ReLU, która pomaga w propagacji gradientu przez sieć. Innym omówionym problemem jest przetrenowanie (overfitting), które występuje, gdy model uczy się zbyt dokładnie wzorców treningowych, co prowadzi do słabego generalizowania na nowych danych. Aby przeciwdziałać przetrenowaniu, zaproponowano metodę regularyzacji jaką jest dropout.

Słowa kluczowe: Sztuczna inteligencja, CNN, konwulacja, gradient błędu.

1. Wprowadzenie

Sztuczna inteligencja jest dziedziną nauki zajmującą się tworzeniem systemów zdolnych do wykonywania zadać, które wymagają inteligencji ludzkiej, takich jak rozpoznawanie obrazów, przetwarzanie języka naturalnego czy podejmowanie decyzji. Jednym z kluczowych narzędzi w sztucznej inteligencji są głębokie sieci neuronowe, w szczególności sieci konwulacyjne (CNN), które zostały zaprojektowane do efektywnej analizy danych obrazowych. Sieci konwulacujne składają się z warstw konwulacyjnych, które automatycznie wykrywają istotne cechy w obrazach, oraz warstw poolingowych i w pełni połączonych, które pomagają w redukcji wymiarów i klasyfikacji.

Trenowanie głębokich sieci wiąże się jednak z pewnymi wyzwaniami. Jednym z nich jest problem zanikającego gradientu, który pojawia się, gdy gradienty używane do aktualizacji was sieci stają się bardzo małe, co utrudnia efektywne uczenie się głębokich warstw. Aby temu zapobiegać, stosuje się techniki takie jak normalizacja danych oraz funkcje aktywacji typu ReLU, które pomagają w propagacji gradientu błędu przez sieć.

Innym istotnym problemem jest przetrenowanie sieci (overfitting), które ma miejsce, gdy model zbytnio dopasowuje się do danych treningowych kosztem zdolności do generalizacji na nowych danych. W celu zapobiegania przetrenowaniu, używa się metod regularyzacji, takich jak dropout który losowo deaktywuje neurony w trakcie treningu, oraz wczesne zatrzymywanie (early stopping), które kończy trening, gdy model zaczyna wykazywać oznaki przetrenowania.

2. Początki sieci konwulacyjnych (CNN)

W ostatnim dziesięcioleciu obserwujemy dynamiczny wzrost zainteresowania głębokimi sieciami kontulacyjnymi (CNN), co przyczyniło się powstaniu specjalistycznych architektur specjalizujących się w zadaniach, takich jak klasyfikacja zdjęć, detekcja obiektów i wiele innych.

Jednym z pierwszych modeli CNN był LeNet, który został opracowany przez Yanna LeCuna w 1989 roku. Zaproponował konwulacyjną sieć neuronową trenowaną przez algorytm wstecznej propagacji do rozpoznawania odręcznie pisanych liczb na zdjęciach. Ostateczna wersja tej sieci powstała w 1990 roku. Wtedy też została spopularyzowana baza danych MNIST. Zawiera ona około 70 tysięcy skanów zdjęć cyfr pisanych odręcznie w rozdzielczości 28x28 pikseli.

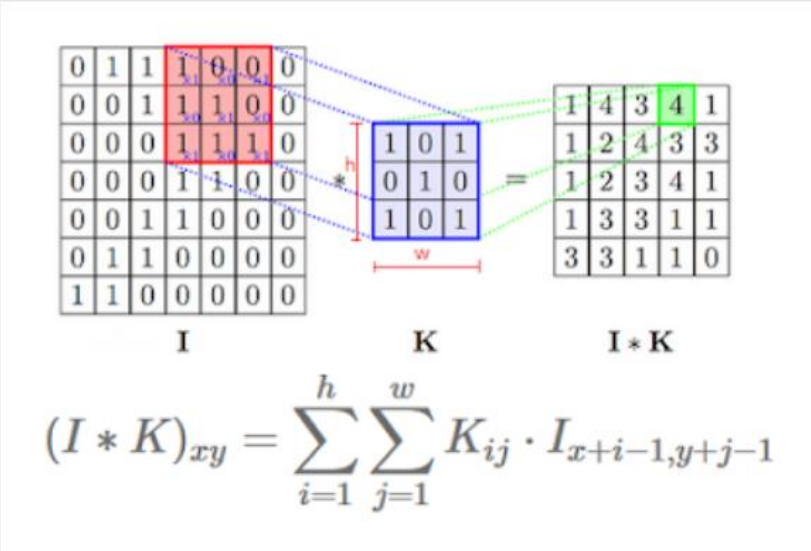
Od 2010 roku sieci CNN szybko się rozwijały z powodu wyzwania ImageNet Large Scale Visual Recognition Challenge (ILSVRC). Było to coroczny konkurs, podczas którego zespoły prezentowały skuteczność swoich algorytmów służących do wykrywania obiektów i klasyfikacji obrazów na zbiorze danych ImageNet.

Rysunek 55. Zbiór danych MNIST[1]

[1] https://datasets.activeloop.ai/docs/ml/datasets/mnist/ (dostęp: 16.05.2024)

3. Budowa sieci konwulacyjnych (CNN)

Konwulacyjne sieci neuronowe są modelem sieci neuronowych, który potrafi automatycznie definiować wzorce (filtry), użyteczne do rozwiązywania danego zadania. Sieć CNN jest zbudowana z warstw, które nazywamy warstwami konwulacyjnymi. Każda z warstw definiuje zestaw filtrów, które są używane do przekształcenia wejściowego obrazu na nowy za pomocą konwulacji.



Rysunek 56. Wynik operacji konwulacji[2]

Stosując dużo warstw, dokonujemy stopniowej modyfikacji obrazu, co jest skutkiem uzyskania docelowej reprezentacji, która finalnie może być wejściem do ostatniej warstwy odpowiedzialnej za klasyfikację.

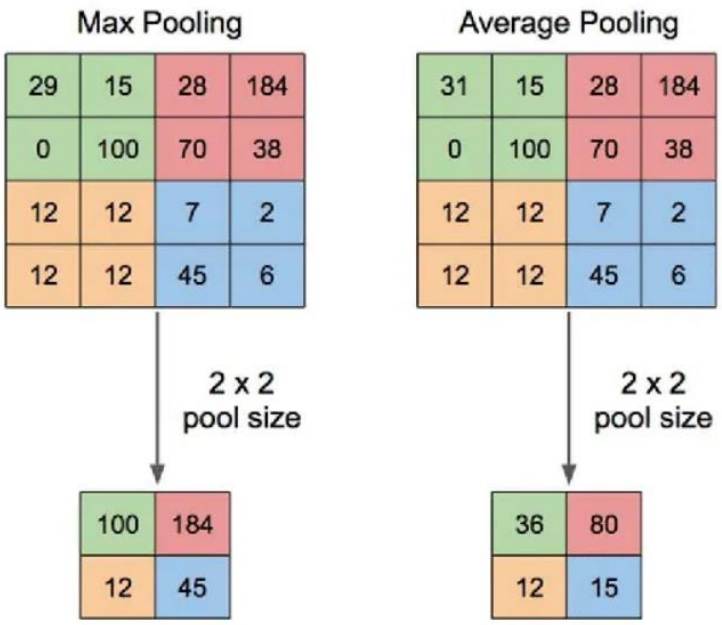
Rozważmy (Rysunek 2.) obraz I, zestaw filtrów K_1, … , K_l oraz funkcję aktywacji *f* definiujące pierwszą warstwę konwulacyjną. Wynikiem działania pierwszej warstwy jest obraz

$I_l = f(I*K_1), \dots , f(I*K_l)$

Następnie na warstwy konwulacyjne nakładane są nieliniowe funkcje aktywacji, gdyż bez zastosowania nieliniowej operacji, złożenie filtrów konwulacyjnych pozostawałoby filtrem konwulacyjnym. W rezultacie uniemożliwiałoby to aproksymowanie skomplikowanych funkcji. Funkcja *f* jest osobno stosowana dla każdego z pikseli.

[2] https://sciagaprogramisty.blogspot.com/2018/01/konwolucja-wstep-do-neuronowych-sieci.html (dostęp: 16.05.2024)

Obraz I_1 jest wynikiem działania warstwy konwulacyjnej. Może być on następnie wejściem do kolejnej warstwy konwulacyjnej. W przypadku stosowania *m* warstw konwulacyjnych otrzymamy ciąg *m* obrazów od I_1 do I_m, które powstaną w kolejnych warstwach. Ostatni zaś obraz, tj. I_m, ukazywał będzie wysokopoziomową reprezentację obrazu początkowego I. W momencie klasyfikacji chcemy dokonać rzutowania I_m do przestrzeni klas. W tym celu po każdej warstwie konwulacyjnej stosuję warstwę splotu. Zadaniem tej warstwy jest grupowanie pobliskich wartości w jedną. Pomaga to zmniejszyć wymiary przestrzenne objętości wejściowej, zmniejszając w ten sposób złożoność obliczeniową.



Rysunek 57. Operacja poolingu[3]

Najpopularniejszym filtrem max poolingu jest 2x2 (Rysunek 3.). Oznacza to, że macierz obrazu jest dzielona na macierze wspomnianego rozmiaru. Następnie wybierana jest najwyższa wartość i zapisywana do nowej macierzy. Zastosowanie poolingu sprawia, że obraz staje się coraz mniejszy wraz z przechodzeniem przez kolejne warstwy sieci, ale równocześnie staje się coraz głębszy i klarowniejszy dla klasyfikatora, dzięki użyciu coraz większej liczby filtrów.

Wyniki warstw konwulacynych i max poolingu wprowadzane są następnie do klasycznej sieci neuronowej, takiej jak na przykład perceptron wielowarstwowy, gdzie są klasyfikowane.

[3] https://www.researchgate.net/figure/Illustration-of-Max-Pooling-and-Average-Pooling-Figure-2-above-shows-an-example-of-max_fig2_333593451 (dostęp: 16.05.2024)

4. Problem zanikającego gradientu

Aby lepiej zrozumieć czym jest problem zanikającego gradientu należy najpierw wprowadzić pojęcia jakie jak propagacja wsteczna i waga w sieciach neuronowych.

Propagacja wsteczna (backpropagation) jest to algorytm wykorzystywany do treningu sieci neuronowych, w tym wspomnianych sieci konwulacyjnych. Polega ona na wyliczaniu gradientu funkcji kosztu (czyli błędu) w odniesieniu do wag sieci i wykorzystaniu tego gradientu do aktualizacji wag w celu minimalizacji błędu. Proces ten odbywa się w kilku krokach:

1. Propagacja w przód (forward pass): Dane wejściowe przechodzą przez sieć neuronową, warstwa po warstwie, aż do warstwy wyjściowej, gdzie obliczany jest wynik (przewidywanie),
2. Obliczanie błędu (loss calculation): Na podstawie wyników przewidywania i rzeczywistych wartości (np. etykiet w przypadku klasyfikacji) oblicza się błąd przy użyciu określonej funkcji kosztu,
3. Propagacja wsteczna (backward padd): Błąd jest propagowany wstecz przez sieć, zaczynając od warstwy wyjściowej, poprzez wszystkie warstwy ukryte, aż do warstwy wejściowej. Podczas tego procesu obliczane są gradienty błędu względem wag,
4. Aktualizacja wag (weight update): Wagi sieci są aktualizowane na podstawie obliczonych gradientów, zazwyczaj przy użyciu algorytmu optymalizacji takiego jak gradient prosty, Adam lub RMSprop.

Wagi (weights) w sieciach neuronowych to wartości, jakie określają siłę połączeń pomiędzy neuronami. W każdej warstwie sieci neuronowej każdy neuron jest połączony z neuronami z poprzedniej warstwy za pomocą wag. Wagi są kluczowe dla funkcjonowania sieci, ponieważ określają, jak silnie sygnał z jednego neuronu wpływa na neuron w następnej warstwie. Podczas treningu sieci neuronowej, wagi są iteracyjnie dostosowywane w celu minimalizacji błędu między przewidywaniami sieci, a rzeczywistymi wynikami.

Problem zanikania gradientu w sieciach konwulacyjnych jest jednym z kluczowych wyzwań w dziedzinie uczenia głębokiego, zwłaszcza gdy sieci stają się coraz głębsze. Zjawisko to polega na tym, że podczas procesu wstecznej propagacji gradienty błędów zmniejszają się

wraz z głębokością sieci. W rezultacie, wagi w początkowych warstwach sieci są aktualizowane bardzo wolno lub praktycznie wcale.

Istnieje kilka czynników, które przyczyniają się do zanikania gradientu w sieciach konwulacyjnych (CNN). Jednym z głównych jest stosowanie funkcji aktywacji, takiej jak sigmoidalna funkcja logiczna, której pochodna ma małe wartości dla dużych lub małych wartości wejściowych, co prowadzi do malejących gradientów w głębokich warstwach. Innym czynnikiem jest inicjalizacja wag sieci, która może również wpłynąć na to, jak szybko gradienty maleją w trakcie propagacji wstecznej. Wzór na zanikający gradient można przedstawić w następujący sposób:

Niech $\frac{\partial J}{\partial W}$ oznacza gradient funkcji kosztu J względem wag W. Wówczas gradient dla danej warstwy można obliczyć jako iloczyn gradientu warstwy powyżej i pochodnej funkcji aktywacji. Przykładowo dla dwóch warstw[4]:

$$\delta_2 = \frac{\partial J}{\partial A_2} \cdot g'(z_2)$$

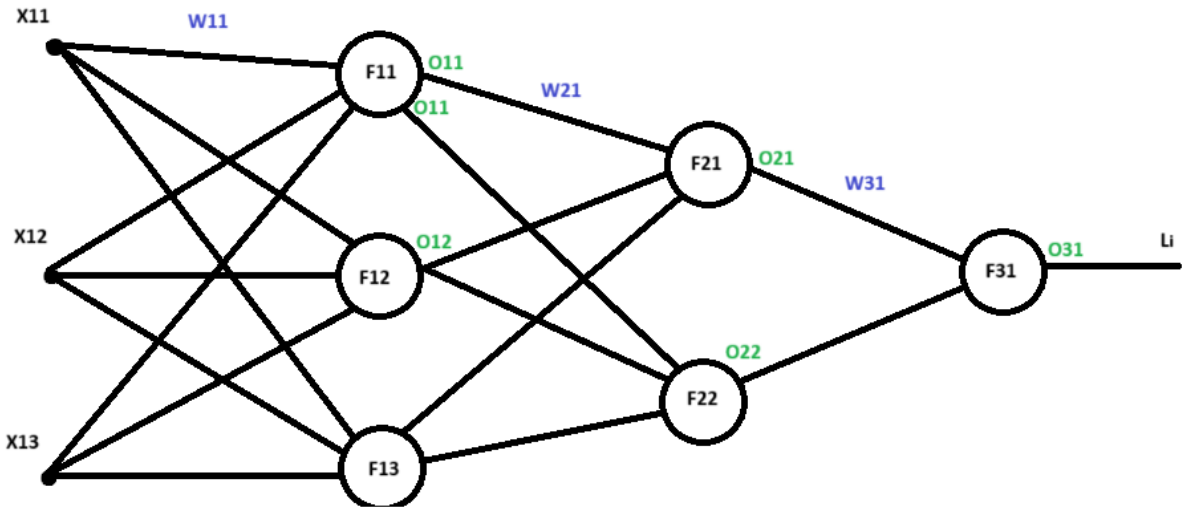
$$\delta_1 = W_2^T \cdot \delta_2 \cdot g'(z_1)$$

Gdzie:

- δ_2 to gradient dla warstwy drugiej (bliżej wyjścia sieci)
- δ_1 to gradient dla warstwy pierwszej (bliżej wejścia sieci)
- $\frac{\partial J}{\partial A_2}$ to gradient funkcji kosztu J względem wyjścia warstwy drugiej
- $g'(z)$ to pochodna funkcji aktywacji względem wejścia z
- W_2 to macierz wag dla warstwy drugiej
- z_1 to wejście do pierwszej warstwy
- z_2 to wejście do drugiej warstwy

Aby lepiej przybliżyć problem zanikającego gradientu rozważmy poniższy rysunek (Rysunek 4).

[4] https://medium.com/@amanatulla1606/vanishing-gradient-problem-in-deep-learning-understanding-intuition-and-solutions-da90ef4ecb54 (dostęp: 16.05.2024)



Rysunek 58. Model prostej sieci neuronowej

Gdzie F to funkcja aktywacji, O to wyjście neuronu, a W to waga (patrz Rysunek 4). W trakcie propagacji wstecznej celem aktualizacji wagi np. W_{11} stosujemy poniższy wzór:

$$W_{11(new)} = W_{11(old)} - \eta \frac{\partial L_i}{\partial W_{11(old)}}$$

- $W_{11(old)}$ – to waga która podlega procesowi uczenia (waga którą aktualizujemy)
- η - to współczynnik uczenia (learning rate)
- $\frac{\partial L_i}{\partial W_{11(old)}}$ – to pochodna cząstkowa funkcji straty względem wagi

Rozważmy równanie $\frac{\partial L_i}{\partial W_{11(old)}}$. Możemy je przekształcić i zapisać w następujący sposób:

$$\frac{\partial L_i}{\partial W_{11}} = \frac{\partial L_i}{\partial O_{31}} * \frac{\partial O_{31}}{\partial W_{11}}$$

Dzięki takiemu zabiegowi jesteśmy w stanie zapisać pochodną wyjścia neuronu O_{31} względem wagi W_{11} to jest $\frac{\partial O_{31}}{\partial W_{11}}$ w taki sposób, aby uwzględnić wszystkie ścieżki od warstwy wyjściowej do wagi którą aktualizujemy:

$$\frac{\partial O_{31}}{\partial W_{11}} = \frac{\partial O_{31}}{\partial O_{21}} * \frac{\partial O_{21}}{\partial O_{11}} * \frac{\partial O_{11}}{\partial W_{11}} + \frac{\partial O_{31}}{\partial O_{22}} * \frac{\partial O_{22}}{\partial O_{11}} * \frac{\partial O_{11}}{\partial W_{11}}$$

Ważną kwestią na którą należy zwrócić uwagę jest to, że w powyższym równaniu bierzemy pod uwagę pochodne wyjść neuronów. Co ważne, stosunki pochodnych wyjść neuronów są niczym innym pochodną funkcji aktywacji. Wynika to z faktu, że wynik jest wynikiem zastosowania sum funkcji aktywacji.

W związku z tym rozważaniem $\frac{\partial O_{31}}{\partial O_{21}}$ pochodna ta jest niczym innym jak pochodną funkcji aktywacji. Z artykule wspomniano, że najpopularniejszą funkcją aktywacji jest funkcja sigmoidalna, której pochodna osiąga maksymalną wartość 0,25. Przyjmując przykładowe wartości załóżmy, że:

$$\frac{\partial O_{31}}{\partial O_{21}} = 0.2 \qquad \frac{\partial O_{21}}{\partial O_{11}} = 0.1 \qquad \frac{\partial O_{11}}{\partial W_{11}} = 0.05 \qquad \frac{\partial O_{31}}{\partial O_{22}} = 0.1 \qquad \frac{\partial O_{22}}{\partial O_{11}} = 0.05,$$

To $\frac{\partial O_{31}}{\partial W_{11}}$ przyjmuje wartość równą 0.00125. Jeżeli teraz wrócimy do wzoru na aktualizację wagi i przyjmiemy że w pierwszym etapie szkolenia sieci neuronowej, waga W_{11} została przyjęta na wartość 1,5 a współczynnik uczenia (learning rate) na 0,01 to:

$$W_{11(new)} = W_{11(old)} - \eta\frac{\partial L_i}{\partial W_{11(old)}} = 1.5 - 0.01*0.00125 = 1.4999875$$

Jak można zauważyć nowa waga względem starej wagi praktycznie nie uległą zmianie. Zjawisko to nazywamy powszechnie zjawiskiem zanikającego gradientu.

Aby zapobiegać problemowi zanikającego gradientu, stosuje się inne techniki takie jak na przykład używanie funkcji aktywacji ReLU, inicjalizacja wag, normalizacja danych czy dodanie połączeń omijających (dropout).

Funkcja aktywacji ReLU jest jedną z najczęściej stosowanych funkcji aktywacji w sieciach neuronowych. Jest to prosta funkcja nieliniowa, która zwraca zero dla wartości ujemnych i równa się wartości wejściowej dla wartości nieujemnej. Matematycznie można ją zdefiniować jako:

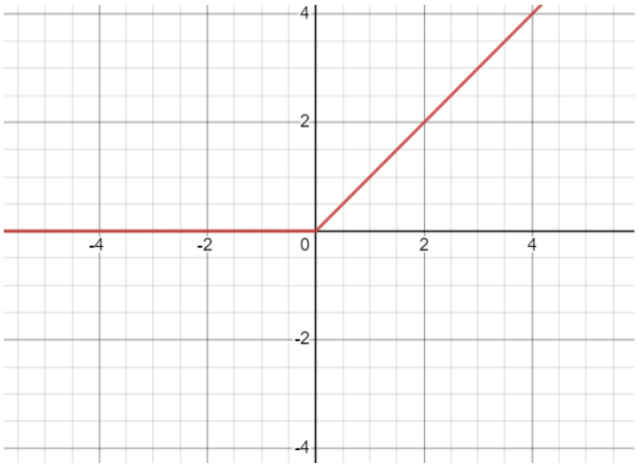
$$f(x) = \max(0,x)$$

Wartości ujemne są przycinane do zera, co sprawia, że funkcja jest niewrażliwa na dużą ilość szumu lub wartości ujemne, co może pomóc w radzeniu sobie z zanikającym gradientem.

Inicjalizacja wag jest jednym z kluczowych kroków w procesie tworzenia i trenowania sieci neuronowej. Polega ona na nadawaniu początkowych wartości wagom w sieci neuronowej, które następnie podlegają procesowi uczenia się w celu zminimalizowania funkcji kosztu. Poprawna inicjalizacja wag może znaczenie wpłynąć na skuteczność uczenia się sieci neuronowej i szybkości zbieżności do optymalnych rozwiązań jak i zapobieganiu problemowi zanikającego gradientu.

Jedną z popularniejszych metod inicjalizacji wag jest inicjalizacja losowa, w której wagi są na początku ustawiane w sposób losowy. Jednak inicjalizowanie losowe może doprowadzić do niestabilności oraz do wolniejszego procesu uczenia, szczególnie w przypadkach gdy sieci są

głębokie. Dlatego istnieją bardziej zaawansowane metody inicjalizacji wag, które charakteryzują się dużo większą dokładnością.



Rysunek 59. Wykres funkcji aktywacji ReLU[5]

Jedną z takich metod jest inicjalizacja zwana inicjalizacją Xaviera. Metoda ta inicjalizuje wagi losowo, ale zgodnie z rozkładem Gaussa, czyli o średniej równej 0 i wariancji zależnej od liczby neuronów w warstwie poprzedniej i następnej. Wagi są skalowane przez pierwiastek z liczby neuronów w poprzedniej warstwie, co pozwala na zachowanie większej stabilności rozkładu aktywacji w trakcie propagacji wstecz oraz w przód.

Natomiast normalizacja danych jest jedną z najpowszechniejszych technik stosowanych w sieciach neuronowych w celu zmniejszenia zakresu wartości wejściowych. Jest to celowy zabieg ułatwiający uczenie się modelu poprzez zapewnienie bardziej stabilnego i efektywnego procesu optymalizacji.

Połączenia pomijające (skip connecting), zwane też połączeniami przeskakującymi, są kluczowym elementem w architekturze głębokich sieci neuronowych. Cała idea polega przekazywaniu informacji z jednej warstwy neuronowej do innej, omijając przy tym część warstw pośrednich. W ten sposób można uniknąć utraty informacji w procesie propagacji wstecznej, zwłaszcza w przypadku bardzo głębokich sieci, które mogą być podatne na zanikanie gradientu. Połączenia przeskakujące umożliwiają również efektywniejsze uczenie się

[5] https://www.capsule.pl/funkcja-akywacji/ (dostęp: 16.05.2024)

przez sieć neuronową, zapewniając połączenia, które umożliwiają przekazywanie istotnych informacji bez niepotrzebnego powtarzania.

5. Implementacja sieci neuronowej

W pierwszym etapie implementacji sieci neuronowej należy ustawić parametry uczenia na których model będzie się szkolił.

```
10    n_epochs = 1
11    batch_size = 64
12    batch_size_test = 1
13    learning_rate = 0.001
14    log_interval = 10
15    random_seed = 123
16
17    out_channels3 = [5,10,25,50,100]
18    out_channels2 = [20,30,50,100,150]
```

Listing 23. Parametry uczenia.

Parametry te określają:

- n_epoch – liczba epok na jakich model będzie się szkolił (inaczej mówiąc, jest to liczba pełnych przejść przez zbiór danych na których model uczy się)
- batch_size – liczba próbek w pojedynczym batchu podczas treningu (oznacza to, że w każdej epoce model będzie uczył się na 64 zdjęciach)
- batch_size_test – liczba próbek na których model będzie testowany
- learning_rate – tempo w jakim model będzie się uczył
- log_interval – parametr ten określa co ile batchów wyświetli się informacja o postępie treningu
- random_seed – ziarno losowości dla powtarzalności wyników
- out_channels3 i out_channels2 – listy liczby neuronów dla różnych warstw

```
device = torch.device("cuda" if torch.cuda.is_available() else torch.device('cpu'))
num_of_devices = torch.cuda.device_count()
print(f"Number of devices is {num_of_devices}")
#print(f"Device name is {torch.cuda.get_device_name()}")

std = (0.3081)

mu = (0.1307,)

#używam dataloader
train_loader = torch.utils.data.DataLoader(
    torchvision.datasets.MNIST( root: '/filse/', train=True, download=True, transform=torchvision.transforms.Compose([
        torchvision.transforms.ToTensor(),
        torchvision.transforms.Normalize(mu,std)
    ])),
    batch_size = batch_size, shuffle = True
)

test_loader = torch.utils.data.DataLoader(
    torchvision.datasets.MNIST( root: '/filse/', train=False, download=True, transform=torchvision.transforms.Compose([
        torchvision.transforms.ToTensor(),
        torchvision.transforms.Normalize(mu,std)
    ])),
    batch_size = batch_size, shuffle = True
)
```

Listing 24. Inicjalizacja urządzenia i ładowanie danych

W kolejnym etapie implementacji należy wybrać urządzenie na którym będą wykonywanie obliczenia (GPU lub CPU). W tym celu wykorzystana została biblioteka PyTorch (patrz Listing 2). Gdy urządzenie zostało poprawnie wybrane, można przystąpić do wczytywania danych na których model będzie się uczył (train_loader) oraz danych na których model będzie testowany (test_loader). W tym etapie dokonano również normalizacji danych w celu zmniejszenia złożoności obliczeniowej sieci.

Mając dostęp do danych uczących można przystąpić do realizacji klasyfikatora (Listing 3). Do poprawnej implementacji klasyfikatora należy utworzyć klasę, która będzie dziedziczyć po klasie torch.nn.Module. W konstruktorze utworzonej klasy należy teraz określić ilość warstw konwulacyjnych wraz z liczbą neuronów na wejściu i na wyjściu oraz ilość warstw liniowych również z liczbą neuronów na wejściu i na wyjściu warstwy. Dodatkowo inicjalizuje się dropout który w kolejnych metodach będzie wykorzystywany do losowego wyłączania neuronów w warstwach. Aby poprawnie obliczyć ilość neuronów na wejściu do pierwszej warstwy liniowej należy od rozmiaru danych wejściowych (w tym przypadku obrazu 28x28 pikseli) odjąć rozmiar filtru (kernel_size) a następnie podzielić całkowicie przez wartość max_poolingu. Ponieważ w kodzie występują dwie warstwy konwulacyjne, proces obliczania należy wykonać dwa razy. Po obliczeniu rozmiaru wyjścia wiadomo, że liczba neuronów w pierwszej warstwie

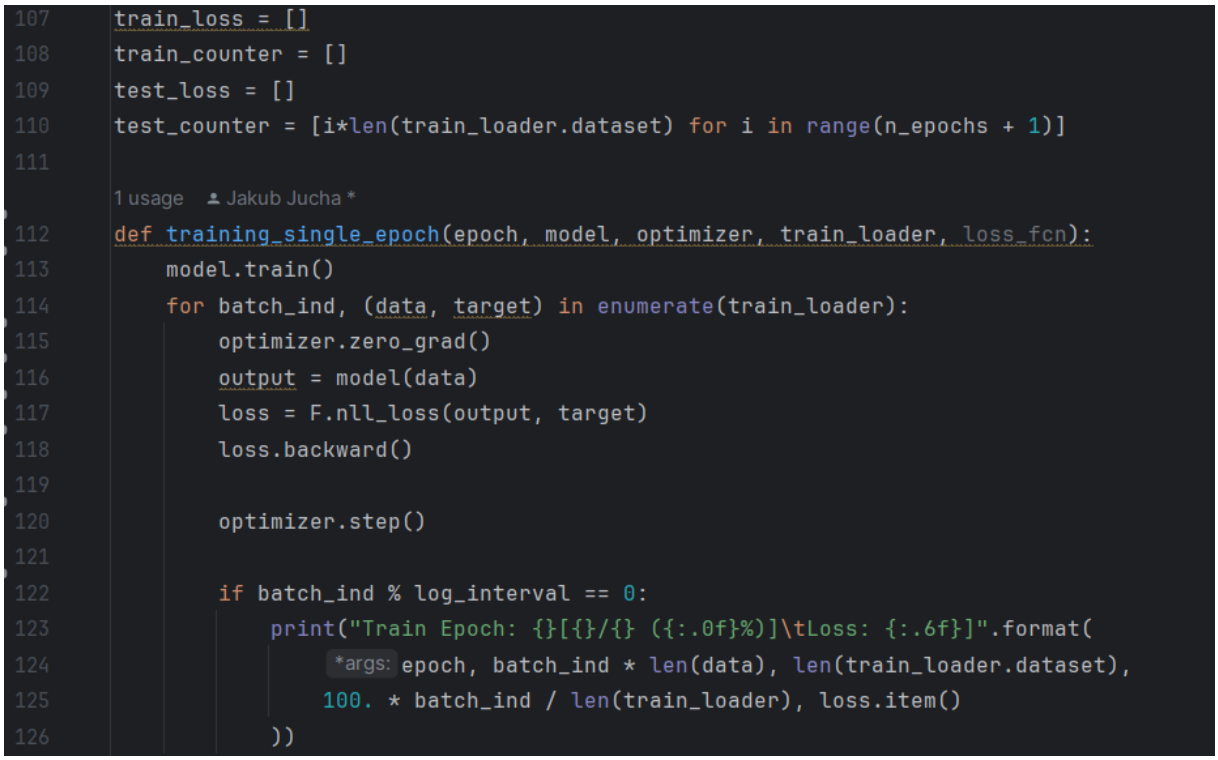
liniowej będzie równa liczbie neuronów na wyjściu drugiej warstwy konwulacyjnen pomnożonej przez kwadrat rozmiaru wyjścia.

```
class Network(nn.Module):
    Jakub Jucha *
    def __init__(self, out_channels3, out_channels2):
        super(Network, self).__init__()
        self.conv1 = nn.Conv2d(in_channels=1, out_channels=40, kernel_size=9)
        self.conv2 = nn.Conv2d(in_channels=40, out_channels=out_channels2, kernel_size=9)

        self.conv2_drop = nn.Dropout2d()

        self._calculate_linear_input_size()
        self.linear1 = nn.Linear(self.linear1_input_size, out_channels3)
        self.linear2 = nn.Linear(out_channels3,  out_features: 10)
    1 usage  Jakub Jucha *
    def _calculate_linear_input_size(self):
        output_size = ((28 - 8) // 2 - 8) // 2
        self.linear1_input_size = self.conv2.out_channels * output_size * output_size
    Jakub Jucha
    def forward(self, x):
        x = self.conv1(x)
        x = F.max_pool2d(x, kernel_size=2)
        x = F.relu(x)
        x = self.conv2(x)
        x = F.max_pool2d(x, kernel_size=2)
        x = F.dropout(x, p=0.5, training=self.training)
        x = x.view(-1, self.linear1_input_size)
        x = self.linear1(x)
        x = F.relu(x)
        x = F.dropout(x, p=0.5, training=self.training)
        x = self.linear2(x)
        return F.log_softmax(x, dim=1)
```

Listing 25. Implementacja klasyfikatora.

Metoda forward definiuje przepływ danych przez sieć neuronową. Na początku dane przechodzą przez pierwszą warstwę konwulacyjną, po której następuje warstwa max pooling zmniejszająca rozmiar danych oraz funkcja aktywacji ReLU dodająca nieliniowość. Następnie dane trafiają do drugiej warstwy konwulacyjnej, po której znów następuje max pooling oraz dropout w celu regularyzacji. Po tych operacjach tensor jest spłaszczany po postaci wektora, który można podać do pierwszej warstwy liniowej. Wektor przechodzi przez kolejną warstwę ReLU oraz dropout, aby następnie trafić do drugiej warstwy liniowej. Na końcu zastosowana jest funkcja logarytmiczna, która przekształca wyjście w prawdopodobieństwa klas.

Posiadając poprawnie zaimplementowany klasyfikator można przejść do implementacji funkcji odpowiedzialnej za trening modelu.



```
train_loss = []
train_counter = []
test_loss = []
test_counter = [i*len(train_loader.dataset) for i in range(n_epochs + 1)]

1 usage  Jakub Jucha *
def training_single_epoch(epoch, model, optimizer, train_loader, loss_fcn):
    model.train()
    for batch_ind, (data, target) in enumerate(train_loader):
        optimizer.zero_grad()
        output = model(data)
        loss = F.nll_loss(output, target)
        loss.backward()

        optimizer.step()

        if batch_ind % log_interval == 0:
            print("Train Epoch: {}[{}/{} ({:.0f}%)]\tLoss: {:.6f}]".format(
                *args: epoch, batch_ind * len(data), len(train_loader.dataset),
                100. * batch_ind / len(train_loader), loss.item()
            ))
```

Listing 26. Funkcja odpowiedzialna za trening modelu.

Na samym początku należy zainicjalizować listy, których zadaniem będzie przechowywanie wartości strat oraz licznik podczas treningu i testowania. Listy te pomagają śledzić, jak dobrze model się uczy w trakcie procesu trenowania oraz jakie są jego osiągi na zbiorze testowym.

Zdefiniowana funkcja odpowiada za przeprowadzenie jednej epoki treningu. Pierwszym krokiem jest ustawienie modelu w tryb treningu poprzez wywołanie „model.train()”, co jest ważne, ponieważ niektóre warstwy, takie jak dropout, działają inaczej podczas treningu niż podczas ewaluacji.

Wewnątrz pętli, która iteruje przez batch’e danych treningowych, kod zeruje gradient z poprzedniej iteracji („optimizer.zero_grad()”). Następnie danę są przepuszczane przez model („output = model(data)”), a funkcja strat oblicza różnicę między przewidywaniami a rzeczywistymi etykietami („loss = F.nll_loss(output, target)”). Strata ta jest następnie propagowana wstecz, aby obliczyć gradienty („loss.backward()”), a optymalizator aktualizuje parametry modelu na podstawie tych gradientów („optimizer.step()”).

Dodatkowo, co pewien interwał określony przez „log_interval”, kod wyświetla informacje o postępie treningu, takie jak numer epoki, liczba przetworzonych przykładów, procentowy

postęp oraz aktualne wartości straty. Pomaga to monitorować, jak dobrze model się uczy i czy proces treningu przebiega prawidłowo.

Aby teraz przetestować sieć neuronową zdefiniujmy następującą funkcję

```
def test():
    model.eval()
    test_loss = 0
    correct = 0
    test_losses = []

    with torch.no_grad():
        for data, target in test_loader:
            output = model(data)
            test_loss += F.nll_loss(output, target).item()

            pred = output.data.max(dim = 1, keepdim = True)[1]

            correct += pred.eq(target.data.view_as(pred)).sum()

    test_loss /= len(test_loader.dataset)
    test_losses.append(test_loss)

    print("\nTest set: Acg. loss: {:.4f}, Accuracy: {}/{} ({:.0f}%)\n".format(
        *args: test_loss, correct, len(test_loader.dataset),
        100. * correct / len(test_loader.dataset)
    ))

    accuracy = 100. * correct / len(test_loader.dataset)
    return test_loss, accuracy
```

Listing 27. Funkcja testująca sieć neuronową

Na początku funkcja „test()” ustawia model w tryb ewaluacji za pomocą „model.eval()”, co deaktywuje niektóre specyficzne dla treningu operacje, takie tak dropout. Następnie inicjalizowane są zmienne „test_loss” do sumowania strat, „correct” do liczenia poprawnych predykcji oraz „test_losses” do przechowywania wartości strat dla każdego batcha.

Blok „with torch.no_grad():” wyłącza obliczanie gradientów, co przyspiesza obliczanie i zmniejsza zużycie pamięci, ponieważ gradienty nie są potrzebne podczas testowania. Wewnątrz tego bloku, pętla iteruje przez batch’e danych testowych, gdzie każdy batch składa się z par „data” i „target”.

Dane są przepuszczane przez model, a wynikowe predykcje są porównywane z rzeczywistymi etykietami. Strata dla każdego batcha jest obliczona za pomocą funkcji „F.nll_loss(output, target)” i sumowana w zmiennej „test_loss”. Predykcje są następnie przekształcane w przewidywane klasy („pred”), a liczba poprawnych predykcji jest sumowana w zmiennej „correct”.

Po przetworzeniu wszystkich batch’y całkowita strata jest uśredniana przez podzielenie jej przez liczbę przykładów w zbiorze testowym. Wynikowa średnia strata jest dostosowana do listy „test_losses”

Na końcu funkcja drukuje średnią stratę i dokładność modelu, która jest obliczana jako procent poprawnych predykcji w stosunku do całkowitej liczby przykładów testowych. Dokładność i strata są również zwracane przez funkcję, co umożliwia dalsze wykorzystanie tych wartości do analizy wyników modelu.

```
for channels3 in out_channels3:
    for channels2 in out_channels2:
        model = Network(channels3, channels2)
        optimizer = torch.optim.Adam(params=model.parameters(), lr=learning_rate)

        print(f"\nTesting model with out_channels1={channels3} and out_channels2={channels2}")
        test_accuracy = []
        for epoch in range(1, n_epochs + 1):
            training_single_epoch(epoch, model=model, optimizer=optimizer, train_loader=train_loader, loss_fcn=F.nll_loss)
            test_loss, accuracy = test()  # pobranie dokładności i straty testowej
            test_accuracy.append(accuracy)  # zapisanie dokładności w liście
            print(f"Test Loss: {test_loss}, Accuracy: {accuracy}%")
        results.append((channels3, channels2, test_accuracy))
```

Listing 28. Testowanie modelu

Dla każdej kombinacji parametrów, tworzony jest nowy model i optymalizator Adam z ustawioną wartością współczynnika uczenia („learning_rate”). Następnie, rozpoczyna się trenowanie modelu dla zadanej liczby epok („n_epochs”), gdzie w każdej epoce wywoływana jest funkcja „training_single_epoch” do trenowania modelu oraz „test” do oceny jego wydajności na zbiorze testowym. Wynikiem testowania jest strata i dokładność, które są drukowane i zapisywane dla danej konfiguracji modelu.

6. Podsumowanie

Niniejszy artykuł skupia się na zastosowaniu głębokich sieci neuronowych do rozpoznawania ręcznie pisanych cyfr ze zbioru danych MNIST, z naciskiem na sieci

konwulacyjne (CNN). CNN są efektywne w analizie obrazów dzięki zdolności automatycznego wykrywania istotnych cech w różnych regionach obrazu. Przedstawiono typową strukturę sieci konwulacyjnej, obejmującą warstwy konwulacyjne, poolingowe i w pełni połączone. Omówiono wyzwania związane z treningiem głębokich sieci, takie jak problem zanikającego gradientu, który utrudnia uczenie się głębokich warstw. W artykule zaprezentowano również techniki, które pomagają w rozwiązywaniu tego problemu, w tym normalizacje batchową i funkcję aktywacji typu ReLU. Zwrócono również uwagę na problem przetrenowania, gdzie model zbytnio dopasowuje się do danych treningowych, co pogarsza jego zdolność do generalizacji na nowych danych. Aby przeciwdziałać przetrenowaniu, zastosowano metodę regularyzacji dropoutu, która losowo deaktywuje neurony podczas treningu.

Literatura

1. Tabor J., Śmieja M., Struski Ł., Spurek P., Wołczyk M., „Głębokie uczenie wprowadzenie”, wydawnictwo Helion

Źródła internetowe

1. https://sciagaprogramisty.blogspot.com/2018/01/konwolucja-wstep-do-neuronowych-sieci.html (dostęp 18.05.2024)
2. https://www.analyticsvidhya.com/blog/2021/08/all-you-need-to-know-about-skip-connections/ (dostęp 18.05.2024)
3. https://miroslawmamczur.pl/jak-dzialaja-konwolucyjne-sieci-neuronowe-cnn/ (dostęp 18.05.2024)
4. https://medium.com/@a01642207/convolutional-neural-networks-for-image-classification-461306f4e7f9 (dostęp 18.05.2024)

Magdalena Matuła, Katarzyna Maternia, Aleksandra Sawicka, Aleksandra Rokita, Wiktor Kuczek
SKNI „Kod”

dr inż. Bartosz Trybus
Opiekun Koła Naukowego

Zrozumienie uczenia maszynowego: kluczowa rola algorytmu wstecznej propagacji w trenowaniu sieci neuronowych

Streszczenie

Uczenie maszynowe to dynamicznie rozwijająca się dziedzina sztucznej inteligencji, umożliwiająca tworzenie modeli matematycznych, które samodzielnie uczą się i doskonalą swoje predykcje. Kluczowym narzędziem w tym procesie jest algorytm wstecznej propagacji, pozwalający na optymalizację wag i biasów sieci neuronowych. Artykuł omawia podstawowe pojęcia uczenia maszynowego, historię jego rozwoju, oraz szczegółowo przedstawia działanie i znaczenie sieci neuronowych. Przedstawiono również korzyści i wyzwania związane z algorytmem wstecznej propagacji, a także jego szerokie zastosowania, takie jak rozpoznawanie obrazów, przetwarzanie języka naturalnego i robotyka. Omówione zostały również metody przyspieszające proces uczenia: metodę momentu oraz adaptacyjny dobór współczynnika uczenia. Mimo pewnych ograniczeń, algorytm ten pozostaje fundamentalnym narzędziem w rozwoju nowoczesnych technologii.

Słowa kluczowe: sztuczna inteligencja, uczenie maszynowe, wsteczna propagacja, metoda momentum, adaptacyjny współczynnik uczenia

1. Wprowadzenie

W ciągu ostatnich dekad, uczenie maszynowe stało się jednym z najważniejszych i najszybciej rozwijających się obszarów w dziedzinie sztucznej inteligencji. Stanowi fundament dla licznych innowacji technologicznych, od rozpoznawania mowy i obrazu, po autonomiczne pojazdy i systemy rekomendacji. Kluczowym elementem tego sukcesu jest zdolność modeli matematycznych do samodzielnego uczenia się i doskonalenia swoich predykcji na podstawie danych, bez konieczności bezpośredniego programowania. W centrum tego procesu znajduje się algorytm wstecznej propagacji, który umożliwia trenowanie sieci neuronowych poprzez optymalizację wag i biasów, minimalizując błędy predykcji.

W niniejszym artykule przyjrzymy się roli algorytmu wstecznej propagacji w trenowaniu sieci neuronowych oraz jego znaczeniu dla rozwoju uczenia maszynowego. Omówimy podstawowe pojęcia związane z uczeniem maszynowym, historię rozwoju tej dziedziny, a także szczegółowo przedstawimy działanie sieci neuronowych i algorytmu wstecznej propagacji. Ponadto, zanalizujemy korzyści i wyzwania związane z tym algorytmem, a także zastosowania,

które zmieniają nasze codzienne życie. Artykuł ma na celu dostarczenie kompleksowego wglądu w kluczowe aspekty uczenia maszynowego oraz zrozumienie, jak algorytm wstecznej propagacji przyczynia się do jego skuteczności i rozwoju.

2. Uczenie maszynowe – czym jest ?

Uczenie maszynowe to dziedzina sztucznej inteligencji, która polega na tworzeniu modeli matematycznych zdolnych do uczenia się i dokonywania prognoz lub podejmowania decyzji na podstawie danych, bez konieczności bezpośredniego programowania. Kluczową cechą uczenia maszynowego jest zdolność systemów do ulepszania swoich predykcji wraz z napływem nowych danych. Im więcej danych otrzymają, tym dokładniejsze będą ich przewidywania. Systemy te nie są jawnie programowane, lecz same uczą się rozpoznawać wzorce i korelacje w danych, aby podejmować decyzje lub formułować prognozy. Wyróżniamy trzy główne typy uczenia: nadzorowane, nienadzorowane i przez wzmacnianie. W uczeniu nadzorowanym maszyna uczy się na podstawie danych które zostały już „opisane" przez człowieka. Na przykład, jeśli chcemy nauczyć maszynę rozpoznawania zdjęć psów, dostarczamy jej zestaw zdjęć, z których część jest oznaczona jako „pies”. W uczeniu nienadzorowanym system samodzielnie identyfikuje wzorce w danych, które nie są wcześniej oznaczone. Ten rodzaj uczenia jest często stosowany w analizie klastrów, wykrywaniu anomalii oraz redukcji wymiarowości danych. W uczeniu przez wzmacnianie maszyna zdobywa wiedzę na podstawie doświadczeń, podejmując działania i otrzymując za nie nagrody lub kary. Jest to fundamentalna metoda w tworzeniu systemów takich jak autonomiczne pojazdy i gry komputerowe.

3. Historia uczenia maszynowego

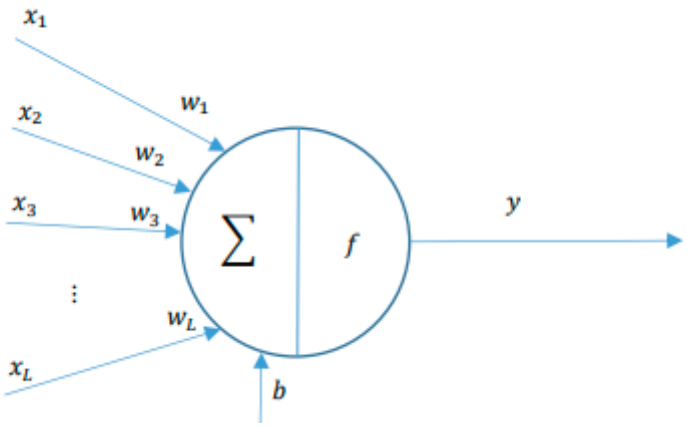
Teoretyczne postawy uczenia maszynowego zostały ukształtowane przez Alana Turinga w 1950 roku. Opublikował on wtedy prace, w której rozważał czy maszyny mogą myśleć, praca ta zapoczątkowała test Turinga. W 1952 roku Arthur Samuel opracował samo uczący program, który grał w warcaby oraz zdefiniował uczenie maszynowe jako zdolność komputerów do nauki bez bycia jawnie zaprogramowanym. W latach 90 zaczęto stosować uczenie maszynowe w praktyce m.in. dzięki rozwojowi teorii oraz narzędzi takich jak maszyna wektorów nośnych. W 1997 roku program Deep Blue firmy IBM pokonał mistrza świata w szachy, Garriego Kasparowa, co dowiodło rosnącej mocy obliczeniowej i skuteczności uczenia maszynowego. Na początku XXI wieku pojawił się termin "głębokie uczenie", który zyskał na popularności dzięki pracom Geoffreya Hintona i jego zespołu. Opracowali oni bardziej efektywne techniki

treningu głębokich sieci neuronowych, co przyczyniło się do szybkiego wzrostu znaczenia głębokiego uczenia. Dominację tej metody w szkoleniu modeli potwierdziło zwycięstwo modelu AlexNet w konkursie ImageNet w 2012 roku. Warto również podkreślić, że rewolucja AI, która miała miejsce w 2022 roku, byłaby niemożliwa bez rozwoju różnych metod uczenia maszynowego. Dzięki tym technikom możemy dziś korzystać z zaawansowanych narzędzi, takich jak ChatGPT, Midjourney czy Google Gemini.

4. Rola sieci neuronowej w uczeniu maszynowym

Uczenie maszynowe często wykorzystuje sieci neuronowe, które są zaawansowanymi modelami obliczeniowymi inspirowanymi biologicznymi sieciami neuronowymi w mózgu. Sieci neuronowe składają się z wielu połączonych jednostek obliczeniowych, zwanych neuronami, które są zorganizowane w warstwy. Każdy neuron w sieci odbiera sygnały wejściowe, przetwarza je i przekazuje dalej, co pozwala sieci na uczenie się złożonych wzorców i relacji w danych.

Neuron:



Rysunek 1 Model neuron z strony [1]

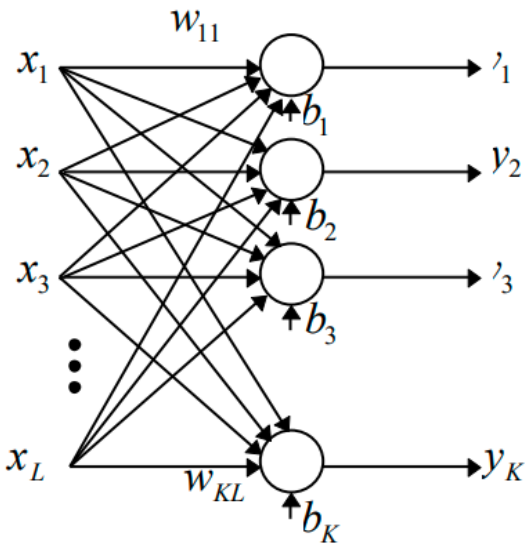
x- sygnały wejściowe

w- wagi czyli współczynników przy wejściach neuronów

b-bias czyli wartości przesunięcia, które dodaje się do sumy ważonej wejść

y- sygnały wyjściowe

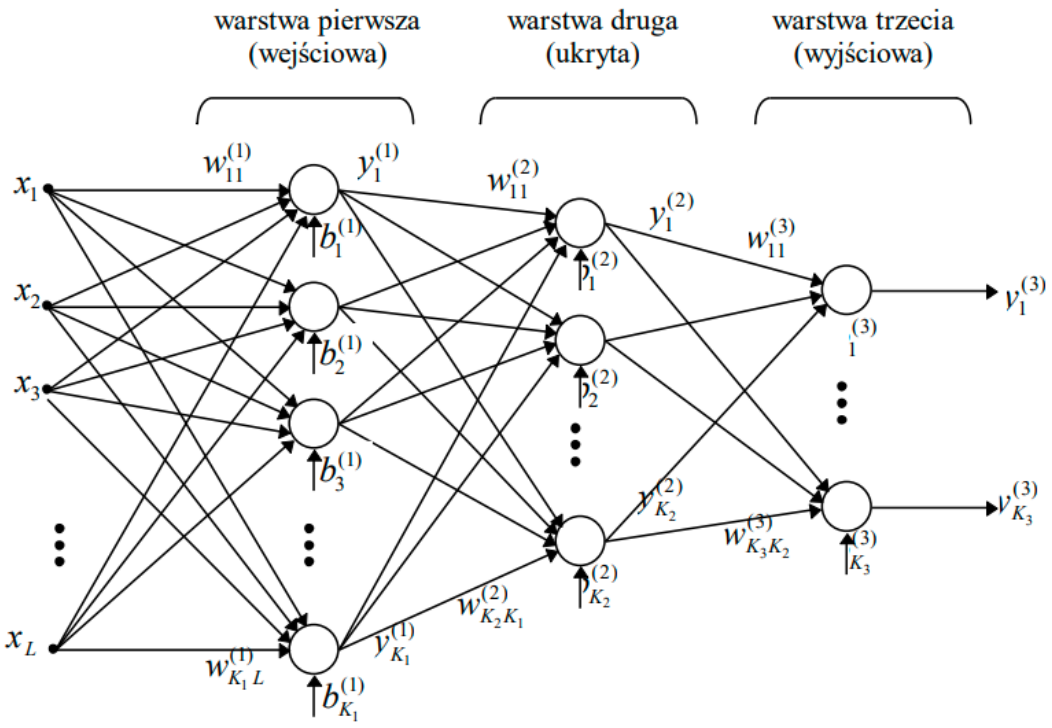
Siec neuronowa jednowarstwowa:



Rysunek 2 Siec jednowarstwowa z strony [2]

Neurony w tej sieci są rozmieszczone w jednej warstwie. Sygnały przepływają jednokierunkowo, $x = [x_1, x_2, \dots, x_L]^T$ zaczynając od wejścia i kończąc na wyjściu. Każdy neuron otrzymuje wektor sygnałów wejściowych, co oznacza, że każde wejście jest połączone z każdym neuronem. Ponadto każdy neuron ma swój własny zestaw wag.

Sieć neuronowa wielowarstwowa:



Rysunek 3 Sieć neuronowa wielowarstwowa ze strony [3]

Sieć ta składa się z następujących warstw:

- Warstwa wejściowa: Na tę warstwę podawane są sygnały wejściowe.
- Warstwa ukryta: Ta warstwa przetwarza dane wejściowe, wydobywając istotne cechy i wzorce. Warstwa ta jest odpowiedzialna za przechwytywanie nieliniowych zależności w danych.
- Warstwa wyjściowa: Przetwarza sygnały pochodzące z warstwy ukrytej i generuje końcowy wynik.

W tej strukturze występują połączenia typu "każdy z każdym" pomiędzy neuronami sąsiednich warstw, co oznacza, że każdy neuron z jednej warstwy jest połączony z każdym neuronem w warstwie następnej. Taka konfiguracja pozwala na efektywne przetwarzanie i propagację informacji przez sieć.

Działanie algorytmu wstecznej propagacji

Jedną z metod uczenia sieci neuronowych jest propagacja wsteczna. Umożliwia ona optymalizację wag i biasów neuronów w taki sposób, aby minimalizować błąd pomiędzy przewidywanym a rzeczywistym wyjściem sieci.

Działanie algorytmu można podzielić na kilka kroków:

1. Propagacja w przód

Dane wejściowe przechodzą przez każdą kolejną warstwę, gdzie każdy neuron otrzymuje sygnał wejściowy **x** z poprzedniej warstwy lub bezpośrednio z danych wejściowych. Następnie sygnały są mnożone przez odpowiadające im wagi **w**, co tworzy jedno skumulowane pobudzenie dla neuronu. Po zsumowaniu ważonych sygnałów, dodawana jest do nich wartość progu aktywacji, czyli bias **b**. Próg aktywacji stanowi decyzyjną wartość, która determinuje aktywację neuronu. Kolejnym krokiem jest zastosowanie funkcji aktywacji **f**, w tym przypadku funkcji tangens hiperboliczny (tansig) z biblioteki nnet, do obliczenia wyniku pobudzenia. Wynik ten jest uznawany za wyjście neuronu i przekazywany do kolejnej warstwy neuronów.

2. Obliczanie błędu

Otrzymany wynik na wyjściu sieci neuronowej jest porównywany z wartością oczekiwaną. Aby oszacować, jak bardzo różnią się uzyskane wyniki od tych, których oczekujemy, używa się funkcji kosztu. Celem tej funkcji jest wyrażenie odchylenia jako pojedynczej wartości liczbowej, którą minimalizujemy podczas treningu. W moim kodzie jako funkcji kosztu używam metody SSE (Sum of Squared Errors), czyli sumy kwadratów różnic między rzeczywistymi wartościami a przewidywanymi przez model.

3. Propagacja wsteczna

Błąd jest propagowany wstecz przez kolejne warstwy sieci neuronowej, co oznacza, że jest on cofany w procesie treningu. Podczas tego cofania każda warstwa oblicza swój wkład do błędu poprzez obliczanie gradientów funkcji kosztu względem swoich wag i propagowanie tych gradientów wstecz.

Gradienty są pochodnymi cząstkowymi funkcji wielu zmiennych, a w kontekście sieci neuronowych obliczane są one względem parametrów modelu, takich jak wagi w poszczególnych warstwach sieci. Ogólnie mówiąc, gradienty wskazują kierunek najszybszego wzrostu lub spadku funkcji w przestrzeni parametrów modelu. Dzięki nim możliwe jest określenie, jak bardzo zmiana każdego parametru wpływa na zmianę wartości funkcji kosztu. Jest to kluczowe dla optymalizacji modelu poprzez dostosowywanie wag w celu minimalizacji błędu. Odpowiednie składniki gradientu funkcji celu względem wag neuronów poszczególnych warstw otrzymuje się przez różniczkowanie zależności.

4. Aktualizacja wag

W celu minimalizacji błędu wagi są modyfikowane. W standardowej metodzie wstecznej propagacji błędu stosuje się algorytm gradientu prostego. W tej metodzie, aktualizacja wag (oznaczona jako $\Delta wij(t)$) następuje na podstawie gradientu funkcji kosztu, który wskazuje kierunek największego spadku funkcji błędu.

5. Iteracja

Proces ten jest powtarzany wielokrotnie, aż do uzyskania wyników najbardziej zbliżonych do oczekiwanych. Zazwyczaj używa się w tym celu wielu epok treningowych, gdzie w każdej epoce sieć jest uczona na wszystkich dostępnych danych treningowych.

5. Przyspieszanie procesu uczenia

Metoda gradientowa, która została przedstawiona powyżej, może być bardzo czasochłonna. Istnieje wiele technik, które pozwalają na znaczne przyspieszenie procesu uczenia, jednak skupie się na dwóch najpopularniejszych metody. Pierwsza metoda polega na dodaniu do wzoru na korektę wag dodatkowego składnika zwanego "momentum", który mierzy "bezwładność" zmiany wag. Druga metoda polega na adaptacyjnym dostosowywaniu współczynnika uczenia η\etaη w zależności od tendencji zmiany miary błędu uczenia.

6. Metoda momentum

Klasyczny algorytm wstecznej propagacji aktualizuje wagi sieci neuronowej wyłącznie na podstawie bieżącego gradientu funkcji błędu. Metoda momentum wprowadza dodatkowy

czynnik - tzw. momentum, który uwzględnia poprzednie zmiany wag.. Współczynnik momentum mc jest zazwyczaj ustalany na wartość między 0 a 1, gdzie większe wartości oznaczają większy wpływ poprzednich kroków na aktualizację wag. Standardowo stosuje się wartość około 0.95.

Wprowadzenie momentum przyspiesza proces uczenia, ponieważ pomaga uniknąć zatrzymywania się w lokalnych minimach funkcji kosztu poprzez utrzymanie momentum w kierunku globalnego minimum. Dzięki temu algorytm jest mniej podatny na oscylacje i może szybciej zbliżyć się do optymalnego rozwiązania.

7. Adaptacyjny współczynnik uczenia

Adaptacyjny współczynnik uczenia (adaptive learning rate) to modyfikacja algorytmu wstecznej propagacji błędu, w której współczynnik uczenia η nie jest stały, lecz dostosowywany dynamicznie w trakcie procesu uczenia sieci neuronowej. Kluczową motywacją jest fakt, że stała wartość współczynnika uczenia może być nieoptymalna na różnych etapach treningu. Zbyt mała wartość spowalnia proces uczenia, a zbyt duża może prowadzić do niestabilności i oscylacji wokół minimum. Dlatego w metodzie adaptacyjnego współczynnika uczenia η jest modyfikowany w każdej iteracji w celu przyspieszenia zbieżności i poprawy stabilności procesu uczenia.

Mechanizmy adaptacji współczynnika uczenia:

- Minimalizacja funkcji błędu względem η - η jest dobierany tak, aby minimalizować funkcję błędu sieci w danej iteracji,
- Reguła heurystyczna - η jest zwiększany, gdy kolejne kroki zmniejszają błąd, lub zmniejszany w przeciwnym przypadku,
- Metoda Chana i Falsidiego - η jest zwiększany, gdy kolejne zmiany wag mają ten sam znak (zgodny kierunek), lub zmniejszany przy zmianach znaku.

Adaptacyjny współczynnik uczenia pozwala na automatyczną regulację wielkości kroków w przestrzeni wag, co przyspiesza zbieżność algorytmu i zwiększa jego stabilność. Stanowi ważne udoskonalenie klasycznej wstecznej propagacji.

Ponadto, w połączeniu z momentem, adaptacyjny współczynnik uczenia umożliwia efektywne uczenie głębokich sieci neuronowych na dużych zbiorach danych, co jest kluczowe w uczeniu głębokim.

8. Zastosowanie algorytmu wstecznej propagacji

Algorytm wstecznej propagacji błędu (backpropagation) jest fundamentalnym algorytmem uczenia nadzorowanego stosowanym w sztucznych sieciach neuronowych. Odgrywa kluczową rolę w wielu zastosowaniach uczenia maszynowego i sztucznej inteligencji.

Główne zastosowania algorytmu wstecznej propagacji:

- Rozpoznawanie obrazów

Sieci neuronowe z propagacją wsteczną są szeroko wykorzystywane w systemach rozpoznawania obrazów, takich jak optyczne rozpoznawanie znaków (OCR), wykrywanie twarzy czy klasyfikacja obiektów na obrazach.

- Przetwarzanie języka naturalnego

Propagacja wsteczna znajduje zastosowanie w zadaniach przetwarzania języka, takich jak tłumaczenie maszynowe, analiza sentymentu, generowanie tekstu czy systemy dialogowe.

- Rozpoznawanie mowy

Algorytmy propagacji wstecznej są wykorzystywane w systemach rozpoznawania mowy do konwersji sygnału audio na tekst.

- Prognozowanie i klasyfikacja

Sieci neuronowe z propagacją wsteczną są powszechnie stosowane do prognozowania szeregów czasowych, klasyfikacji danych oraz wykrywania anomalii w różnych dziedzinach, np. finansach, medycynie, telekomunikacji.

- Sterowanie i robotyka

Propagacja wsteczna znajduje zastosowanie w systemach sterowania i robotyce, umożliwiając uczenie się optymalnych strategii sterowania na podstawie danych.

- Rekomendacje i personalizacja

Sieci neuronowe z propagacją wsteczną są wykorzystywane w systemach rekomendacji produktów, treści multimedialnych oraz personalizacji reklam i interfejsów użytkownika.

9. Wady i zalety algorytmu wstecznej propagacji

Zalety algorytmu wstecznej propagacji:

- Efektywność uczenia - umożliwia skuteczne uczenie sieci neuronowych o wielu warstwach ukrytych, co zwiększa ich zdolność modelowania złożonych zależności,
- Uniwersalność - może być stosowany do różnych zadań, takich jak klasyfikacja, regresja, przetwarzanie sygnałów itp.,
- Prostota implementacji - algorytm jest stosunkowo prosty w realizacji i zrozumieniu,

- Globalna optymalizacja - dąży do znalezienia globalnego minimum funkcji błędu, a nie tylko lokalnego,

Wady algorytmu wstecznej propagacji:

- Zbieżność do minimum lokalnego - istnieje ryzyko utknięcia w minimum lokalnym funkcji błędu, co może być ograniczane przez odpowiednie inicjalizacje wag i metody jak momentum,
- Czas uczenia - dla dużych sieci i zbiorów danych proces uczenia może być czasochłonny,
- Konieczność obliczeń wstecznych - wymaga przechowywania danych pośrednich i wykonywania obliczeń wstecz przez całą sieć, co zwiększa zapotrzebowanie na pamięć,
- Podatność na znikanie/eksplozję gradientu - problemy te mogą utrudniać uczenie głębokich sieci i są częściowo rozwiązywane przez modyfikacje jak LSTM,
- Dobór hiperparametrów - skuteczność uczenia zależy od odpowiedniego doboru hiperparametrów, takich jak współczynnik uczenia, momentum itp.

Pomimo tych wad, algorytm wstecznej propagacji pozostaje fundamentalnym i niezwykle ważnym narzędziem uczenia maszynowego, umożliwiającym rozwój głębokich sieci neuronowych i sztucznej inteligencji. Jego modyfikacje i udoskonalenia pozwalają na przezwyciężanie niektórych ograniczeń.

10. Podsumowanie

Algorytm wstecznej propagacji odgrywa kluczową rolę w trenowaniu sieci neuronowych, będąc fundamentalnym narzędziem w dziedzinie uczenia maszynowego. Uczenie maszynowe, będące gałęzią sztucznej inteligencji, polega na tworzeniu modeli matematycznych, które uczą się i dokonują prognoz na podstawie danych bez bezpośredniego programowania. W ramach tego procesu sieci neuronowe, inspirowane biologicznymi sieciami neuronowymi, umożliwiają efektywne przetwarzanie i analizę danych dzięki skomplikowanej strukturze połączeń między neuronami.

Historia uczenia maszynowego sięga prac Alana Turinga i Arthura Samuela, którzy wprowadzili podstawowe koncepcje i stworzyli pierwsze samo-uczące się programy. W kolejnych dekadach rozwój tej dziedziny przyspieszył dzięki rosnącej mocy obliczeniowej, pojawieniu się nowych algorytmów oraz sukcesom takim jak zwycięstwo Deep Blue w szachach i rozwój głębokiego uczenia.

Algorytm wstecznej propagacji umożliwia optymalizację wag i biasów w sieci neuronowej, minimalizując błąd między przewidywanym a rzeczywistym wynikiem. Proces ten składa się z kilku kroków: propagacji w przód, obliczania błędu, propagacji wstecznej, aktualizacji wag i iteracji. W celu przyspieszenia uczenia stosuje się techniki takie jak metoda momentum i adaptacyjny współczynnik uczenia.

Wsteczna propagacja ma szerokie zastosowanie, od rozpoznawania obrazów i mowy, przez przetwarzanie języka naturalnego, aż po sterowanie i robotykę. Algorytm ten umożliwia skuteczne modelowanie złożonych zależności w danych, choć ma również swoje wady, takie jak ryzyko zbieżności do minimum lokalnego, czasochłonność procesu uczenia oraz problemy znikania i eksplozji gradientu.

Pomimo tych wyzwań, wsteczna propagacja pozostaje nieodzownym elementem rozwoju sztucznej inteligencji, umożliwiając tworzenie zaawansowanych systemów i narzędzi, które rewolucjonizują różne dziedziny życia.

Źródła internetowe:

1. http://materialy.prz-rzeszow.pl/pracownik/pliki/34/%C4%86WICZENIE%204.pdf (dostęp: 25.05.2024) [1]
2. http://materialy.prz-rzeszow.pl/pracownik/pliki/34/%C4%86WICZENIE%206.pdf (dostęp: 25.05.2024) [2]
3. http://materialy.prz-rzeszow.pl/pracownik/pliki/34/%C4%86WICZENIE%207.pdf (dostęp: 25.05.2024) [3]
4. http://materialy.prz-rzeszow.pl/pracownik/pliki/34/%C4%86WICZENIE%208.pdf (dostęp: 25.05.2024) [4]
5. https://www.sas.com/pl_pl/insights/analytics/neural-networks.html (dostęp: 26.05.2024) [5]
6. https://predictivesolutions.pl/sieci-neuronowe (dostęp: 26.05.2024) [6]
7. https://www.deeptechnology.ai/czym-jest-uczenie-maszynowe/ (dostęp: 27.05.2024) [7]
8. https://miroslawmamczur.pl/czym-jest-uczenie-maszynowe-i-jakie-sa-rodzaje/ (dostęp: 27.05.2024) [8]
9. https://www.guru99.com/pl/backpropogation-neural-network.html (dostęp: 27.05.2024) [9]

Sebastian Cwynar, Jakub Jucha, Maciej Karczmarz, Hubert Kraus, Adam Krawczyk
SKNI „Kod”

dr inż. Bartosz Trybus
Opiekun Koła Naukowego

Rodzaje ataków DDOS (Distributed Denial of Service) i strategie obronne

Artykuł opisuje działanie ataków DDoS (Distributed Denial of Service) oraz strategie obronne, mające na celu zminimalizowanie podatności na ten rodzaj ataków. Ataki DDoS stanowią istotny temat w dziedzinie cyberbezpieczeństwa od którego wiele osób zaczyna swoje zainteresowanie tą dziedziną. Artykuł ten ma na celu przybliżenie tego tematu gdzie zostanie opisane czym jest atak DDoS oraz proces jak zostaje ten atak przeprowadzony. W artykule zostanie także poruszony temat botnetów i sposobu w jaki biorą udział w ataku DDoS. Kolejnym ważnym elementem jest przedstawienie historii ataków tego rodzaju, co pozwala na zrozumienie ich ewolucji i skali problemu. Ostatnią częścią artykułu będzie przeanalizowanie sytuacji potencjalnej ofiary ataku oraz przedstawienie różnorodnych strategii obronnych które można zastosować przed atakiem, w jego trakcie oraz po jego zakończeniu. Badanie opierało się na analizie literatury naukowej oraz źródeł internetowych.

Słowa kluczowe: DDoS, Cyberbezpieczeństwo , Botnet.

1. Wprowadzenie

DDoS jest to kolejna forma ataku DoS, gdzie DoS jest atakiem przeprowadzanym z pojedynczego źródła, które przesyła ogromną ilość żądań do ofiary, w celu przeciążenia jej zasobów i uniemożliwienia obsługi normalnych żądań. DDoS to skoordynowany atak na określone cele z wielu źródeł – zainfekowanych hostów zwanych botami lub zombie połączonymi w botnety, co czyni go bardziej niszczycielskim i trudniejszym do powstrzymania. Możemy rozróżnić kilka rodzajów ataków takich jak:

- ataki wolumetryczne,
- ataki na poziomie protokołów,
- ataki na poziomie aplikacji.

Liczba ataków DDoS stale rośnie, według przewidywań Cisco liczba ataków miała z 7,9 miliona w 2018 roku wzrosnąć do 15,4 miliona w 2023. Dlatego też zmniejszanie podatności na ataki DDoS poprzez strategie obronne staje się coraz ważniejsze. Artykuł przedstawi zasadę działania botnetu oraz metody jego wykrywania która jest jedną z strategii obrony przed atakami DDoS a następnie zostanie opisany DDoS i wybrane jego rodzaje oraz strategie obrony

przed tymi atakami co pozwoli zminimalizować wciąż rosnące ryzyku stania się ofiarą ataku DDoS.

2. Botnet

Botnet jest to sieć zainfekowanych hostów, na których działa złośliwe oprogramowanie które włącza zainfekowane urządzenie do botnetu przez co staje się tzw. botem lub zombie. Botmaster może wydawać polecenia botom w botnecie np. wskazując nowe cele do ataku DDoS i każdy bot wykonuje te polecenia. Właściciele zainfekowanych urządzeń też są pośrednio ofiarami w atakach DDoS ponieważ w tym czasie z powodu że urządzenie dostało polecenie wykonania ataku to zasoby danego urządzenia są wykorzystywane do tego celu co powoduje znaczący spadek wydajności danego urządzenia. Nie tylko komputery mogą zostać zainfekowane w ten sposób, narażone są także urządzenia IoT.

Jedną z najważniejszych cech botnetów jest to że zapewniają one atakującemu anonimowość poprzez wykorzystanie wielopoziomowej architektury dowodzenia. Boty znajdujące się w jednym botnecie są rozproszone po całym świecie a różnice pomiędzy strefami czasowymi i językami dodatkowo utrudniają śledzenie działań botnetów. Cykl życia botnetów można podzielić na pięć faz :

- początkową infekcję,
- wtórną iniekcję,
- połączenie,
- dowodzenie przez botmastera,
- aktualizacja i utrzymanie.

W pierwszej fazie atakujący znajduje luki w urządzeniach a następnie infekuje je za pomocą różnych metod. Następnie na zainfekowanym urządzeniu w fazie wtórnej iniekcji wykonuje się skrypt używający FTP , HTTP lub P2P który instaluje bota na danym urządzeniu i od tego momentu z każdym uruchomieniem urządzenia zombie uruchamia się także bot. W fazie połączenia program bota ustawia połączącenie C&C i łączy zombie z serwerem poprzez który może otrzymywać polecenia od botmastera. Po nawiązaniu połączenia rozpoczyna się czwarta faza w której botmaster może wydawać polecenia ataku jednocześnie wszystkim botom w jego botnecie. Ostatnią fazą jest aktualizowanie i utrzymanie botnetu w której boty otrzymują co jakiś czas polecenie zaktualizowanego pliku binarnego co jest spowodowane tym że botnet musi się dostosowywać do zmieniających się technik wykrywania w celu pozostania w ukryciu. Czasem aktualizacje to powodują przeniesienie botów do innego serwera C&C co nazywane jest migracją. Botmasterzy wykorzystują Dynamiczny DNS (DDNS) do zmiany lokalizacji serwerów, co pozwala na utrzymanie botnetu w sposób niewykrywalny. Gdy działanie serwera C&C zostanie zakłócone, botmaster

może błyskawicznie skonfigurować nowy serwer, korzystając z tego samego DDNS. Dzięki krótkim czasom TTL, zmiany adresów IP są szybko propagowane, co umożliwia botom natychmiastową migrację do nowego serwera C&C i kontynuowanie swojej działalności. Istnieje wiele botnetów które osiągnęły dużą liczbę botów, jednym z nich jest botnet Mirai który w znaczącym stopniu składa się z urządzeń IoT i w 2016 roku osiągnął 600 000 zainfekowanych urządzeń włączonych do tego botnetu. Jedną ze strategii obrony przez atakami DDoS jest zapobieganie atakom poprzez wykrywanie botnetów, istnieje kilka sposób na ich wykrywanie przez co istnieje szansa na ich wyeliminowanie gdy zostaną już wykryte

3. Metody wykrywania botnetów

Jedną z strategii obrony przed atakami DDoS jest wykrywanie botnetów co daje możliwość obrony przed atakami DDoS z ich strony.

Oto znane metody wykrywania botnetów:

5. Wykrywanie oparte na sygnaturach – Jest to technika wykrywania botnetów poprzez sygnatury i zachowania istniejących botnetów które zostały już wcześniej odkryte. Technika ta sprawdza się tylko w przypadku znanych botnetów, w przypadku nowych lub wcześniej nieodkrytych nie jest skuteczna
6. Wykrywanie oparte na anomaliach – Jest to technika identyfikująca botnety na podstawie wysokiej latencji, nietypowego zachowania sieci, komunikacji na nietypowych portach oraz dużego ruchu w sieci co może sugerować na obecność botów w sprawdzanej sieci. Jednym z narzędzi wykorzystujących tą technikę jest Botsniffer który dodatkowo identyfikuje boty na podstawie tego że są one ze sobą silnie zsynchronizowane jeśli chodzi o komunikację co pozwala na wykrycie botnetu z dużą skutecznością jednocześnie minimalizując możliwość wystąpienia fałszywych detekcji.
7. Wykrywanie oparte na DNS – Boty komunikują się z serwerami C&C poprzez zapytania DNS co umożliwia ich wykrywania na podstawie analizy ruchu DNS i wykrywania nietypowych wzorców. Jednym ze sposobów jest wykrywanie skoncentrowanych zapytań DDNS co pozwala na identyfikację botnetów ale może powodować fałszywe wykrycia. Kolejnym sposobem jest identyfikacja na podstawie zapytań DNSBL botmasterów do botnetu gdy próbują sprawdzić czy ich boty zostały przez kogoś odkryte. Inną metodą wykorzystującą DNS jest metoda identyfikująca botnety na podstawie, jest to wykrywanie grupowych działań w DNS gdy wiele botów jednocześnie wykonuje zapytanie DNS lub w krótkim odstępie czasu w celu komunikacji z serwerem C&C, jest to metoda która wykazuje się znaczną skutecznością ponieważ takie zachowanie w sieci jest łatwe do odróżnienia od normalnych zapytań

8. Wykrywanie oparte na identyfikacji ruchu C&C – jest to jedna z trudniejszych metod ponieważ ruch C&C nie powoduje typowych dla botnetów anomalii jak np. wysoka latencja, w tych metodach wykorzystuje się uczenie maszynowe czy też klasteryzację. W przeszłości jedną z metod było wykrywanie nietypowej komunikacji wykorzystującej protokół IRC (Internet Relay Chat) gdzie boty komunikują się z serwerem IRC który jest pośrednikiem do serwera C&C. Obecnie wykrywanie komunikacji po protokole IRC może nie być skuteczne ponieważ protokół IRC nie jest już popularny i obecne botnety mogą korzystać z innych protokołów takich jak HTTP czy HTTPS. Jednym z narzędzi do identyfikacji ruchu C&C jest program Botminer który za pomocą klasteryzacji i korelacji podobnych złośliwych działań w sieci jest w stanie wykryć komunikację botów z serwerem C&C

Każda z wyżej wymienionych metod ma swoje zalety jak i wady a ich skuteczność może zależeć od specyfikacji botnetu oraz sytuacji w jakiej zostały te metody zastosowane. Połączenie wielu metod zwiększa prawdopodobieństwo wykrycia botnetu i minimalizacji ryzyka związanego z możliwym atakiem DDoS.

4. DDoS

Ataki DoS pojawiły się we wczesnych latach 80 ale dopiero w 1999 został zgłoszony przez Computer Incident Advisory Capability (CIAC) pierwszy przypadek ataku DDoS i po tym wydarzeniu większość przyszłych ataków była już atakami typu DDoS, powstało wiele rodzajów ataków tego typu i w tej części artykułu zostaną opisane wybrane z tych metod. Ogólne działanie ataku DDoS : atak DDoS polega na tym że wykorzystując botnet atakujący określa cel ataku który jest np. serwerem i wszystkie urządzenia w botnecie atakują określony i poprzez duże natężenie ruchu ofiara nie jest w stanie wykonywać normalnych operacji ponieważ występuje przeciążenie. Celem ataku zwykle jest wyczerpanie zasobów ofiary takich jak przepustowość czy moc obliczeniowa.

Oto znane metody ataków DDoS:

1. Ataki typu Smurf – ten typ ataku wykorzystuje najczęściej komunikaty ICMP (Internet Control Message Protocol), które normalnie stosuje się jako narzędzie diagnostyczne w sieciach. W ataku Smurf atakujący wysyła żądanie ICMP do niezabezpieczonej domeny rozgłoszeniowej co powoduje wzmocnienie ataku gdzie jeżeli w domenie znajduje się N komputerów to ofiara będzie otrzymywać N odpowiedzi ICMP od wszystkich komputerów w domenie. Jest to atak na poziomie sieci, komunikaty ping posiadają

sfałszowany adres komputera ofiary przez co odpowiedzi trafiają do ofiary a nie do atakującego. Do ataku na ofiarę może zostać wykorzystana więcej niż jedna domena rozgłoszeniowa co prowadzi do wyczerpania przepustowości i zasobów komputera ofiary

2. Ataki typu Refleciton – są to ataki pośrednie w którym wykorzystuje się np. routery czy serwery do ataku na ofiarę. Wysyłane zostają pakiety wymagające odpowiedzi ze sfałszowanym adresem ofiary przez co pakiety służące do ataku zostają odbite jako normalne pakiety w kierunku ofiary w celu zalania jego łącza. Jednym z typów ataków typu Reflection jest poprzednio wspomniany w pierwszym punkcie atak typu Smurf. Ataki Reflection opierają się na zdolności generowania wiadomości odpowiedzi przez urządzenia wykorzystywane do odbicia , przez co każdy protokół w którym występuje generowanie wiadomości może zostać wykorzystany do przeprowadzenia ataku tego typu jak np. pakiety TCP i UDP, komunikaty ICMP, pakiety SYN-ACK, RST. Rozmiar ataku określany jest przez przez pulę urządzeń odbijających oraz częstotliwość i rozmiar odbitych pakietów, są to normalne pakiety które nie mogą być łatwo filtrowane.
3. Ataki typ Reflection ze wzmocnieniem na warstwę aplikacji – w tym przypadku stosuje się zapytania DNS z sfałszowanymi adresami IP gdzie generowana jest odpowiedź i wysyłana do ofiary. Te odpowiedzi są zazwyczaj znacznie większe niż zapytania, co prowadzi do zalania ofiary dużą ilością ruchu sieciowego. Drugim przykładem ataku jest wykorzystanie sfałszowanych pakietów VoIP przez SIP z bardzo dużą szybkością i dużym zakresem adresów IP przez co zostaje użyta znaczna część zasobów ofiary ponieważ jej serwer VoIP musi odróżniać prawidłowe połączenia od tych sfałszowanych. Ten typ ataku z wykorzystaniem VoIP naśladuje duże serwery VoIP przez co jest trudny do zidentyfikowania i sprawia wrażenie prawidłowego ruchu.
4. Ataki typu Session flooding – jest to typ ataku skierowany na serwer WWW ofiary gdzie zostają wysłane z dużą częstotliwością żądania połączenia sesji która w dużo większym stopniu przewyższa częstotliwość żądań od zwykłych użytkowników przez co serwer zostaje przeciążony. Jednym z ataków tego typu jest atak z wykorzystaniem HTTP GET/POST gdzie atakujący wykorzystuje botnety które wysyłają prawidłowe żądania GET lub POST na serwer, niewielka liczba botów jest w stanie przeciążyć serwer ponieważ każdy bot może wygenerować dużą liczbę zapytań na sekundę. Oprócz HTTP, atakujący mogą wykorzystywać inne protokoły sieciowe, takie jak FTP, SMTP, SSH czy protokoły baz danych, w celu wyczerpania zasobów serwera i utrudnienia obsługi prawdziwych sesji. Wykorzystanie botnetów w tym przypadku jest podane jako przykład,

do tego typu ataków wykorzystane mogą być także skrypty czy narzędzia do testowania penetracyjnego

5. Ataki typu Request flooding – są to ataki oparte na zalewaniu żądaniami, takie jak atak pojedynczej sesji HTTP GET/POST, są formą ataku DDoS, którego celem jest zasłonięcie serwera lub usług internetowych poprzez przesłanie nieprawidłowych lub nadmiernych żądań. Podstawową strategią napastników jest wysyłanie sesji zawierających więcej żądań niż zwykle, co prowadzi do ataku DDoS. Wykorzystując możliwości protokołu HTTP 1.1, który pozwala na równoległe wysyłanie wielu żądań w ramach jednej sesji TCP/IP, napastnik może ograniczyć szybkość sesji ataku HTTP i ominąć mechanizmy obronne. Skutkiem tego ataku jest otrzymanie przez serwis ogromnej ilości danych wejściowych, co może spowodować jego awarię lub znaczne spowolnienie działania
6. Ataki asymetryczne – ten typ ataków wykorzystuje się mechanizmy gdzie odpowiedzi serwera są dużo większe niż zapytania które zostały skierowane do serwera, kluczową cechą tej metody jest to że atakujący nawet przy wygenerowaniu niewielkiej ilości żądań jest w stanie spowodować obciążenie serwera dużą ilością danych która pojawi się w odpowiedzi.
7. Ataki typ Multiple HTPP GET/POST flood - znany jako multiple VERB single request, polega na masowym wysyłaniu wielu żądań HTTP w jednym pakiecie. Umożliwia to napastnikowi utrzymanie wysokiego obciążenia serwera ofiary przy niskiej częstotliwości wysyłania pakietów, co utrudnia wykrycie przez systemy monitorowania ruchu sieciowego. Jedną z głównych zalet tego ataku jest możliwość unikania technik wykrywania anomalii. Napastnicy mogą wybierać mniej wykrywalne VERB HTTP, co pozwala im ukryć działania przed tradycyjnymi metodami wykrywania, takimi jak filtracja na poziomie protokołu czy analiza zachowań użytkowników.
8. Ataki typu Slow request/response – jest to typ ataku w którym do ofiary wysyłane są żądania charakteryzujące się tym że zużywają dużo zasobów i powodują duże obciążenie. Jednym z ataków tego typu jest Slow POST, gdzie do serwera WWW ofiary wysyłane jest prawidłowe zapytanie POST ale jego zawartość jest przesyłana z bardzo małą prędkością, nawet do pojedynczych bajtów na minutę co powoduje że serwer oczekuje na to aż atakujący zakończy przesyłać dane i zapytanie jest cały czas obsługiwane co pochłania zasoby i w przypadku wielu takich zapytań serwer zostaje obciążony na tyle że nie może obsługiwać normalnych zapytań. Kolejnym przypadkiem ataków tego typu jest atak Slowreading gdzie odpowiedzi serwera są celowo odbierane dużo wolniej nawet bajt

po bajcie co wymaga na serwerze ciągłego utrzymania połączenia z atakującym ponieważ odbiera on informacje. Ten atak wymusza na serwerze utrzymanie wielu aktywnych połączeń przez co w pewnym momencie serwer przestanie wysyłać odpowiedzi do normalnych zapytań. Następnym z ataków tego typu jest atak Slowloris gdzie atakujący wykorzystując żądania GET wysyła na serwer ofiary niekompletne żądania HTTP przez co połączenie zostaje nawiązane ale serwer nie może wykonać polecenia, kolejne żądania są wysyłane w ten sposób aż do momentu gdy maksymalna liczba możliwych połączeń serwera zostanie wyczerpana. W tej metodzie atakujący może wykorzystać nawet pojedyncze urządzenie do przeprowadzenia ataku. Innym atakiem tego typu jest HTTP Fragmentation attack gdzie do ofiary wysyłane są pakiety HTTP podzielone na małe fragmenty które powolnie wysyłane są na serwer co powoduje że połączenie zostaje utrzymywane z serwerem i serwer zostaje przeciążony w podobny sposób jak zaprezentowano w poprzednich atakach w tym podpunkcie. Ten typ ataku obecnie nie jest popularny ponieważ aktualizacje oprogramowania potencjalnych ofiar nie pozwalają serwerowi na odpowiedź na ten typ próby połączenie co uniemożliwia atak DDoS. Ataki wymienione w tym podpunkcie charakteryzują się tym że są trudne do wykrycia i wymagają od atakującego stosunkowo małej liczby urządzeń potrzebnych do ataku.

5. Wybrane strategie obronne przed atakami DDoS

Na przestrzeni lat pojawiały się kolejne rodzaje ataków DDoS i tak samo z upływem lat powstawały kolejne strategie obronne które chronią przed atakiem lub minimalizują jego wpływ na ofiarę albo też minimalizują szansę na jego wystąpienie.

Oto wybrane strategie obronne:

1. IP hopping - jest to technika zwiększająca bezpieczeństwo sieci polegająca na regularnej zmianie adresu IP urządzenia. Gdy urządzenie ofiary zostanie zaatakowane atakiem DDoS to zmieni swój adres IP co może utrudnić atak jeżeli atakujący nie przygotował się na taką sytuację, w przypadku gdy atakujący posiada funkcję śledzenia DNS metoda staje się nieskuteczna ponieważ atak zostanie przeniesiony na nowy adres
2. Wyłączanie adresu broadcast – wyłączając rozgłaszanie w całej sieci podczas ataku komputery nie mogą być już używane jako wzmacniacze ataku w przypadku typu ICMP Flood czy Smurf
3. Stosowanie poprawek bezpieczeństwa - aktualizowanie oprogramowania komputerów w sieci pozwala zwiększyć bezpieczeństwo ponieważ twórcy oprogramowania dodają poprawki zapobiegające atakom DDoS, w przypadku obrony SYN Flood jedną z

poprawek bezpieczeństwa jest zmniejszenie czasu oczekiwania na nawiązanie połączenia TCP

4. Mechanizm Push-back w routera – jest to mechanizm umożliwiający routerom izolowanie źródeł ataku. Kiedy router wykrywa atak na określoną przestrzeń nazw ogranicza wszystkie nowe żądania dotyczące tej przestrzeni nazw a następnie przekazuje informację o tym routerom połączonym z tym interfejsem. Kolejne routery mogą powtarzać tę czynność ograniczając liczbę żądań i przekazując informację dalej co prowadzi do ograniczenia możliwości atakującego i zminimalizowanie ataku.
5. DDoS-Shield – mechanizm ten wykorzystuje metody statystyczne do analiz sesji HTTP i na tej podstawie ogranicza częstotliwość żądań co jest jego głównym mechanizmem obrony.
6. Filtrowanie Ingress/Egress na routerach – wiele ataków DDoS wykorzystuje sfałszowane adresy IP ofiar, cele ataku nie są w stanie odróżnić fałszywych adresów od właściwych ale może to zostać wykryte z poziomu routera, gdzie adresy są analizowane czy mieszczą się w odpowiednim zakresie adresów co utrudnia sfałszowanie adresu IP. Niestety filtrowanie ingress/egress powoduje dodatkowe zużycie zasobów na analizę przez co nie jest często stosowane w starszych sieciach
7. Packet dropping – jest to mechanizm polegający na odrzucaniu pewnych pakietów danych w celu zmniejszenia obciążenia serwera lub sieci oraz zapobieżenia ich przeciążeniu. Podczas ataku DDoS, gdzie cel ataku jest zalewany ogromną ilością niechcianego ruchu, odrzucanie niektórych pakietów może być kluczowe dla utrzymania dostępności i sprawności usług. Jedną z metod tego typu jest Packetscore gdzie pakiety są analizowane i przypisywany jest im pewien priorytet i na tej podstawie podejmowana jest decyzja czy pakiet zostanie porzucony czy nie

6. Podsumowanie

Na przestrzeni lat powstało wiele rodzajów ataków DDoS, a zagrożenie nimi stale rośnie. Atak DDoS jest jednym z najniebezpieczniejszych ataków, który może spowodować duże straty. Kluczowa jest zatem obrona przed tego typu atakami, aby zminimalizować ryzyko. Mimo że opracowano wiele strategii obronnych, żadna z nich nie zapewnia pełnego bezpieczeństwa, ponieważ wiele rodzajów ataków DDoS jest trudnych do wykrycia. Istniejące metody są skuteczne przeciwko określonym rodzajom ataków, ale dopiero ich połączenie zapewnia większe bezpieczeństwo. Ze względu na to, że po stronie potencjalnej ofiary nie da się zapewnić pełnej obrony przed atakiem DDoS, ważnym elementem zapobiegania atakom

jest wykrywanie i eliminacja istniejących botnetów, które w większości ataków pełnią główną rolę.

Literatura

Felix Lau, Stuart H. Rubin, Michael H. Smith, Ljiljana Trajković., Distributed Denial of Service Attacks, DOI: 10.1109/ICSMC.2000.886455

Rocky K. C. Chang., Defending against Flooding-Based Distributed Denial-of-Service Attacks: A Tutorial, DOI: 10.1109/MCOM.2002.1039856

Saman Taghavi Zargar, James Joshi, David Tipper., A Survey of Defense Mechanisms Against Distributed Denial of Service (DDoS) Flooding Attacks, DOI: 10.1109/SURV.2013.031413.00127

Sanjeev Kumar., Smurf-based Distributed Denial of Service (DDoS) Attack Amplification in Internet, DOI: 10.1109/ICIMP.2007.42

Maryam Feily, Alireza Shahrestani, Sureswaran Ramadass., A Survey of Botnet and Botnet Detection, DOI: 10.1109/SECURWARE.2009.48

Moheeb Abu Rajab, Jay Zarfoss, Fabian Monrose, Andreas Terzis., A Multifaceted Approach to Understanding the Botnet Phenomenon, DOI: 10.1145/1177080.1177086

Źródła internetowe

https://www.stationx.net/ddos-statistics/ (dostęp: 20.05.2024).

https://www.usenix.org/conference/usenixsecurity17/technical-sessions/presentation/antonakakis (dostęp 25.05.2024)

Magdalena Matuła, Katarzyna Maternia, Aleksandra Sawicka, Aleksandra Rokita, Łukasz Książek
Koło naukowe SKNI „Kod”

dr inż. Bartosz Trybus
Opiekun Koła Naukowego

Bezpieczeństwo w erze sztucznej inteligencji: strategie obrony przed nowoczesnymi cyberatakami

Sztuczna inteligencja, kiedyś zarezerwowana dla świata science fiction, dzisiaj stała się dynamicznie rozwijającą się dziedziną nauki, znajdującą zastosowanie w różnych obszarach. Jednakże, wraz z jej postępem, pojawiają się nowe wyzwania i zagrożenia. Wzrasta liczba incydentów zgłaszanych przez agencje monitorujące bezpieczeństwo teleinformatyczne, takie jak NASK czy CERT, co wskazuje na to, że sztuczna inteligencja staje się narzędziem wykorzystywanym przez przestępców do łamania prawa. Artykuł omawia współczesne zagrożenia, takie jak złośliwe oprogramowanie, deepfakes, ataki phishingowe i ataki DDoS, oraz sposób, w jaki sztuczna inteligencja zasila te cyberataki. Tego typu zagrożenia najskuteczniej zwalcza się używając mechanizmów korzystających ze sztucznej inteligencji, podobnych do tych, które są wprowadzane przez cyberprzestępców. Ważnym elementem obrony jest edukacja zarówno pracowników korporacji, jak i najbliższych, poprzez zastosowanie zasad bezpieczeństwa.

Słowa kluczowe: sztuczna inteligencja, uczenie maszynowe, złośliwe oprogramowanie, deepfaki, phising, ataki DDoS

1. Wprowadzenie

Sztuczna Inteligencja, znana również jako AI to obszar badawczy, który wywodzi się z dziedzin naukowych takich jak informatyka, kognitywistka oraz nauk przyrodniczych. Celem AI jest tworzenie maszyn oraz systemów, które działały by w sposób zbliżony do umysłu ludzkiego. Narzędziem który zasila sztuczną inteligencje jest uczenie maszynowe (UM), które poprzez rozwój algorytmów umożliwia ”uczenie się” maszyn na podstawie dostarczonych im danych bez konieczność programowania w sposób tradycyjny. Istnieją trzy sposoby nauki maszyn:

- Nadzorowany – maszyna uczy się z danych, które zostały wcześniej opracowane przez człowieka
- Nienadzorowany- maszyna stara się samodzielnie odkrywać lub identyfikować wzorce w danych, które nie są opisane ani oznaczone przez wcześniejsze wskazówki czy etykiety.

- Przez wzmacnianie- maszyna uczy się przez doświadczenie, zostając nagrodzona lub ukarana za swoje działanie

Dzięki samodzielnej nauce maszyny mogą odkrywać nowe wzorce, relacje, których programiści mogli nie przewidzieć, co czyni sztuczną inteligencję potężnym narzędziem, które wprowadza rewolucjonizm w niemal wszystkich dziedzinach życia i nauki. AI wspomaga na przykład: ochronę zdrowia poprzez analizę obrazów medycznych takich jak tomografii komputerowej, aby wykrywać choroby, oraz militaryzm rozwijając autonomiczne systemy broni.

Związek pomiędzy AI a bezpieczeństwem teleinformatycznym w sieci jest złożony i wielowymiarowy. Z jednej strony, sztuczna inteligencja jest wykorzystywana do ochrony informacji np. w sieciach komputerowych poprzez analizę ruch sieciowego oraz wykrywanie potencjalnych zagrożeń. Z drugiej strony ze względu na dostęp do dużych zbiorów danych oraz zdolności analityczne, sztuczna inteligencja może stać się celem ataków cybernetycznych. AI może również być wykorzystane do przeprowadzania ataków, gdy zostanie przejęta przez cyberprzestępców. Istnieje wiele sposobów by użyć sztuczną inteligencje jako narzędzie do ataku. Najbardziej popularne to: złośliwe oprogramowanie zasilane sztuczna inteligencja, deepfakes czyli fałszywe konta, generowanie przez AI e-maili phishingowych oraz ataki DDoS wspomagane sztuczną inteligencją.

W tym artykule przybliżę na czym polegają ataki w cyberprzestrzeni oraz sposoby jak się przed nimi bronić.

2. Nowoczesne zagrożenia cybernetyczne

2.1. Złośliwe oprogramowanie

Złośliwe oprogramowanie to szkodliwe kody lub aplikacje, które mają na celu zakłócenie lub uniemożliwienie normalnego funkcjonowania urządzenia, na którym się znajdują. Gdy urządzenie zostaje zainfekowane takim oprogramowaniem, może to poprowadzić do niepożądanego dostępu, kradzieży danych lub zablokowania urządzenia, wymagając okupu za przywrócenie jego funkcjonalności. Cyberprzestępcy, którzy rozpowszechniają złośliwe oprogramowanie, motywami są głównie chęcią osiągnięcia zysku. Wykorzystują zainfekowane urządzenia do różnych działań, takich jak kradzież danych finansowych, zbieranie informacji osobistych w celu ich sprzedaży, handel dostępem do zasobów komputerowych lub wymuszanie opłat ofiar.

Złośliwe oprogramowanie występuje pod wieloma postaciami:

- Wyłudzanie informacji: Cyberprzestępcy podszywają się pod wiarygodne źródła, wysyłając fałszywe wiadomości e-mail lub tekstowe, aby wyłudzić poufne dane, takie jak nazwy użytkowników, hasła lub informacje bankowe.
- Rootkity: Programy, które ukrywają złośliwe oprogramowanie na urządzeniu, aby umożliwić cyberprzestępcom nieograniczony dostęp do systemu operacyjnego, maskując swoją obecność i działalność.
- Konie trojańskie: Oprogramowanie udające przydatne aplikacje, które podstępnie implementują złośliwe funkcje, mogą być przemycone w różnych typach programów i plików.
- Programy szpiegujące: Instalują się na urządzeniu bez zgody właściciela, monitorując online zachowanie, rejestrując klawisze klawiatury i inne dane, które są przekazywane do kontrolerów w celach szpiegowskich lub kradzieży tożsamości.
- Robaki: Złośliwe oprogramowanie rozpowszechniane przez różne kanały, takie jak wiadomości e-mail czy programy do udostępniania plików, zdolne do samodzielnego kopiowania się i rozprzestrzeniania przez sieć.
- Wirusy: Fragmenty szkodliwego kodu dołączane do plików na komputerze, uruchamiające złośliwe działania po otwarciu "zainfekowanego" pliku, mogące nawet usuwać pliki z dysku twardego lub irytować użytkowników poprzez różne działania.

Mechanizmy SI stosowane w złośliwym oprogramowaniu:

Cyberprzestępcy wykorzystują algorytmy uczenia maszynowego w celu analizy dużej ilości danych i wykrycia słabości w systemach informatycznych co pomaga w opracowaniu skutecznych strategii ataków. SI wykorzystywana jest do mechanizmów ukrywających złośliwy kod w legalnych aplikacjach i analizuje środowisko aby w odpowiednim momencie uruchomić złośliwe oprogramowanie które jest ukryte w paczce. Takie oprogramowanie może być zaszyte w systemie nawet miesiącami i gromadzić dane o środowisku po to aby zaatakować z najmniej oczekiwanym momencie.

` W 2018 roku organizacja TaskRabbit stała się celem ataku, który operował w całkowitej ukrytości. Złośliwe oprogramowanie działało poza zasięgiem w środowisku informatycznym, niezauważalne na żadnym z poziomów sieciowych ani aplikacyjnych. Po wykryciu kodu okazało się, że operuje on za pomocą inteligentnych algorytmów, adaptując się do otoczenia poprzez naukę zaufanych protokołów, dostępnych zasobów oraz cykli aktualizacji, aby zaatakować w najmniej oczekiwanym monecie i podszyć się pod zaufane aplikacje. W wyniku

tego incydentu 3,7 miliona użytkowników znalazło się narażonych na kompromitację i wyciek danych.

Innym przykładem wykorzystującym złośliwe oprogramowanie wspomagane sztuczną inteligencją jest robot TrickBot. Jego działanie polega na rozsyłaniu szkodliwych wiadomości z już skompromitowanych skrzynek e-mailowych do nowych celów, a następnie usuwaniu tych wiadomości z folderu "wysłane" w skrzynkach użytkowników. Nowe cele oraz treść wiadomości były dobierane na podstawie istniejącej korespondencji w przejętych skrzynkach e-mailowych. Według raportu specjalistów z Deep Instinct, robot zebrał bazę 250 adresów e-mailowych wraz z poświadczeniami z różnych domen i organizacji.

2.2. Deepfakes

Nazwa pochodzi od połączenia dwóch słów z języka angielskiego.:”Deep learning” czyli głębokie uczenie oraz ”fake”, oznaczające fałszywy. Określa ona technikę obróbki obrazów, którą zapoczątkował Ian Goodfellow 2014r poprzez opracowanie technologii GAN (Generative Adversial Network), która dzieli algorytm sztucznej inteligencji na dwa segmenty: pierwszy odpowiada za generowanie sztucznego obrazu, a drugi sprawdza realistyczność działania pierwszego. W rezultacie uzyskany obraz jest bardzo realistyczny. Połączenie tej technologii z generowaniem wypowiedzi głosowych, które polega na analizie tekstu przez komputer za pomocą uczenia maszynowego oraz przekształceniu go na dźwięk, umożliwia tworzenie bardzo autentycznych filmów. Zastosowanie tej techniki można zaobserwować w filmach takich jak Gwiezdne Wojny lub Szybcy i Wściekli, gdzie obraz zmarłych aktorów został wygenerowany w celach fabularnych.

To potężne narzędzie które w niepowołanych rękach służy do manipulacji ludźmi. Fałszywy wizerunek słynnych aktorów został już wielokrotnie użyty do siania dezinformacji lub kompromitacji aktorów, których twarze zostały poddane obróbce. Powstało wiele nagrań z sztucznie wygenerowanymi politykami, które miało wpłynąć na wyniki wyborów. Wiele aktorów zostało skompromitowanych oraz narażonych na złą opinie publiczną gdy ich wizerunki zostały użyte w nielegalnych filmach.

W 2020 roku zostało nadane fałszywe bożonarodzeniowe przemówienie Królowej Elżbiety II, w którym ”monarchini” miała mieć możliwość swobodnego wyrażania swoich myśli. Wygenerowana królowa wspomniała o kontrowersjach które miały miejsce w tamtym czasie w rodzinie królewskiej a na koniec zaczęła tańczyć na biurku. Nagranie zostało wydane w celu ostrzeżenia Brytyjczyków, przed łatwością dezinformacji, która może rozprzestrzeniać się w erze cyfrowej.

Obecnie tworzenie deepfaków jest bardzo powszechne. Istnieje wiele darmowych aplikacji i narzędzi, które umożliwiają tworzenie realistycznych fałszywych z łatwością, nie wymagając przy tym zaawansowanej wiedzy technicznej. Z tego powodu nie tylko już celebryci i wysoko postawieni politycy ale również zwykli obywatele mogą być wykorzystywani do manipulacji.

2.3. Ataki phishingowe

Dźwiękowe skojarzenie z fishingiem czyli z angielskiego łowieniem ryb doskonale obrazuje działanie tego mechanizmu. Cyberprzestępcy niczym rybacy stosują odpowiednio przygotowaną ”przynętę” w postaci e-mail lub SMS podszywając się pod firmy kurierskie, operatorów telekomunikacyjnych, urzędy administracji czy nawet znajomych. Chcą w ten sposób wymusić dane logowania do m.in. kont bakowych, kont społecznościowych, systemów biznesowych.

CERT w rocznym raporcie poinformowało, że w 2023 roku najbardziej popularnymi formami ataków phisingowych wymierzonych w Polaków było rozpowszechnianie fałszywych ankiet. W okresie wakacyjnym cyberprzestępcy podszywali się z PKP Intercity zapewniając, że po wypełnieniu ankiety, która miała na celu badanie satysfakcji z usług przewoźnika, użytkownik mógł otrzymać nagrodę pieniężną. Jednak by ją odebrać należało wpisać dane karty płatniczej na fałszywej stronie.

Oszuści internetowi nie ograniczali się tylko do wysyłania sms i e-mailów. Zaczęto posługiwać się naklejkami QR, które naklejone na miejski parkometrach w Krakowie miały wyłudzać dane karty płatniczej. Kierowcy chcący uiścić opłatę za parking zostali przekierowani na fałszywą stronę z logiem Krakowa, która przekazywała dane ofiar w ręce przestępców.

AI wspomaga phishing poprzez generowanie wiadomości wysyłanych do potencjalnych ofiar ataku. Sztuczna inteligencja poprzez analizę danych takich jak często wyszukiwane strony internetowe, historię zakupów może personalizować treść wysyłanych treści. AI może zmieniać adresy nadawców wiadomości e-mail na bliźniaczo podobne do znanych już odbiorcom. To sprawia, że odbiorcy mogą być bardziej skłonni uwierzyć, że wiadomość pochodzi od rzeczywistego źródła, co zwiększa skuteczność ataku.

W ostatnich latach AI-Phising podszywał się pod Facebooka. Sztuczna inteligencja wygenerowała wiadomość z informacją, że strona odbiorcy na Facebooku została uznana za naruszającą standardy społeczność i cofnięto je z publikacji. Aby rozwiązać problem użytkownik miał kliknąć w dołączany do wiadomości link by złożyć odwołanie. Link prowadził do strony, która wyłudzała dane uwierzytelniające użytkownika. Po wypełnieniu formularza atakujący otrzymywał dostęp do konta oraz powiązanych stron ofiary.

Cyberprzestępcy posługują się narzędziami podobnymi do ChatGPT w celu podszywania się pod dostawców. Atak polega na kompromitacji poczty elektronicznej dostawcy, wykorzystując już istniejące zaufanie pomiędzy klientem a dostawcą. Konwersacje z dostawcami często dotyczą kwestii związanych z fakturami i płatnościami co sprawia, że jest trudniej dostrzec ataki, które próbują naśladować te rozmowy.

2.4. Ataki DDoS

Z angielskiego distributed denial of service czyli rozproszona odmowa usługi. To ataki nakierowane bezpośrednio na systemy komputerowe lub usługi sieciowe, mają na celu zajęcie dostępnych zasobów tak by uniemożliwić funkcjonowanie całej usługi w sieci Internet. Atak przeprowadzany jest głównie na komputerach, gdzie została przejęta kontrola przy użyciu specjalnego oprogramowania (np. trojany). Właściciele tych komputerów mogą nawet nie wiedzieć, że ich urządzenia mogą być wykorzystywane do przeprowadzania ataku. Atak rozpoczyna się, gdy wszystkie przejęte komputery zaczynają jednocześnie atakować usługę WWW lub system wściela urządzenia. Ofiara ataku DDoS jest zasypywana fałszywymi próbami skorzystania z usługi. Każda taka prośba wymaga przydzielenia odpowiednich zasobów do obsługi tego żądania przez zaatakowany komputer. Duża ilość takich żądań doprowadza do wyczerpania dostępnych zasobów, co skutkuje przerwą w działaniu lub zawieszeniem systemu.

AI wspomaga takie cyberprzestępstwa poprzez zautomatyzowanie organizacji ataków. Korzystając z uczenia maszynowego analizowane są wzorce ruchu sieciowego i dostosowywane strategię ataków w czasie rzeczywistym. Uczenie maszynowe może również imitować legalny ruch w sieci co sprawia, że takie ataki są coraz trudniejsze w rozróżnieniu między złośliwymi i autentycznymi żądaniami.

Urządzenia IoT, z angielskiego Internet of Things, to wszelkiego rodzaju urządzenia, które są połączone z internetem i mogą komunikować się między sobą oraz z innymi systemami. Mechanizmy te są niewystarczająco chronione przez co łatwo można je przejąć i wykorzystać do rozszerzenia potężnych sieci botnetów na potrzeby ataków DDoS. Taki atak miał miejsce w 2016 roku, gdzie Botnet Mirai zainfekował 2,5 miliona urządzeń takich jak inteligentne lodówki, maszyny przemysłowe. Botnet atakował urządzenia IoT oparte na systemie operacyjnym Linux z zainstalowanym pakietem narzędzi unikowych o nazwie BusyBox oraz otwartym portem usługi Telnet, przez który następowała infekcja. CERT Polska w swoim raporcie z 2016 roku poinformował, że w Polsce zaobserwowano nawet 14 tysięcy przejętych urządzeń dziennie.

3. Strategie obronne przed cyberatakami w erze SI

Wzrost wykorzystania sztucznej inteligencji przez cyberprzestępców staje się coraz większym wyzwaniem. Dlatego konieczna jest stała czujność oraz inwestowanie w nowoczesne rozwiązania cyberbezpieczeństwa, aby skutecznie bronić się przed stale ewoluującymi zagrożeniami. Wykorzystując zaawansowane technologie bezpieczeństwa, firmy mogą skutecznie odpierać ataki i zabezpieczać swoje dane. W obliczu ciągłego rozwoju sztucznej inteligencji ważne jest, aby środki bezpieczeństwa rozwijały się równie szybko, aby zapewnić niezbędną ochronę dla kluczowych informacji i zasobów.

3.1.Usługi i rozwiązania w zakresie bezpieczeństwa zarządzane w oparciu o sztuczną inteligencję

Rozwój sztucznej inteligencji sprzyja nie tylko cyberprzestępcom ale również systemom bezpieczeństwa, które oferują coraz bardziej zaawansowane rozwiązania do zwalczania przestępczości w sieci. Uczenie maszynowe analizując ogromne ilości danych i ruch w sieci identyfikuje wzorce i anomalie, które mogą wskazywać na atak. Dodatkowo AI pomaga ustalić priorytety reakcji przez określenie, które zdarzenie związane z bezpieczeństwem wymaga natychmiastowej reakcji a które może zostać obsłużone przez zautomatyzowane systemy reagowania.

Przy zwalczaniu zagrożenia ze strony złośliwego oprogramowania, wykorzystuje się sztuczną inteligencje do analizy ruchu sieciowego i identyfikacji wskaźników potencjalnego naruszenia bezpieczeństwa. Specjaliści ds. bezpieczeństwa otrzymując wyniki tych analiz mogą szybko zareagować i uniknąć zagrożenia. Ponadto AI analizując zachowanie złośliwego oprogramowania identyfikuje jego rodzaj co umożliwia dobranie skutecznych metod i środków zaradczych.

Nowe technologie oparte na sztucznej inteligencji rewolucjonizują sposób, w jaki rozpoznawane i blokowane są wiadomości phishingowe jeszcze przed dotarciem do swoich celów. Zaawansowane algorytmy są w stanie wykrywać fałszywe e-maile na podstawie różnych cech, takich jak treść wiadomości czy adres nadawcy. Dodatkowo, sztuczna inteligencja może być wykorzystywana do ciągłego monitorowania mediów społecznościowych oraz innych kanałów internetowych w poszukiwaniu prób oszustw typu phishing. W ten sposób możliwe jest szybkie reagowanie na tego rodzaju zagrożenia w czasie rzeczywistym.

Na podstawie uczenia maszynowego powstał system Adaptive DDoS Protection, który w celu ochrony przed atakami DDoS tworzy profile ruchu na podstawie maksymalnej prędkości ruchu klienta z ostatnich siedmiu dni, uwzględniając różne wymiary takie jak kraj źródłowy,

agent użytkownika i protokół IP. Wartości te są aktualizowane codziennie z użyciem percentyla 95, eliminując skrajne wartości. Dodatkowo, system wykorzystuje uczenie maszynowe do identyfikowania ruchu pochodzącego od botów. Profilowanie ruchu odbywa się asynchronicznie, minimalizując opóźnienia w ruchu klientów, a następnie profile są rozpraszane po sieci w celu wykrywania i łagodzenia ataków DDoS.

3.2. Strategie obronne w korporacjach

Ważnym aspektem obrony przez cyberatakami w firmach jest uświadamianie i edukacja pracowników w formie szkoleń. Istnieje wiele programów, które mają na celu wspieranie firm w budowie solidnych i skutecznych systemów bezpieczeństwa.

Narodowy Instytut Standaryzacji i Technologii (NIST) opracował Ramy Bezpieczeństwa Cybernetycznego czyli zbiór wytycznych, praktyk i zasad mających na celu zwiększenie odporności na zagrożenia w sieci. NIST klasyfikuje swoje podejście w pięciu funkcjach, które są rozdzielone na 23 kategorie Każda kategoria ma swoje podkategorię, których w sumie jest 108. Pierwsza z funkcji to Zidentyfikuj, moduł ten oparty jest o zrozumienie zasobów, danych firmy oraz potencjalnych zagrożeń. Kolejna funkcji to Chroń, skupia się na implementacji odpowiednich zabezpieczeń, aby chronić zasoby firmy. Trzecia funkcja to Wykryj, polega ona na opracowaniu metod wykrywania zdarzeń cybernetycznych w organizacji. Czwarta to Odpowiedź, funkcja ta skupia się na szybkiej reakcji firmy na atak. Ostatnia funkcja to Odzyskaj, opracowuje ona sposób na przywrócenie normalnego funkcjonowania firmy po ataku oraz działać by zapobiegać podobnym incydentom w przyszłości. Każda z tych funkcji obejmuje szereg kategorii, takich jak zarządzanie aktywami, kontrola dostępu, świadomość i szkolenie, analiza ryzyka, planowanie reagowania i inne. Te kategorie zawierają podkategorie, które precyzują konkretne działania, jakie organizacja powinna podjąć, aby spełnić cele bezpieczeństwa cybernetycznego.

Kolejną pomocą dla firm jest standard ISO 27001, który opiera się na systematycznym podejściu do bezpieczeństwa informacji, koncentrując się na ustanowieniu i utrzymaniu Systemu Zarządzania Bezpieczeństwem Informacyjnym. Ten system dostarcza strukturalnej ramy do wdrażania i zarządzania kontrolami bezpieczeństwa, w tym tych dotyczących łagodzenia zagrodzeń opartych na sztucznej inteligencji. Standard ten skupia się na ocenie ryzyka, kontrolach dostępu, szkoleniach z zakresu świadomości bezpieczeństwa, zarządzaniu incydentami oraz zarządzaniu lukami w zabezpieczeniach.

Wdrażając zarówno NIST CSF, jak i ISO 27001, firmy mogą zbudować kompleksową i wielowarstwową ochronę przed cyberatakami wykorzystującymi sztuczna inteligencję.

Elastyczność NIST CSF pozwala firmą dostosować swoje podejście w oparciu o ich specyficzne potrzeby, podczas gdy uporządkowana systematyzacja ISO 27001 zapewnia spójne i skuteczne wdrażanie kontroli bezpieczeństwa

3.3. Zasady bezpieczeństwa zwykłych użytkowników

W dzisiejszych czasach cyberataki napędzane przez sztuczną inteligencję są skierowane także przeciwko zwykłym obywatelom. Stanowią oni łatwy cel dla cyberprzestępców, którzy wykorzystują ich nieuwagę oraz zaabsorbowanie codziennymi sprawami. Dlatego tak istotne jest uświadamianie ludzi, prowadzenie szkoleń w szkołach, programach telewizyjnych, a przede wszystkim rozmowy z najbliższymi, aby pomóc im zrozumieć, jak mogą się ustrzec przed atakami stosując się do prostych zasad.

Pierwszą z zasad jest bezpieczne korzystanie z skrzynki pocztowej. Należy uważnie sprawdzać adresy mailowe, ponieważ AI jest w stanie wygenerować łudząco podobne nazwy adresów mailowych do adresów znanych nam ludzi. Ważne jest również nie wchodzenie w linki, niewiadomego pochodzenia. W takich załącznikach może kryć się złośliwe oprogramowanie.

Kolejną pomocą może być ustalenie wśród najbliższych hasła, w celu ochrony przed deepfakami. Gdy ktoś z rodziny lub znajomych zadzwoni prosząc np. o gotówkę lub poufne informacje nie używając wcześniej ustalonego hasła może być to podpowiedź, że znajomy nam głos może być wygenerowany przez AI. Warto uprzedzić najbliższych, żeby ograniczali udostępnianie swojego wizerunku w internecie aby sami nie zostali przedmiotem obróbki AI.

Ważną rzeczą jest ustalanie silnych haseł do kont na różnych portalach społecznościowych, skrzynek pocztowych oraz innych platform. Hasło takie powinno mieć co najmniej 10 znaków zawierające duże i małe litery oraz znaki specjalne. Powinno się unikać ustalania haseł zawierające nasze dane osobowe. Warto stosować dwuetapowe uwierzytelnianie, które jest procesem zabezpieczania konta lub dostępu do systemu, który wymaga potwierdzenia tożsamości użytkownika za pomocą dwóch różnych czynników. Te czynniki mogą obejmować coś, co użytkownik zna (np. hasło), coś, co użytkownik ma (np. urządzenie fizyczne) lub coś, czym użytkownik jest (np. biometryczne dane, takie jak odcisk palca).

4. Podsumowanie

W Erze Sztucznej Inteligencji, rozwój techniczny wspomaga wiele dziedzin życia oraz nauki. Rozwój ten wpływa również na cyberbezpieczeństwo. Sama sztuczna inteligencja jest narzędziem neutralnym, sposób w jaki wykorzystywane jest to potężne narzędzie przez ludzi

deteminuję skutki jej użycia. Zdolności analityczne, analiza ruch sieciowego dostęp do ogromnych ilości danych jakie posiada AI wykorzystywane nierzadko przez cyberprzestępców, którzy wykorzystują sztuczna inteligencje do zasilania złośliwego oprogramowania, deepfaków, phisingu, ataków DDoS. Jednak w odpowiedzi na rosnące zagrożenia powstaje wiele systemów bezpieczeństwa opartych właśnie na AI, które wykorzystują zdolności Sztucznej Inteligencji do obrony. Można zatem stwierdzić, że rozwój AI w cyberbezpieczeństwie jest samonapędzający, im większe zagrożenie tym lepsze systemy bezpieczeństwa są wprowadzane, jednak bardziej rozwinięta ochrona sprawia, że cyberprzestępcy tworzą coraz bardziej skuteczne strategię ataków.

W dzisiejszych czasach bardzo ważna jest edukacja o zagrożeniach cybernetycznych oraz sposobach ochrony przed nimi. Powstają specjalne pomoce dla firm takie jak struktury NIST CSF oraz ISO 27001, które pomagają firmom wdrażanie systemów zabezpieczających przed cyberatakami oraz prowadzonych jest wiele szkoleń dla pracowników. Ponieważ ofiarami cyberataków coraz częściej są zwykli obywatele należy edukować także naszych znajomych, rodzinę, dzieci przedstawiając im podstawowe zasady bezpieczeństwa w internecie.

Źródła internetowe:

1. https://blog.theprotocol.it/artykul/wprowadzenie-do-sztucznej-inteligencji-i-uczenia-maszynowego?utm_source=google&utm_medium=cpc&utm_campaign=protocol_blog&utm_term=dsa_artykuly&campaignid=20534266769&adgroupid=150131427101&keyword=&gad_source=1&gclid=CjwKCAjw26KxBhBDEiwAu6KXt2YYhJ28egPnsqHruE2se2iILcjq0mkJpMOZGGxqYY9awSv6LTeQ3xoCd0AQAvD_BwE(dostęp: 05.07.2024).
2. https://kancelarierp.pl/sztuczna-inteligencja-to-zarowno-szansa-jak-i-potencjalne-zagrozenie/(dostęp: 05.07.2024)
3. https://www.microsoft.com/pl-pl/security/business/security-101/what-is-malware(dostęp: 05.07.2024)
4. https://kapitanhack.pl/2020/01/14/ai-kampania/jak-sztuczna-inteligencja-wykorzystywana-jest-w-zlosliwym-oprogramowaniu/(dostęp: 05.07.2024)
5. https://www.xcubelabs.com/blog/artificial-intelligence-in-cybersecurity-ai-cyberattacks-securing-your-ecosystem-with-ai-and-more/(dostęp: 05.07.2024)
6. https://www.bbc.com/news/world-europe-68931214(dostęp: 05.07.2024)
7. https://managerplus.pl/sztuczna-inteligencja-ulatwi-cyberprzestepcom-ataki-phishingowe-89475(dostęp: 05.07.2024)

8. https://cert.pl/uploads/docs/Raport_CP_2023.pdf(dostęp: 05.07.2024)
9. https://kapitanhack.pl/2023/06/23/ai-kampania/sztuczna-inteligencja-wykorzystywana-w-phishingu/(dostęp: 05.07.2024)
10. https://pomoc.home.pl/baza-wiedzy/co-to-jest-atak-ddos(dostęp: 05.07.2024)
11. https://www.link11.com/en/blog/threat-landscape/how-artificial-intelligence-is-changing-ddos-attacks/(dostęp: 05.07.2024)
12. https://cert.pl/posts/2020/03/analiza-bota-mirai-oraz-jego-wariantow/(dostęp: 05.07.2024)
13. https://arctiq.com/blog/ai-vs-ai-fighting-ai-powered-phishing-and-malware-with-ai-powered-cybersecurity-solutions(dostęp: 05.07.2024)
14. https://blog.cloudflare.com/adaptive-ddos-protection-pl-pl(dostęp: 05.07.2024)
15. https://scytale.ai/resources/defending-against-ai-based-cyber-attacks/(dostęp: 05.07.2024)

Łukasz Książek, Katarzyna Maternia, Magdalena Matuła, Aleksandra Sawicka, Aleksandra Rokita
SKNI „Kod”

dr inż. Bartosz Trybus
Opiekun Koła Naukowego

Bezpieczeństwo i prywatność w obliczeniach w chmurze

Streszczenie

Celem artykułu jest uświadomienie czytelnika w kontekście cyberbezpieczeństwa w aplikacjach chmurowych oraz zapoznanie go ze sposobami ataku na dane prywatne oraz obrony przed nimi. Omawia on zagadnienia związane z wykorzystaniem chmury obliczeniowej oraz metody ochrony danych w tym środowisku. Artykuł został rozpoczęty od przedstawienia podstawowych informacji dotyczących chmury obliczeniowej prezentując jej definicję oraz rodzaje. W kontekście zabezpieczeń danych omówiono różne metody zabezpieczeń, w tym szyfrowanie danych oraz zarządzanie tożsamością i dostępem. Przytoczone zostały współcześnie stosowane algorytmy oraz protokoły. Omówiono także stosowane sposoby ataku na usługi chmurowe, wraz ze współczesnymi metodami obrony przed nimi. Prezentowane są przykłady dotyczące nie tylko kradzieży danych poprzez serwer ale także ingerowania bezpośrednio w innego użytkownika poprzez jego przeglądarkę czy stosując inżynierię społeczną. Tekst uświadamia, że analizowany problem dotyczy każdego człowieka, niezależnie od tego w jakim stopniu używana on danej technologii.

Słowa kluczowe: cyberbezpieczeństwo, aplikacje chmurowe, atak na dane prywatne, metody zabezpieczeń, problem dotyczy każdego człowieka

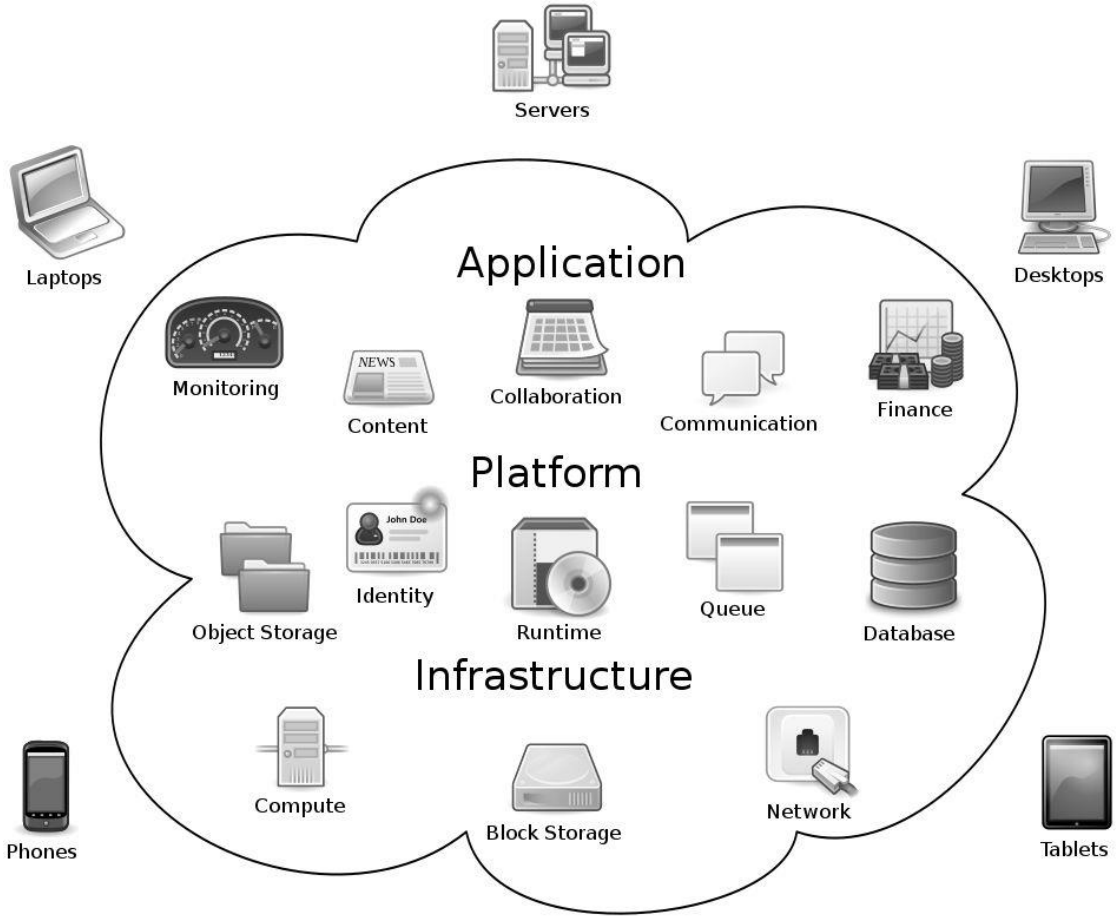
1. Wprowadzenie

Wraz z dynamicznym rozwojem technologii obliczeniowych, obliczenia w chmurze stały się integralną częścią życia codziennego oraz działalności biznesowej. Zapewniają one niezwykłą elastyczność, skalowalność i dostępność zasobów informatycznych, które mogą być udostępniane i wykorzystywane przez użytkowników na całym świecie. Jednakże, w miarę jak coraz więcej danych i aplikacji przechodzi do chmury, nieuniknione stają się również kwestie dotyczące bezpieczeństwa i prywatności. Razem z rosnącą liczbą ataków hakerskich, wycieków danych i naruszeń prywatności, konieczne jest podejmowanie środków zaradczych oraz ciągłe doskonalenie praktyk bezpieczeństwa, aby zapewnić użytkownikom optymalne warunki ochrony ich danych i prywatności w środowisku chmurowym.

2. Czym jest chmura obliczeniowa?

Chmura obliczeniowa, zwana także przetwarzaniem w chmurze, to model korzystania z usług dostarczanych przez zewnętrznego dostawcę, który może być zarówno działem wewnętrznym firmy, jak i zewnętrzną organizacją. W tym modelu, użytkownicy korzystają z usług, takich jak oprogramowanie i infrastruktura, bez konieczności zakupu i instalacji własnych licencji czy oprogramowania. Zamiast tego, płacą za korzystanie z określonych usług, na przykład za dostęp do programu do tworzenia arkuszy kalkulacyjnych, bez potrzeby inwestowania w sprzęt czy

oprogramowanie. Umowy dotyczące korzystania z usług w chmurze są zazwyczaj standardowe i niezwiązane z konkretnym dostawcą. Termin "chmura obliczeniowa" wywodzi się z idei wirtualizacji.



Rysunek 1. Diagram przedstawiający „chmurę”
Źródło: https://pl.wikipedia.org/wiki/Chmura_obliczeniowa

System ten polega na przeniesieniu danych, oprogramowania lub mocy obliczeniowej na serwer, który udostępnia zasoby swoim klientom. Z tego powodu bezpieczeństwo oraz szybkość procesów nie zależą od urządzenia użytkownika.

Chmury można podzielić na:

- prywatne – prywatne chmury danych przedsiębiorstw udostępniające usługi działom biznesowym i partnerom.

- publiczne – udostępniane za pośrednictwem internetu po pobraniu opłat od klienta. Opłaty różnią się w zależności od stopnia wykorzystanych zasobów. Obecnie najpopularniejsze z nich to Amazon Web Services (AWS), Microsoft Azure i Google Cloud Platform (GCP)

- hybrydowe – są połączeniem chmury prywatnej i publicznej. Pewna część aplikacji i infrastruktury danego klienta pracuje w chmurze prywatnej, a część w przestrzeni chmury publicznej.

Technologia ta niesie za sobą wiele zagrożeń, takich jak nieautoryzowany dostęp do infrastruktury i danych wrażliwych, przechwytywanie kont, usług i ruchu w sieci, a także kradzież usługi, Liczba ataków na systemy chmurowe wzrosła z 2021 roku na 2022 aż o 228 procent.

3. Zabezpieczenia danych w chmurze

3.1 Szyfrowanie danych

Uznaje się to rozwiązanie jako najlepszy sposób na ochrone informacji. Jest to proces polegający na konwersji niezabezpieczonych danych na inny bezpieczniejszy format. Używane są w tym celu algorytmy, które według przyjętej metody zmieniają informację tak, aby były one nieczytelne. Aby je rozszyfrować należy użyć unikalnego klucza, który może być przechowywany lokalnie, przez dostawców chmury lub przez dostawców IT. W przypadku szyfrowania lokalnie wymagana jest weryfikacja dwustopniowa – przykładowo dane logowania takie jak hasło są przechowywane na urządzeniu użytkownika, a po jego ponownym użyciu wysyłane jest potwierdzenie np. poprzez SMS. W drugim przypadku całość szyfrowania i deszyfrowania odbywa się natomiast na serwerze dostawcy. Użytkownicy działają wyłącznie na poziomie chronionego interfejsu, który nie jest dostępny lub nie może zostać obsłużony offline. Nie powinno się jednak przechowywać klucza i danych na tej samej instancji, ponieważ w przypadku nieautoryzowanego dostępu do danych można uzyskać również dostęp do klucza, co sprawia że dwuetapowa weryfikacja dostępu znacząco zwiększa bezpieczeństwo danych użytkownika.

Można wyróżnić dwa procesy szyfrowania:

- W spoczynku - dotyczy plików zapisanych na serwerze/kliencie, które aktualnie nie są przesyłane w sieci lub pomiędzy klientem a serwerem.
- W locie - dotyczy momentu komunikacji czy też przesyłu danych w sieci lub pomiędzy klientem a serwerem.

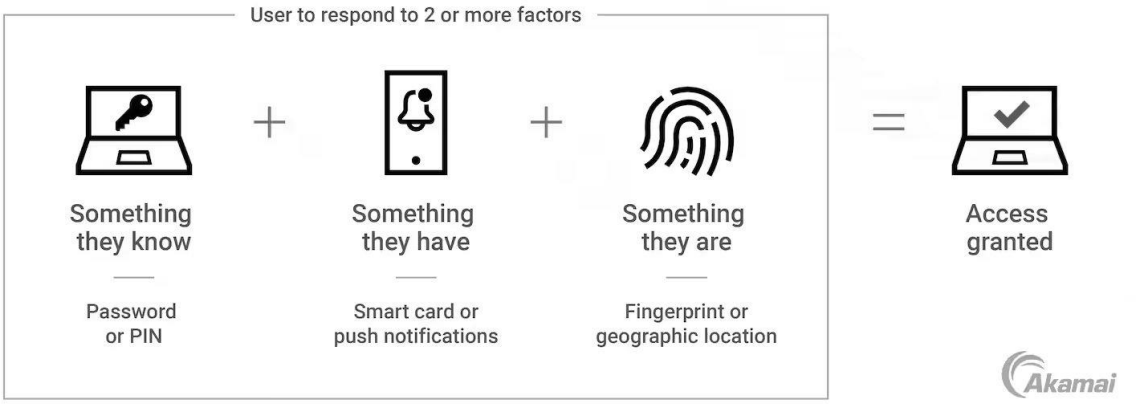
Dane w spoczynku w zależności od modelu działania chmury mogą być przechowywane w przestrzeni dysku wirtualnego lub na dysku lokalnym. W przypadku danych w locie ochronie podlega więcej elementów jak np. API, połączenie sieciowe, instancja chmurowa czy sam dysk fizyczny dostawcy. Z tego powodu dane w locie są łatwiejszym celem do przechwycenia.

3.2 Zarządzanie tożsamością i dostępem

Zarządzanie tożsamością i dostępem (Identity and Access Management, IAM) to zestaw narzędzi i procedur, które zapewniają kontrolę nad dostępem do danych czy zasobów organizacji. Kontrola ta opiera się na identyfikowaniu użytkowników w celu przydzielenia im określonych zasobów i uprawnień do aplikacji, usług chmurowych i lokalnych. Pomimo, że IAM jest kojarzone z chmurą, ma zastosowanie w lokalnych systemach informatycznych. IAM obejmuje takie pojęcia jak:

- MFA (Multi-Factor Authentication),
- SSO (Single Sign-On),
- JWT (JSON Web Token),
- IDaaS (Identity as a Service),
- UEBA (User and Entity Behavior Analytics),
- CASB (Cloud Access Security Broker),
- RBAC - Role-Based Access Control,
- SIEM - Security Information and Event Management.

Multi-Factor Autentication



Rysunek 2. Schemat uwierzytelniania wielostopniowego.
Źródło: https://www.akamai.com/glossary/what-is-cloud-mfa

MFA jest często spotykanym sposobem ochrony dostępu do zasobów chmurowych. Jest to zabezpiecznie już na etapie logowania użytkownika do systemu. Jest on wtedy zobowiązany potwierdzić swoją tożsamość nie tylko znajomością hasła ale także wymagane jest od niego aby podał dodatkowy składnik uwierzytelniania. Ta metoda zapobiega sytuacjom, kiedy nieautoryzowana osoba zyskała dostęp do danych logowania innego użytkownika.

Formy uwierzytelniania MFA:

Do uwierzytelniania MFA mogą zostać wykorzystane metody, które można skategoryzować wg czterech następujących kryteriów:

- „coś, co użytkownik wie”: dane, które zna tylko użytkownik np. hasło, kod PIN,
- „coś, czym użytkownik jest”: dane biometryczne, unikalne cechy fizyczne użytkownika np. odcisk palca, skan tęczówki oka, skan twarzy itp.,
- "coś, co użytkownik posiada”: przedmiot lub urządzenie do uwierzytelniania np. token, aplikacja na smartfonie, klucz, karta bankomatowa itp.,
- „miejsce, w jakim użytkownik aktualnie się znajduje”: np. wykorzystanie GPS, albo IP do potwierdzenia typowych lokalizacji użytkownika.

Dodatkowe składniki uwierzytelniania użytkownika mogą przybierać różne formy, takie jak:

- ograniczone czasowe hasło jednorazowe (TOTP) z aplikacji zainstalowanej na urządzeniu mobilnym,
- token sprzętowy,
- hasło lub link wysłany za pośrednictwem e-mail,
- kod wysłany za pomocą SMS,
- kod z powiadomienia push,
- uwierzytelnianie biometryczne (skanowanie źrenicy oka, linii papilarnych, twarzy, weryfikacja tonu głosu),
- autoryzacja behawioralna,
- pytanie zabezpieczające (użytkownik musi odpowiedzieć na pytanie wg określonego wzorca).

Single Sign-On

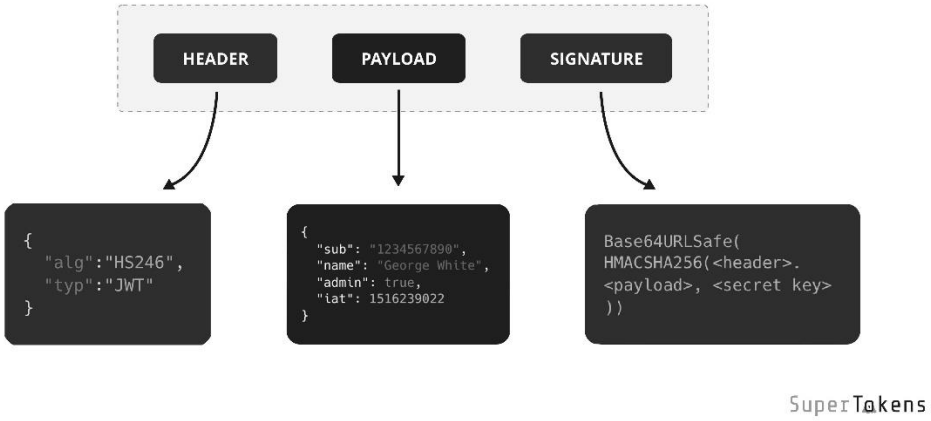
SSO to możliwość jednorazowego zalogowania się do usługi sieciowej i uzyskania dostępu do wszystkich autoryzowanych zasobów zgodnych z tą usługą. Technikę pojedynczego logowania zaimplementowano m.in. w Koncie Google, Facebook Connect, czy Koncie Microsoft. Stworzenie takiego logowania ułatwiają standardy potwierdzania tożsamości m.in. OpenID, OpenID Connect, czy OAuth.

Współczesna technologia umożliwia tworzenie zróżnicowanej architektury systemów, gdzie logowanie odbywa się w jednym miejscu, przy braku wspólnej bazy danych. Przykładem takiej technologii może być JSON Web Token (JWT).

JSON Web Token

Technologia ta polega na odbieraniu zakodowanego biletu (z ang. token) przez przeglądarkę czy aplikacje mobilną. Token zawiera wszystkie dane potrzebne do uwierzytelnienia w usługach, wraz z podpisem gwarantującym autentyczność i nienaruszalność tych danych. Swoje działanie opiera na trzech głównych segmentach:

- nagłówku,
- danych użytkownika,
- podpisie cyfrowym.



Rysunek 3. Struktura JSON Web Token
Źródło: https://supertokens.com/static/b0172cabbcd583dd4ed222bdb83fc51a/9af93/jwt-structure.png

Nagłówek przeważnie zawiera w sobie typ tokenu oraz algorytm kodujący. Drugi segment jest nazywany *payloadem*, który przechowuje dane użytkownika. Trzeci, najważniejszy element to podpis cyfrowy. Jest to zabezpieczenie przed nieautoryzowanym dostępem, zmianą danych czy innymi tego typu zagrożeniami. Dzięki temu podpisowi serwer może zweryfikować token.

Identity as a Service

Zarządzanie tożsamościami w chmurze jest ułatwione dzięki usłudze znanej jako Identity as a Service (IDaaS). Jest to usługa oparta na modelu chmurowym, która umożliwia zarządzanie tożsamościami (IAM) oraz kontrolę dostępu do danych, bez konieczności tworzenia i utrzymywania własnej infrastruktury. IDaaS umożliwia przechowywanie, zarządzanie i zabezpieczanie informacji o użytkownikach w jednym centralnym punkcie dostępu.

Identity as a Service obejmuje:

- SSO,
- MFA,
- usługi katalogowe w chmurze,
- zarządzanie poziomem dostępu pracowników.

User and Entity Behavior Analytics

Jednym z najbardziej niedocenianych elementów bezpieczeństwa IT jest czynnik ludzki. Pomimo swobodnego dostępu do wiedzy na temat kradzieży danych dalej firmy narażone są na atak z powodu błędów pracowników takich jak:

- Brak rzetelności pracowników,
- Niewystarczająca edukacja w zakresie bezpieczeństwa IT,
- Nadużywanie uprawnień przez członków zespołu.

Problemem są również działania przestępcze I szantaże członków zespołu. Tymi problemami zajmują się narzędzia UEBA (user and entity behavior analytics). Koncentrują się one na codziennych zachowaniach użytkowników, dzięki temu wykrywane są zagrożenia wewnętrzne ze strony pracowników. UEBA to analiza zachowań użytkowników, robotów serwisowych oraz innych urządzeń należących do firmowej sieci. System ten gromadzi informacje w sieci pozostawiane przez pracowników i na tej podstawie określa ich normalną aktywność. Następnie monitorowana jest aktywność użytkowników I na podstawie pewnych wyuczonych standardowych zachowań wykrywane są”anomalie”, czyli zachowania odbiegające od normy. UEBA wykorzystuje uczenie maszynowe, dzięki czemu jest w stanie analizować ogromne ilości pobranych informacji, takich jak:

- odwiedzane strony internetowe,
- ilość pobieranych danych,
- używane aplikacje.

Jeśli system wykryje taką anomalię, jest w stanie zablokować podejrzaną aktywność lub przekazać informację do pracownika ochrony.

Dzięki integracjii UEBA z innym systemem wykrywania zagrożeń (np. Security Information and Event Management, SIEM) możliwe jest ustalenie hierarchii zagrożeń wykrytych anomalii. Bardziej zaawansowane modele UEBA potrafią wykrywać oszustwa, kradzieże własności intelektualnej, działania szpiegowskie czy szeroko rozumiane nadużycia.

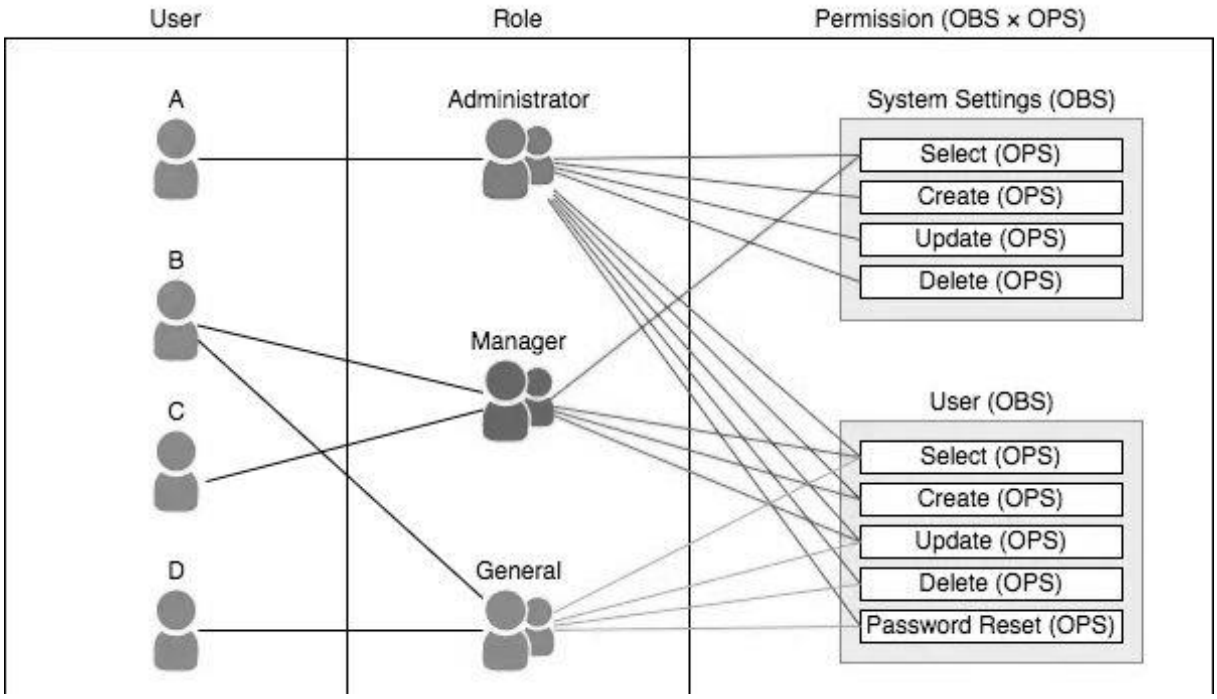
Kolejnym ważnym elementem systemów zabezpieczeń chmury jest CASB (Cloud Access Security Broker). Narzędzie to zabezpiecza dane transportowane pomiędzy przedsiębiorstwem a chmura. System ten skupia się głównie ochronie przed szeroko pojętymi cyberatakami. Broker CASB szyfruje dane w ten sposób aby złośliwe oprogramowanie nie mogło się przedostać do serwerów przedsiębiorstw oraz aby strumienie danych były nieczytelne dla zewnętrznych przedmiotów. Model działa według czterech filarów:

- widoczność – zapewnia wgląd a aktywność (kto używa aplikacji, jaka jest lokalizacja oraz jakie urządzenia biorą udział w tym procesie) użytkowników podczas korzystania z konta aplikacji w chmurze,
- spełnienie – w momencie przenoszenia danych do chmury, organizacje chcą zapewnić zgodność przepisów o prywatności I bezpieczeństwie danych. CASB może pomóc skonfigurować wymogi bezpieczeństwa z wymaganymi regulacjami,

- ochrona danych – gwarantem jest szyfrowanie, tokenizacja, kontrola dostępu oraz zarządzanie dostępem do informacji,
- ochrona przed zagrożeniami – wykrywanie niezidentyfikowanych oraz złośliwych zagrożeń wewnętrznych i zewnętrznych.

Role-based access control

Mechanizm RBAC (Role-based access control) umożliwia kontrolę dostępu operatego na rolach. Role tutaj oznaczają przydzielony zakres obowiązków dla każdego z użytkowników, co daje im uprawnienia do wykonywania określonych dla tych ról czynności. RBAC sprawdza się bardzo dobrze tam, gdzie stosuje się zasadę podziału obowiązków. Przykładowo, gdy w celu zapobiegania nadużyciom niektóre operacje wymagają akceptacji dwu niezależnych użytkowników.



Rysunek 4. Przykładowa organizacja ról.
Źródło: https://dsonoda.medium.com/role-based-access-control-overview-257de64534c

4. Metody ataku na usługi chmurowe

Pishing

Jest to jeden z najpopularniejszych typów ataków opartych o wiadomości e-mail lub SMS. Nazwa *phishing* nawiązuje w swojej nazwie do *fishing* (ang. łowienie ryb) – przestępcy podobnie jak wędkarze stosują pewne *przynęty* na ludzi jak wędkarze na ryby. Metoda ta wykorzystuje inżynierię społeczną I manipulację ofiary aby podjęła ona określone działania. Oszuści podszywają się pod np.

- firmy kurierskie,
- urzędy administracji,
- operatorów telekomunikacyjnych,
- przyjaciół,
- członków rodziny.

Ich celem jest zdobycie danych logowania użytkowników do np. kont bankowych czy portali społecznościowych. Tworzą w tym celu wiadomości, które sprawiają wrażenie autentycznych, ale w rzeczywistości nie są. Ma to skłonić użytkowników do ujawnienia poufnych informacji lub kliknięcie na link przypięty do wiadomości zawierający szkodliwe oprogramowanie.

Istnieje również pojęcie *spear-pushingu*. Jest to również podszywanie się pod inne osoby w celu jej zmanipulowania ale w tym wypadku jest to ukierunkowane tylko na jednego konkretnego adresata. Wiadomości mogą być spersonalizowane co zwiększa prawdopodobieństwo wyłudzenia.

Brute Force

Jest to prosta technika łamania haseł polegająca na sprawdzeniu wszystkich możliwych kombinacji. W teorii da się złamać tą metodą każde hasło – w praktyce jest ona bezużyteczna. Jest to spowodowane bardzo dużą złożonością obliczeniową tego algorytmu. Liczbę takich *n* kombinacji można wyrazić matematycznie:

$$n = 2^k$$

Gdzie *k* to ilość bitów w haśle.

Zakładając, że w rzeczywistym komputerze sprawdzenie takiego hasła zajmuje pewien czas, można napisać:

$$T_p = 2^k t$$

Gdzie *t* to czas wykonywania tej pojedyńczej operacji, a T_p to cały czas poświęcony na jego odgadywaniu.
Jest to jednak zapis dla przypadku pesymistycznego, gdzie hasło zostaje złamane przy ostatniej próbie. Statystycznie hasło zostaje łamane w połowie:

$$T_{śr} = 2^{k-1} t$$

Dla przykładu – łamanie standardowego 128-bitowego szyfru AES, przy założeniu, że komputer potrafi wykonać million prób na sekundę otrzymujemy średnią:

$$2^{127} 10^{-6} = 1.7 \cdot 10^{32} [s]$$

Jest to $5.5 \cdot 10^{23}$ lat, gdzie wiek wszechświata szacowany jest na $13.82 \cdot 10^{9}$ lat. Pokazuje to niską skuteczność tego algorytmu.

Wariantem ataku brute force jest atak słownikowy. Polega on na sprawdzaniu wszystkich możliwych kombinacji haseł, ale jedynie z podanej bazy haseł. Wykorzystuje się do tego np. bazę wyrazów występujących w danym języku lub historyczną bazę popularnych haseł.

Dlatego korzystanie z długich fraz, zawierających kombinacje liter, cyfr i znaków specjalnych, zwiększa liczbę możliwych kombinacji, co znacznie utrudnia zadanie hakerom. Dodatkowo, regularna zmiana haseł oraz unikanie stosowania tych samych danych do różnych kont i usług również zwiększa poziom bezpieczeństwa. Pamiętanie o tych podstawowych zasadach może zdecydowanie zminimalizować ryzyko naruszenia bezpieczeństwa i utraty danych.

DDoS

Kolenym atakiem, nie prowadzącym do kradzieży danych ale uniemożliwiający korzystaniem z chmury jest DDoS (Distrubuted Denial of Service). Polega na zajęciu wsztstkich dostępnych zasobów, z wielu komputerów jednocześnie.

Atak DDoS jest odmianą ataku DoS polegającą na zaatakowaniu celu z kilku miejsc w tym samym czasie. Służą temu komputery, nad którymi przejęto kontrolę przy użyciu oprogramowania takiego jak boty lub trojany. Przy ataku każdy komputer zasypuje system fałszywego skorzystania z usług. Dla każdego takiego wywołania komputer musi przydzielić zasoby klientowi, co przy bardzo dużej liczbie żądań prowadzi do ich wyczerpania.

Groźba atakiem DDoS bywa używana do szantażowania firm aukcyjnych np. firm brokerskich, serwisów aukcyjnych, gdzie przerwa w działaniu systemu przekłada się na duże straty finansowe dla firmy i klientów.

Man-in-the-Middle (MitM)

Jest to atak kryptologiczny polegający na podsłuchu i modyfikacji danych przesyłanych pomiędzy dwiema stronami bez ich wiedzy.

Przykładem takiego ataku jest narzucenie nadawcy własnego klucza szyfrującego przy transmisji chronionej szyfrem asymetrycznym. Atakujący w tym wypadku musi początkowo przekierować ruch z komputera wysyłającego na swój komputer. Może to zrobić poprzez modyfikację danych podawanych przez DNS lub podsłuchując zapytania DNS i przekierowując je na swój komputer. Po połączeniu atakujący przekazuje swój klucz publiczny odbiorcy podszywając się pod osobę wysyłającą, jednocześnie nawiązując z nią kontakt. Otrzymane dane od celu przekazuje osobie, do której miały dotrzeć. Jeżeli podsłuchiwana osoba będzie chciała się zalogować używając np. loginu i hasła, to dane przechodzą przez komputer osoby atakującej, która te dane przechwytuje.

Oszust może monitorować całą komunikację, zbierając wszystkie przesłane informację, np.

- stan konta,
- dane osobowe,
- numery karty płatniczej.

W ten sposób można przechwycić dane pozornie bezpiecznym kanałem, bez potrzeby łamania szyfrów zabezpieczających.

Atak ten zabiezpiecza się poprzez weryfikację klucza – np. gdy jest on podpisany przez organizację certyfikującą. Gdy strona internetowa ma zabezpieczenia w postaci protokołow SSL i TLS atak MitM jest niemożliwy, a strona zamiast przedrostka *http://* ma *https://*.

Injections

Atak ten jest stosowany w przypadku braku odpowiednich zabezpieczeń. Możliwe jest dodać własne zaoytania SQL do formularzy strony internetowej, aby skomunikować się z bazą danych. Przykładowo, oszust zamiast wprowadzić hasła do konta może użyć zapytania SQL. Jeżeli formularz nie jest zabezpieczony, aplikacja użyje tego zapytania, które zostanie wykonane z poziomu bazy danych.

Można się zalogować tym sposobem jako administrator, uzyskać informację na temat zapisanych loginów i haseł, lub w najgorszym przypadku usunąć bazę danych.

Ataki SQL są przeprowadzane na poziomie formularzy wbudowanych w aplikacje webowe, np.:

- w polach logowania,
- w wyszukiwarkach produktów,
- w formularzach kontaktowych,
- w formularzach rejestracyjnych do newsletterów.

Taki sposób ataku można podzielić na kilka typów:

- In-band SQLi (klasyczny SQLi)
    - Error-based SQLi
    - Union-based SQLi
- Inferential SQLi (Blind SQLi)
    - Boolean-based (content-based) blind SQLi
    - Time-based blind SQLi
- Out-of-band SQLi

\Pierwszym sposobem jest In-band SQLi, gdzie napastnik wykorzystuje tylko jeden kanał do przeprowadzenia ataku. Jest to najprostszy i najpopularniejszy typ SQL injection. Pierwszym podtypem tego sposobu jest Error-based SQLi, gdzie wykorzystując błędy bazy danych można uzyskać informacje typ, wersja bazy danych lub jej struktura. Dzięki zebranym informacjom jest się w stanie ustalić strategię dalszego ataku na sysem. Drugim podtypem jest Union-based SQLi. Polega on na wykorzystaniu polecenia UNION w SQL, aby połączyć wyniki kilku zapytań w jeden. Jest on wyjątkowo niebezpieczny, ponieważ można tym sposobem zdobyć wszystkie przechowywane informacje.

Drugim sposobem jest Inferential SQLi. Tutaj napastnik nie widzi bezpośrednio wyników zapytań, a obserwuje zachowanie bazy danych. Można na tej podstawie odtworzyć jej structure. W przypadku metody Boolean-based blind SQLi wykorzystuje się zapytań w celu uzyskania odpowiedzi *True* lub *False*. W przypadku Time-based blind SQLi zmusza się bazę danych do odczekania pewnego czasu zanim wyśle odpowiedź. Sugeruje to czy baza zwróciła wynik *True* lub *False*. Można w tej metodzie odgadnąc np. nazwę tablicy literka po literce, jednak jest to czasochłonne.

Out-of-band SQLi jest ostatnim sposobem SQL injection i nie posiada żadnych podtypów. Nie używa się tutaj żadnego kanału w celu uzyskania wyników, a zmusza się aplikacje do przesyłania odpowiedzi do własnego kontrolowanego punktu końcowego.

Zabezpieczyć sie można przed tym poprzez:

- segregację danych,
- rzutowanie danych,
- prepared statements (ang. Gotowe zapytania).

Segregacja danych polega na decentralizacji bazy danych. W ten sposób dane nie są przechowywane w jednej baize danych, co utrudnia uzyskanie jej zawartości. Rzutowanie danych (konwersja) to przekształcenie danych na docelowy format. Oznacza to, że w przypadku gdy użytkownik będzie chciał wpisać w polu tekstowym liczbę, jest ona zamieniana na inny typ danych. Pozwala to uniknąć niektórych ataków SQL injection. Prepared statements oznacza, że aplikacja używa gotowych części zapytań SQL, zamiast generować je pojedyńczo. Polega to na rozdzieleniu instrukcji na polecenie oraz argumenty, które zostają dodane do niego później. Zmienne wtedy są zastępowane znakami zapytania. Ten sposób chroni przed większością ataków na bazę danych, jednak wadą jest pogorszona optymalizacja aplikacji.

Cross-Site Scripting (XSS)

W odróżnieniu od SQL injection, ta metoda atakuje przede wszystkim klienta korzystającego z aplikacji. Atak polega na wstrzyknięcoi do przeglądarki ofiary fragmentu kodu języka skryptowego, np. JavaScript. W efekcie można wykonać ten kod w przeglądarce ofiary.

Napastnik może wykorzystywać tę metodę w celu:

- wykradania plików cookies,
- dynamicznej podmiany zawartości strony,
- ruchomienia keyloggera w przeglądarce,
- hostowania aplikacji malware.

Sposób można podzielić na dwie kategorie:

- Persistent (lub stored) XSS,
- Reflected XSS.

Persistent XSS jest najbardziej niebezpieczna. Polega na umieszczeniu kodu javascript na serwerze. Może to wykradać np. pliki cookie administratora. Przykładowo, atakujący umieszcza skrypt w komentarzu pod postem na blogu, który jest wysyłany do moderacji. Gdy moderator odczytuje komentarz skrypt jest uruchamiany I wykradane są określone dane. Taka sytuacja miała miejsce w systemie WordPress. Reflected XSS polega na zaszyciu skryptu w linku, który jest przesyłany do ofiary. Po kliknięciu tego linku klient łączy się z aplikacją przekazując jej fragment HTML zawierający wcześniej umieszczony skrypt, który jest wykonywany w przeglądarce.

Można się bronić przed tym atakiem filtrując dane przesyłane przez użytkownika, używając znaków kontrolnych. Inną metodą jest zastosowanie parametru HttpOnly. Flaga HttpOnly blokuje próby odczytu cookie z tą flągą przez API inne niż HTTP.

Podsumowanie

Zagadnienia związane z bezpieczeństwem i prywatnością w obliczeniach w chmurze obejmują szeroki zakres technologii, procedur i narzędzi mających na celu ochronę danych użytkowników. Szyfrowanie danych, zarządzanie tożsamością i dostępem, a także środki obronne przeciwko różnym metodą ataków, takim jak phishing, brute force czy XSS, stanowią kluczowe elementy w budowaniu solidnych systemów ochrony danych w chmurze.

Ważne jest, aby zarówno dostawcy usług chmurowych, jak i użytkownicy byli świadomi zagrożenia kradzieży informacji. Edukacja w zakresie bezpieczeństwa cybernetycznego, stosowanie silnych haseł, regularna aktualizacja oprogramowania i wiedza o zagrożeniach są kluczowe dla zapewnienia bezpieczeństwa w chmurze.

Ochrona informacji w chmurze ma kluczowe znaczenie zarówno dla jednostek, jak i dla firm i instytucji. Utrata danych lub ich nieuprawnione ujawnienie może prowadzić do poważnych konsekwencji finansowych, reputacyjnych i prawnych, dlatego należy podejmować wszelkie możliwe środki, aby dbać o ich bezpieczeństwo.

Źródła internetowe:

1. https://pl.wikipedia.org/wiki/Chmura_obliczeniowa
2. https://www.microsoft.com/pl-pl/security/business/security-101/what-is-cloud-security
3. https://www.politykabezpieczenstwa.pl/pl/a/szyfrowanie-w-chmurze
4. https://www.netia.pl/Netia/media/Netia/dokumenty-do-pobrania/Raport-Chmura-Cyberbezpieczenstwo-2021.pdf#page=28&zoom=100,0,0
5. https://fordata.pl/szyfrowanie-danych-w-chmurze-co-warto-wiedziec-zanim-wybierzemy-dostawce/
6. https://pl.wikipedia.org/wiki/Uwierzytelnianie_wieloskładnikowe#Przypisy
7. https://pl.wikipedia.org/wiki/Pojedyncze_logowanie
8. https://boringowl.io/blog/jwt-klucz-do-bezpieczenstwa-w-aplikacjach-internetowych
9. https://www.ekransystem.com/pl/blog/co-to-jest-ueba
10. https://www.keepersecurity.com/pl_PL/resources/glossary/what-is-identity-as-a-service/
11. https://pl.m.wikipedia.org/wiki/CASB
12. https://pl.m.wikipedia.org/wiki/Role-based_access_control
13. https://www.gov.pl/web/baza-wiedzy/czym-jest-phishing-i-jak-nie-dac-sie-nabrac-na-podejrzane-widomosci-e-mail-oraz-sms-y
14. https://pl.wikipedia.org/wiki/Atak_brute_force
15. https://pl.wikipedia.org/wiki/Atak_słownikowy
16. https://pl.wikipedia.org/wiki/DDoS
17. https://pl.wikipedia.org/wiki/Atak_man_in_the_middle
18. https://nordvpn.com/pl/blog/sql-injection-co-to/
19. https://sekurak.pl/czym-jest-xss/
20. https://www.blog.omegasoft.pl/jak-zaszyfrowac-dane/

Dominika Fergisz, Maja Jaszowska, Mateusz Fesz, Piotr Laskowski, Filip Skawiński
SKNI KOD

Dr inż. Bartosz Trybus
Opiekun Koła Naukowego

Rola badań użytkowników w procesie projektowania aplikacji

Streszczenie

Przedmiotem artykułu jest omówienie roli użytkowników w procesie projektowania aplikacji, a także wszelkich innych produktów. Wyjaśniono pojęcia user experience oraz user interface. Wymienio najważniejsze i najpopularniejsze narzędzia stosowane w trakcie procesu projektowego, a także wytłumaczono w jaki sposób są one wykorzystywane we współczesnym podejściu.

Słowa kluczowe: użytkownik, projektowanie, ux, ui, aplikacja.

1. Wprowadzenie

UX (User Expirience) oraz UI (User Interface) uchodzą za pojęcia, które od jakiegoś czasu zaczęły opanowywać szeroko pojęty świat nowych technologii. Można powiedzieć, że wpływ na to miały przede wszystkim wysoka konkurencyjność oraz ogromna ilość produktów na rynku. Zaspokajanie potrzeb użytkownika zaczęło liczyć się o wiele bardziej niż kiedykolwiek wcześniej. Z tak wielką dostępnością produktów na rynku, pojawiło się również przekonanie, że kupowane są one oczami, zarówno w kontekście przedmiotów użytku codziennego, jak również i aplikacji. Jednak, nie tylko wygląd okazał się ważny. Wraz z nim idą uczucia użytkownika, którymi również można go ”kupić”. Można to zobrazować na prostym przykładzie. Posłuży do tego telefon komórkowy, który stał się nieodzownym elementem codziennego życia. Kupujący cieszy się z nowo zdobytego przedmiotu, ale emocje można nasycić jeszcze na etapie samego rozpakowywania, co stało się również popularne na portalach społecznościowych, w formie krótkich nagrań. To, że znajduje się w atrakcyjnym pudełku, idealnie dopasowanym do rozmiaru smarfona, zabezpieczenia dają się łatwo usunąć, a także nie ma konieczności mocowania się z produktem, aby móc wyciągnąć go z opakowania. Każde nasze działanie zostało dokładnie opracowane przez wyspecjalizowaną grupę już wcześniej, aby uczynić podróż użytkownika jak najbardziej przyjemną.

2. UX vs UI

Definicje UX oraz UI bywają mylące, zdarza się również, że traktowane są one jako wspólna całość. Nie do końca tak jest. Oczywiście, zazwyczaj idą one ze sobą w parze, jednakże oznaczają coś całkowicie innego.

UX, czyli User Experience, możemy przetłumaczyć na projektowanie doświadczeń użytkownika. Nazwa ta niestety odbiega od rzeczywistości. Każdy człowiek jest inny, ma inne spojrzenie na świat i inaczej odbiera bodźce. Ciężko więc jest zaprojektować wspólną, jedyną słuszną drogę, a nawet narzucić jakieś doświadczenie dla tak różnorodnej grupy odbiorców, jest to wręcz niemożliwe.

Czym zatem zajmuje się projektowanie UX? Wpływaniem na to doświadczenie. Prawie każda podróż użytkownika na swojej drodze zawiera tak zwane Pain Pointy, czyli momenty trudne, nieprzyjazne, budzące negatywne emocje. Należy więc je zminimalizować, aby miały jak najmniejszy wpływ na końcowy odbiór całego produktu. Trzeba je odnaleźć, zrozumieć, zidentyfikować, a także zaproponować rozwiązanie.

Co ważne, UX zawsze działa na ścisłych danych, polega na ich analizie. Nie ma w nim miejsca na gdybanie i zastanawianie się. Jest to najprostsza droga do porażki.

Od czego w takim razie mamy UI? Czy nie ma on już związku z użytkownikiem i jego emocjami? Oczywiście, że ma, ale oddziałuje na nie w zupełnie inny sposób. UI Designer zajmuje się tym wszystkim, co jest graficzne, bo UI jest niczym innym jak projektowaniem interfejsów użytkownika. Dla UI designera liczą się obecne trendy, technologie i umiejętne wykorzystywanie GUI (graficznego interfejsu użytkownika).

W jaki sposób zatem oddziałuje na emocje użytkownika? Poprzez dobieranie odpowiedniej typografii, przestrzeni między elementami, zdjęć, a także przyciągania za pomocą ruchu i animacji.

Tutaj, bardzo ważny jest wpływ barw na percepcję człowieka, znany inaczej jako ”psychologia koloru”. Wiedza z tej dziedziny jest szeroko wykorzystywana przy projektowaniu aplikacji, budynków, czy kampanii reklamowych. Łącząc kolory z produktami, klienci tworzą pewną emocjonalną relację z produktem.

Emocje związane z kolorami pochodzą zarówno z podłoża biologicznego, jak i kultury. W ten sposób członkowie różnych cywilizacji mogą ująć ten sam kolor w różny sposób. Podczas gdy dla Europejczyków biel jest kolorem czystości, Japończycy postrzegają ją jako uosobienie śmierci. Duży wpływ mają również osobiste doświadczenia i preferencje jednostek.

3. Persona

Najważniejszą rzeczą w projektowaniu jest persona. Jej zidentyfikowanie to podstawa, bowiem ukierunkuje ona zespół do pracy na twardych danych. Należy poświęcić jej sporo czasu, bo to właśnie od niej zależeć będą działania nie tylko projektanta, ale również marketingu, technologii developmentu czy strategii sprzedaży. Nie obejdzie się bez danych

analitycznych, raportów, charakterystyki grupy osobowej, czy analizy ruchu w produkcie cyfrowym. Na ich podstawie możliwe jest stworzenie obrazu docelowego użytkownika, reprezentującego jego potrzeby, zachowania a także pozwalającego na nawiązanie empatii z obrazem odbiorcy. Zazwyczaj budowany jest on w sposób iteracyjny - użytkownicy aplikacji dzieleni są na mniejsze grupy dla których możliwe jest wyróżnienie reprezentantów. Proces ten można przedstawić w formie prostego algorytmu, reprezentującego nie tylko samą agregację danych, lecz także sposób ich wykorzystania w projekcie:

1. Zbieranie danych o użytkowniku docelowym
2. Określenie cech i różnic pomiędzy użytkownikami
3. Opracowanie hipotezy odnośnie wymagań użytkowników
4. Konsultacje opracowanych hipotez
5. Zbudowanie kilku person dla projektu, w zależności od zróżnicowanie grupy odbiorców
6. Nazwanie i opisanie każdej persony
7. Przedstawienie w jaki sposób mogą one chcieć korzystać z projektowanego produktu
8. Ponowne konsultacje przygotowanych założeń z zespołem
9. Zapoznanie zespołu z końcowym efektem
10. Nadzór nad poprawnym budowaniem i wdrażaniem scenariuszy użytkowania
11. Wprowadzanie bieżących modyfikacji wraz z rozwojem projektu

Przykładową personę możemy stworzyć dla prostego projektu aplikacji społecznościowej dla studentów. “Adam” może być 20 letnim mężczyzną, chcącym nawiązać znajomości wśród ludzi o podobnych zainteresowaniach, ale jednocześnie dopiero adaptującym się do nowego środowiska. Możemy sobie wyobrazić że wcześniej rzadko opuszczał rodzinną miejscowość, mieszka w akademiku i pracuje przez weekend. Jego cele i ograniczenia łączą się z przygotowanym scenariuszem - nie ma czasu na udział w imprezach i woli wieczorne rozmowy od weekendowych wyjść. Warto skojarzyć z utworzonym, mentalnym obrazem również cechy typowe dla każdej osoby - zdjęcie, wartości, zainteresowania, cele czy pewne wzorce zachowań, co pozwala uważać taką personę za bardziej żywą i zbliżoną do realnej osoby. W tym kontekście bardzo dobrze sprawdzi się stworzenie krótkiej historii życia - będzie to znacznie bardziej przystępne niż wypunktowanie cech czy jedynie potrzeb.

Po Personie utworzona zostaje Empathy Map, czyli mapa empatii, która z kolei ułatwia nakreślenie User Journey Map - mapy podróży użytkownika. Wiedząc jak przebiega ta podróż, pojawia się możliwość identyfikacji największych bolączek persony, co z kolei pomaga w stworzeniu możliwych scenariuszy, które pozwolą rozpocząć prototypowanie.

Klienci bardzo często nie są zainteresowani jej tworzeniem. Uważają, że to strata czasu i pieniędzy. Nic bardziej mylnego! Dobra persona jest gwarancją oszczędności czasu i pieniędzy, a także realizacji projektu zgodnie z określonymi wymaganiami.

4. Rozwiązania idealne

Rozwiązania idealne są mitem, nie istnieją i nie ma możliwości, żeby kiedykolwiek się pojawiły, dlatego słowo idealne należy zastąpić słowem optymalne. Co na to wpływa? Każdy projekt niesie za sobą pewne ograniczenia. Składają się na nie budżet, strategia marketingowa, wybór takiej, a nie innej technologii, umiejętności zespołu developerskiego. Cała sztuka polega na znalezieniu rozwiązania mądrego, zważając na przeróżne przeszkody.

Jednym z wielu problemów przed którymi musi stanąć projektant jest wielkość docelowej grupy odbiorców danego rozwiązania. Im ta jest bardziej zróżnicowana pod względem wieku, etniczności czy nawet upodobań i poglądów, tym większe jest ryzyko przedstawienia dalekiego od optymalności. Co więcej, w niektórych przypadkach odmienne oczekiwania mogą całkowicie uniemożliwić znalezienie kompromisu, zadowalającego wszystkich użytkowników. W omówionym wcześniej przykładzie Adama, musimy pamiętać że nie reprezentuje on wszystkich studentów z danej grupy wiekowej. Część z nich nie będzie zainteresowana tego typu aplikacją już na poziomie celu jej wykorzystania, a zatem najlepszym rozwiązaniem może być pominięcie tej grupy w definiowaniu oczekiwanych funkcjonalności oraz jako celu kampanii promocyjnych

Projektanci w krótkim czasie są w stanie stworzyć produkt oryginalny, wywołujący na początku nagły skok euforii. Niesie to jednak za sobą spore ryzyko klęski. Użytkownicy po wdrożeniu mogą nie być w stanie tego używać, uczyć się nowych rzeczy. Polegają na znanych im, wypracowanych wcześniej schematach. Rynek aplikacji na ten moment jest ogromny, przez co łatwo znaleźć alternatywę, a nowy produkt szybko zostanie wycofany.

5. Mapy empatii

Mapa empatii stanowi rozszerzenie dla podstawowych danych. Nazwa w tym przypadku okazuje się zobowiązująca - empatia jest zawsze na pierwszym miejscu, bowiem bez niej nie powstaną dobre rzeczy. Przepływ stanie się skomplikowany, interfejsy - nieużyteczne, wyroby- słabej jakości, a obsługa klienta - wręcz nietrafiona, czy nawet zła. Liderzy światowych przedsiębiorstw zdają sobie sprawę z tego, jak ważną umiejętnością jest wczuwanie się w rolę użytkownika, rozumienie jego potrzeb i emocji. Dowodem na to jest wpisanie empatii na listę najistotniejszych umiejętności, jakie będą wymagane od ludzi w przyszłości.

Największą zaletą Mapy Empatii jest prosta, komunikatywna, przekonująca forma. Zrozumienie jej celów, założeń, zalet, zakresów stosowania nie wymaga żadnej eksperckiej wiedzy. Nie wymaga także doświadczenia. Dzięki swojej użyteczności jest wykorzystywana na spotkaniach z klientami, jest codziennym narzędziem wykorzystywanym przez zespoły projektowe.

Szablon mapy empatii, jako metoda profilowania, skupia się na 4 istotnych kwestiach (z punktu widzenia użyteczności produktu cyfrowego). Wyróżniamy tu: potrzeby, doświadczenia, myśli, emocje. W jej centrum zawsze stoi użytkownik, czyli zdefiniowana wcześniej persona. Cztery pola pozwalają opisać, to co użytkownik w czasie badań: powiedział, zrobił, pomyślał, poczuł. O ile dwa pierwsze elementy nie budzą dużych problemów, o tyle kwestie myśli i emocji wymagają szczególnego wyczulenia na detal, umiejętności "czytania między wierszami", wyciągania wniosków o stanach emocjonalnych z różnych źródeł (np. mowa ciała). W przypadku Wypowiedzi ważne jest skupienie uwagi na słowach kluczowych, często powtarzających się, opiniach, które pokazują stosunek, na wypowiedziach oceniających. Obserwacja Zachowań powinna znaleźć odzwierciedlenie na Mapie Empatii w postaci opisów wymownych zachowań, postaw wyrażających się w zachowaniach. Pożądane okazują się również korzyści, pragnienia, potrzeby użytkownika, a także jego bolączki, obawy i frustracje.

6. User Journey Map

User Journey Map, czyli mapa podróży użytkownika pozwala na wizualną interpretację relacji użytkownika z produktem w odniesieniu do czasu i przestrzeni. Pozwala na jego zrozumienie, jednocześnie odpowiadając na pytania, jak i kiedy korzysta on z produktu. W jej skład wchodzą takie elementy jak persona, czas, miejsce oraz interakcje użytkownika z produktem. Bardzo ważne są również emocje odbiorcy.

Mapa podróży klienta skupia się na różnych etapach, emocjach, a także jego potrzebach. Elementy mapy mogą obejmować punkty kontaktu, emocje, cele, pytania oraz różne kontekstualne informacje. Jej tworzenie zwykle zaczyna się od zidentyfikowania różnych etapów, przez które klient przechodzi. Następnie analizuje się emocje i potrzeby w tych etapach oraz identyfikuje punkty kontaktu z produktem. Takie mapy pozwalają projektantom lepiej zrozumieć, jakie są potrzeby i oczekiwania użytkowników na różnych etapach korzystania z aplikacji. To umożliwia dostosowanie interfejsu, funkcjonalności i treści do realnych potrzeb użytkowników. To również narzędzie dynamiczne. Proces tworzenia mapy może być kontynuowany w miarę jak zmieniają się potrzeby i oczekiwania klientów oraz ewoluują produkty i usług, co pozwala na ciągłe doskonalenie doświadczenia klienta. W rezultacie, mapa

podróży klienta jest nieocenionym narzędziem dla firm i projektantów aplikacji, pomagającym lepiej dostosować swoje oferty i interfejsy do rzeczywistych potrzeb i oczekiwań użytkowników, poprawiając jakość obsługi klienta, zwiększając lojalność i przyczyniając się do sukcesu na rynku.

7. Prototypowanie

Czym jest prototypowanie? To nic innego jak projektowanie rozwiązań, które zaspokajają potrzeby użytkownika. Niestety, jego realizacja bywa uciążliwa. Aktualnie, świat projektowania zaopatrzony został w wiele rodzajów prototypów, a także ogromną ilość narzędzi do jego realizacji. Ponadto, walidacja, czyli ich testowanie, może przybierać przeróżne formy. Najczęściej jednak wybiera się najbardziej klasyczną formę prototypowania, czyli to na kartce papieru. W jego rysowaniu chodzi przede wszystkim o to, aby był efektywny - czytelny dla użytkownika, szybki w realizacji i pozwalający na skuteczne testowanie. Forma ta, najczęściej tworzona już jest w trakcie spotkania z klientem zamawiającym produkt. Umożliwia ona zarówno klientowi, jak i projektantowi, zrozumienie wzajemnych wymagań, dokładniejsze zdefiniowanie potrzebnych funkcji, a także przede wszystkim - uniknięcie nieporozumień już na najwcześniejszej fazie tworzenia produktu.

Tworząc prototyp papierowy, obie strony otrzymują możliwość zostania świadkami pojawiania się nowego wyrobu, budowania go z wypowiedzianych słów i powstania myśli. Daje on swego rodzaju zarys, szkielet, który będzie mógł zostać wypełniony w dalszym procesie budowania.

8. Scenariusze użycia

W zawodzie projektanta liczy się analiza i skrupulatność. Przed przystąpieniem do badania użytkownika, należy wcześniej przygotować scenariusze użycia. Jest to prosty krok, polegający jedynie na przygotowaniu poleceń dla badanego. Najlepiej zebrać je w formie listy i przekazać kopie obserwatorom. Dzięki temu będą mogli zanotować obserwacje dla każdego badania, ułatwi to też sortowanie, analizę i przechowywanie danych. Przy ich tworzeniu, należy pamiętać, że przejrzystość dokumentacji to podstawa! Scenariusze użycia powinny być krótkie, zwięzłe i oczywiste. Zbyt długie polecenia mogą zaburzyć proces badawczy, co ma odbicie w samych wynikach badania.

9. Heurystyki Nielsena

Mówiąc o doświadczeniu użytkownika, nie sposób nie wspomnieć o Jakobie Nielsenie, czyli jednym z najbardziej uznanych specjalistów w dziedzinie użyteczności, która stanowi

punkt wyjściowy do UX. W latach 90. stworzył on dziesięć pryncypiów, do których mimo upływu czasu nadal się wraca w procesie projektowania. Ich waga przestała już być tak duża jak kiedyś, ze względu na posuwającą się do przodu technologie, a także nowe rozwiązania. Jednakże, wypunktowane przez niego zasady, stanowią źródło tz. ‘dobrych praktyk’:

- pokazuj status systemu,
- zachowaj zgodność pomiędzy systemem a rzeczywistością,
- daj użytkownikowi pełną kontrolę,
- trzymaj się standardów i zachowaj spójność,
- zapobiegaj błędom,
- pokaż, zamiast zmuszać do pamiętania,
- elastyczność i efektywność,
- dbaj o estetykę i umiar,
- zapewnij skuteczną obsługę błędów,
- zadbaj o pomoc i dokumentację.

10. Badania

Przystępując do badania, należy kierować się dewizą - użytkownik ma zawsze rację. Produkt wykonywany jest dla niego i to właśnie on doświadczać może ewentualnych nieprzyjemności, związanych ze złym projektem, ale również ze specjalnymi wymaganiami Za każdym razem należy sprawdzać, czy próg wejścia nie jest dla niego zbyt wysoki, czy potrafi z niego skorzystać, a także czy coś go blokuje, zniechęca. Jeżeli pojawi się kilka wersji rozwiązania - należy przetestować je wszystkie, prawdopodobnie najbardziej pożądane okaże się to, które projektanci na wcześniejszym etapie chcieli odrzucić.

Co może posłużyć jako materiał do testów? Wszystko, co zostało stworzone. Prototyp papierowy, klikalny, czarno-biały, kolorowy, a nawet gotowy już produkt. W zależności od etapu zaawansowania projektu, specyfiki produktu oraz budżetu należy wybrać rodzaj testu. Jaki test jest najlepszy? Ten, który w efektywny sposób dostarcza niezbędnych informacji.
Testy dzielą się na dwie grupy: ilościowe i jakościowe. Testy ilościowe realizowane są na większej grupie. Nie zwraca jednak uwagi na wiele szczegółów. W tym teście liczą się statystyki, nie zagłębia się w emocje użytkownika. Testy jakościowe natomiast bardzo mocno skupiają się na człowieku, analizie jego emocji i zachowaniach niewerbalnych. Taki test nie sprawdza się zawsze, nie wszystkie produkty wymagają tego typu sprawdzenia.

Jak może wyglądać przykładowe badanie? Potencjalny użytkownik może być zaproszony osobiście. Przygotowany dla niego zostaje scenariusz użycia, na podstawie którego

będzie testował aplikację. Bardzo często scenariusz odczytywany jest przez jedną z osób prowadzącą badania. Na podstawie poleceń głosowych, osoba badana wykonuje bądź wskazuje kolejne kroki potrzebne do zrealizowania celu np. zaloguj się, dodaj nowy przedmiot do koszyka z kategorii ubrania, sfinalizuj zakup, wyloguj się. Polecenia wydawane są kolejno, a w międzyczasie badany jest bacznie obserwowany. Powszechnym jest prowadzenie takiego badania przez kilka osób, jedna wydaje polecenia, druga zwraca uwagę jak długo wykonywane są zadania, czy badany z łatwością odnajduje się na stronie, czy ma jakieś problemy, kolejna zaś przygląda się i próbuje odczytać jego emocje.

11. Podsumowanie

Bez przeprowadzenia wcześniejszych badań, ryzyko tego, że tworzona aplikacja czy jakikolwiek inny produkt, zakończy się fiaskiem znacząco rośnie. Konkurencja na rynku każdego roku przybiera na sile, co daje potencjalnemu klientowi większą możliwość na zdobycie narzędzia alternatywnego. Dlatego też, tworząc, należy zwrócić największą uwagę na użytkownika, dla którego produkt ten będzie przeznaczony. Przygotowanie wyrobu idealnie sprecyzowanego pod kątem technicznym, opartego na najnowszych technologiach nie gwarantuje sukcesu. Istnieje spore ryzyko, że nie zdobędzie on swojego własnego grona odbiorców. Twórcy nie muszą zgadzać się z ich zdaniem, ale to właśnie oni dają im możliwość zarobku na tych właśnie wyrobach.

Literatura

1. Badura C., *UXUI. Design Zoptymalizowany.* Helion 2022
2. Mościchowska I., Rogoś-Turek B., *Badania jako podstawa projektowania User Experience*, Wydawnictwo Naukowe PWN, 2020.
3. Krutysza E., *Wykłady*, uzyskane podczas Best Design Week, Warszawa 2023.

Źródła internetowe

1. https://thestory.is/pl/journal/mapa-empatii/ (dostęp: 11.06.2024).
2. https://www.interaction-design.org/literature/topics/personas (dostęp: 12.06.2024).
3. https://www.pr.fanfar.pl/aktualnosci/customer-experience-i-mapapodrozy-odkryj-klienta-na-nowo/ (dostęp: 11.06.2024).

Mateusz Skali, Karol Michoński, Łukasz Michnik, Szymon Jabłoński, Maciej Nabożny
SKINI KOD

Opiekun naukowy
dr inż. Bartosz Trybus

Działanie i zastosowanie Stable Diffusion

Streszczenie

Model generatywny Stable Diffusion jest jednym z najnowszych i najszybciej zyskujących popularność narzędzi w dziedzinie sztucznej inteligencji. Jego głównym zadaniem jest generowanie obrazów na podstawie opisów tekstowych. Pozwala to na tworzenie realistycznych obrazów i grafik. Zasada działania tego modelu opiera się na dwóch warstwach kluczowych: „Latent Space" i „Self Attention Layer". Pierwsza warstwa odpowiedzialna jest za przygotowanie zdjęcia przy pomocy dodawania i odejmowania szumu w celu wyszukiwania konkretnych cech obrazu. Druga zaś odpowiada za interpretację i przypisanie tekstu odpowiednim elementom na zdjęciu. Technologia ta znajduje szerokie zastosowanie w różnych branżach. Dzięki temu, iż jego kod jest otwarty i umożliwia modyfikację GUI (ang. graphical user interface), jest dostępny zarówno dla profesjonalistów, jak i amatorów, co dodatkowo przyczynia się do jego rosnącej popularności.

Wprowadzenie Stable Diffusion do procesów twórczych otwiera nowe możliwości w zakresie automatyzacji i kreatywności. Artystom i projektantom umożliwia szybsze realizowanie swoich wizji, podczas gdy firmy mogą wykorzystać go do tworzenia unikalnych kampanii reklamowych i materiałów marketingowych. Co więcej, dzięki zaawansowanym algorytmom, model ten potrafi identyfikować wzorce i generować obrazy, które są nie tylko estetyczne, ale również spójne z dostarczonymi opisami.

Mimo licznych zalet, wykorzystanie Stable Diffusion niesie ze sobą także pewne wyzwania. Obejmują one kwestie etyczne, takie jak ochrona praw autorskich i prywatności, a także dbałość o odpowiednie wykorzystanie wygenerowanych treści. Ważne jest, aby użytkownicy tej technologii byli świadomi potencjalnych zagrożeń i stosowali ją w sposób odpowiedzialny.

Słowa kluczowe: model generatywny, Stable Diffusion, sztuczna inteligencja, warstwy.

1. Wprowadzenie

Stable Diffusion jest jednym z najnowszych osiągnięć w dziedzinie sztucznej inteligencji. Model ten został opracowany w 2022 roku przez zespół badaczy z Maximilian University w Monachium oraz Heidelberg University w Heidelbergu. Aktualne poczynania oraz rozwój projektu można śledzić za pomocą udostępnionego publicznie repozytorium github[1].

W artykule zostanie przedstawiona zasada w jaki sposób algorytm wykorzystywany w Stable Diffusion jest w stanie zbudować zdjęcia na podstawie tekstu wprowadzonego przez użytkownika modelu.

[1] https://github.com/Stability-AI/generative-models (dostęp: 13.06.2024)

Metody badawcze obejmują przegląd literatury naukowej i branżowej, analiza przypadków zastosowania Stable Diffusion, a także omówienie etycznych i prawnych aspektów związanych z tą technologią. Przeprowadzona analiza pozwoli na lepsze zrozumienie roli Stable Diffusion w procesach twórczych oraz identyfikację korzyści i wyzwań związanych z jego użytkowaniem.

Technologia ta rewolucjonizuje sposób, w jaki tworzymy i odbieramy wizualne treści w cyfrowym świecie. Dzięki zaawansowanym algorytmom, model ten potrafi generować obrazy spójne z dostarczonymi opisami, co pozwala artystom i projektantom na szybsze realizowanie swoich wizji, a firmom na tworzenie unikalnych kampanii reklamowych i materiałów marketingowych. Jednakże, pomimo licznych zalet, wykorzystanie Stable Diffusion wiąże się również z pewnymi wyzwaniami, takimi jak ochrona praw autorskich i prywatności, a także odpowiedzialne użytkowanie generowanych treści.

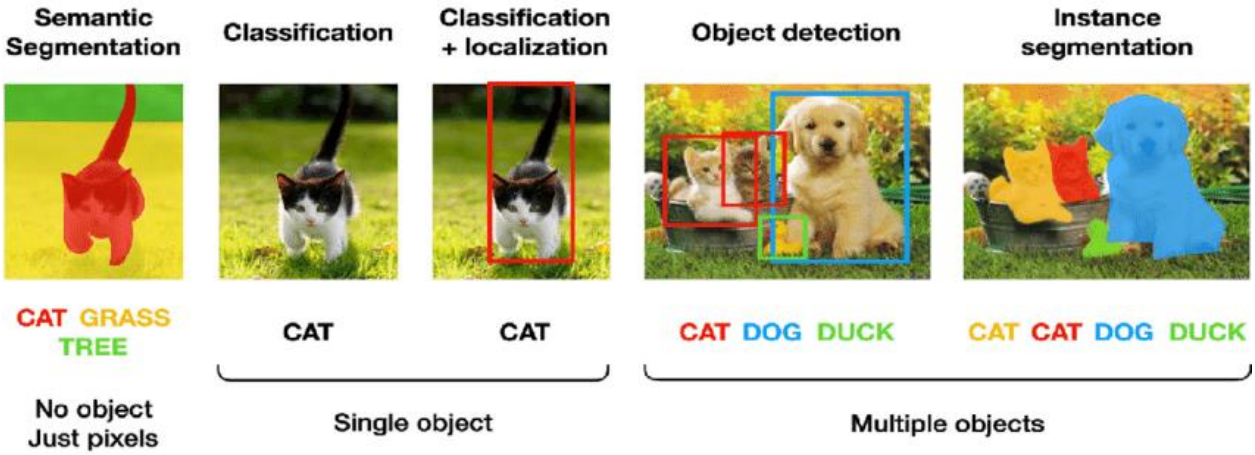
W artykule przyjrzymy się szczegółowo zasadzie działania oraz roli Stable Diffusion w procesach twórczych, analizując metody, narzędzia i technologie wykorzystywane w tym obszarze. Omówimy również praktyczne zastosowania tego modelu, jego korzyści oraz związane z nim wyzwania. Ostatecznym celem jest zwiększenie zrozumienia i świadomości dotyczącej tej innowacyjnej technologii, która ma potencjał zmienić sposób, w jaki tworzymy i odbieramy wizualne treści w cyfrowym świecie.

2. U-Net

Wykrywanie obiektów w dziedzinie sztucznej inteligencji można podzielić na kilka kategorii.

Kategorie zostały wymienione poniżej (Rysunek 1.):

- klasyfikacji – rozpoznanie obiektu, brak informacji o położeniu
- klasyfikacji wraz z lokalizacją obiektu – rozpoznanie obiektu wraz z jego lokalizacją
- klasyfikacji obiektów wraz z lokalizacja – rozpoznanie wielu obiektów z ich lokalizacją
- segmentacji semantycznej – każdy piksel ma przypisaną etykietę opisującą co przedstawia
- segmentacji instancji – każdy piksel ma przypisaną cechę opisującą co przedstawia, dodatkowo kolejne obiekty są numerowane. (wiele tych samych obiektów jest rozróżnianych)

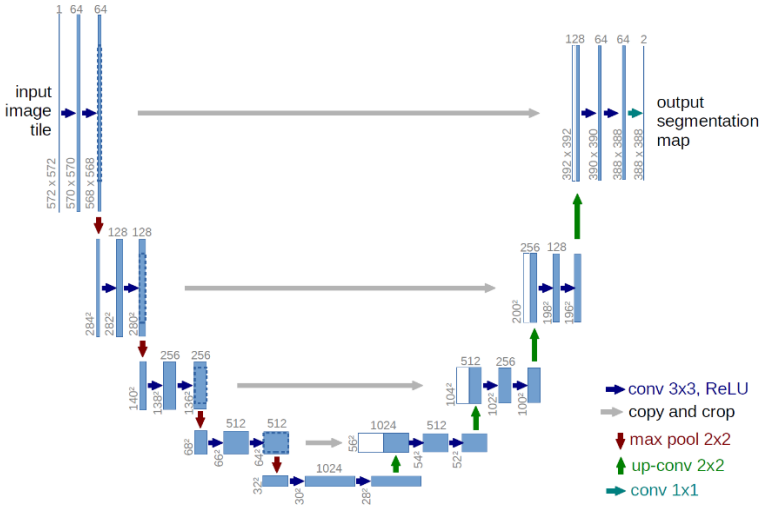


Rysunek 1. Klasyfikacja i detekcja obiektów na obrazach w algorytmach sztucznej inteligencji
Źródło: https://blent.ai [2] (dostęp 13.06.2024)

Stable Diffusion wykorzystuje segmentację semantyczną oraz segmentację instancji. Jest to niezbędne ze względu na to, że model może podmienić bądź usunąć dane piksele wypełniając je tłem bądź innym obiektem.

Przez wiele lat uważano, iż nie jest efektowne przetwarzanie zdjęć, gdyż wymagało to bardzo dużej ilości próbek szkoleniowych. Grupa naukowców w 2015 roku stworzyła sieć U-Net, która była znacząco szybsza i pozwalała na stworzenie modelu przy pomocy o wiele mniejszej ilości danych treningowych. Sieć ta jest opisana w opracowaniu „U-Net: Convolutional Networks for Biomedical Image Segmentation”[3], której wpływ znacząco wpłynął na rozwój sztucznej inteligencji. Głównym założeniem U-Net’u jest wydobycie wszystkich cech z obrazu. Odbywa się to na zasadzie zastosowania macierzy z sieci CNN oraz stopniowe zmniejszanie rozdzielczości obrazu. Na każdym kroku zwiększa się ilość kanałów (początkowo może być ich 3 – RGB), skutkując zwiększeniem informacji o tym co znajduje się na obrazie. Gdy sieć osiągnie zakładaną ilość kanałów (np. 1024) przechodzi w tryb zwiększania rozdzielczości otrzymanych wyników, zmniejszając ich ilość kanałów. Sieć pobiera informacje z procesu zmniejszania rozdzielczości łącząc je z otrzymanymi obrazami w celu powrotu obrazu do pierwotnej rozdzielczości. W wyniku tego otrzymanym obrazem jest maska identyfikująca jednoznacznie położenie każdego piksela danej cechy na obrazie.

[2] https://blent.ai/blog/a/unet-computer-vision (dostęp: 13.06.2024)
[3] https://lmb.informatik.uni-freiburg.de/Publications/2015/RFB15a/ (dostęp: 13.06.2024)



Rysunek 2. Wizualizacja sieci U-Net
Źródło: https://lmb.informatik.uni-freiburg.de [4] (dostęp 13.06.2024)

3. Proces uczenia modelu Stable Diffusion oraz jego zasada działania

Ze względu na specyfikację modelu zastosowanie klasycznych sieci takich jak w pełni połączone sieci stosowane np. w propagacji wstecznej czy też inne, które łączą wszystkie neurony z warstwy poprzednika z warstwą następnika nie odniosło by oczekiwanych skutków. Jest to spowodowane tym, iż obrazy, które wykorzystuje sieć posiadają ogromną ilość informacji takich jak każdy piksel na płaszczyźnie oraz jego kolor. Dlatego w procesie uczenia używana jest sieć CNN (ang. Convolutional Neural Network), która może połączyć tylko małą część neuronów z warstwy wejściowej z warstwą ukrytą. Działanie CNN opiera się na relacji między wieloma pobliskimi pikselami na obrazie. Wykorzystywane jest LRF (ang. Local Receprive Fields), co umożliwia detekcję konkretnej tej samej cechy na obrazie przez wszystkie neurony w warstwie ukrytej. Zastosowanie tych technologii pozwala na wydobycie z obrazu cech za pomocą relacji między pobliskimi pikselami.

Przykładowo dla obrazu o wymiarach 100px na 100px z jednym kanałem koloru (czarno – biały obraz) sieć w pełni połączona musiała by posiadać 100 milinów połączeń między neuronami. Jest to nadmiarowością, gdyż nie istnieje potrzeba wykorzystania relacji każdego piksela z ze sobą. CNN korzystając z macierzy 3x3 bądź 5x5 przechodzi przez cały obraz mnożąc kolejno wartości i zapisując je. Korzystając z macierzy 5x5 otrzymane zostanie na wyjściu 25 wartości, które można wykorzystać zamiast 100 milionów połączeń.

Stable Diffusion kompresując zdjęcia do tzw. „Latent Space” za pomocą enkodera. Następnie przy użyciu U-Net’u wraz z CNN, LRF wprowadza w każdej kolejnej warstwie U-

[4] https://lmb.informatik.uni-freiburg.de/Publications/2015/RFB15a/ (dostęp: 13.06.2024)

Net'u szum korzystając z SDE (Stochastic Differential Equations). Na ogół SDE korzysta z wzoru[5]:

$$dx = f(x,t)dt + g(t)dw$$

Gdzie:

$f(x,t)dt$ – jest funkcją wektorową (ang. drift coefficient)

$g(t)$ – jest funkcją o wartościach rzeczywistych, zwaną współczynnikiem dyfuzji

dw – jest nieskończenie małym białym szumem

Sieć jest uczona na różnej ilości szumu dodanego do obrazów, dzięki czemu może zaklasyfikować obiekty na bardziej lub mniej zaszumionych obrazach. Do określenia ilości dodanego szumu w obrazie używa się kodowania pozycyjnego (ang. positional encoding), który zamienia wartości dyskretne (w przypadku szumu będzie to określenie pozycji sekwencji szumu) na wartości ciągłych wektorów. Kodowanie pozycyjne można zastosować przy użyciu wzoru[6]:

$$PE(pos, 2i) = \sin(pos/100000^{2i/d_{model}})$$

$$PE(pos, 2i+1) = \cos(pos/100000^{2i/d_{model}})$$

gdzie:

pos – pozycja

i – wymiar

d_{model} – wymiar modelu uczącego

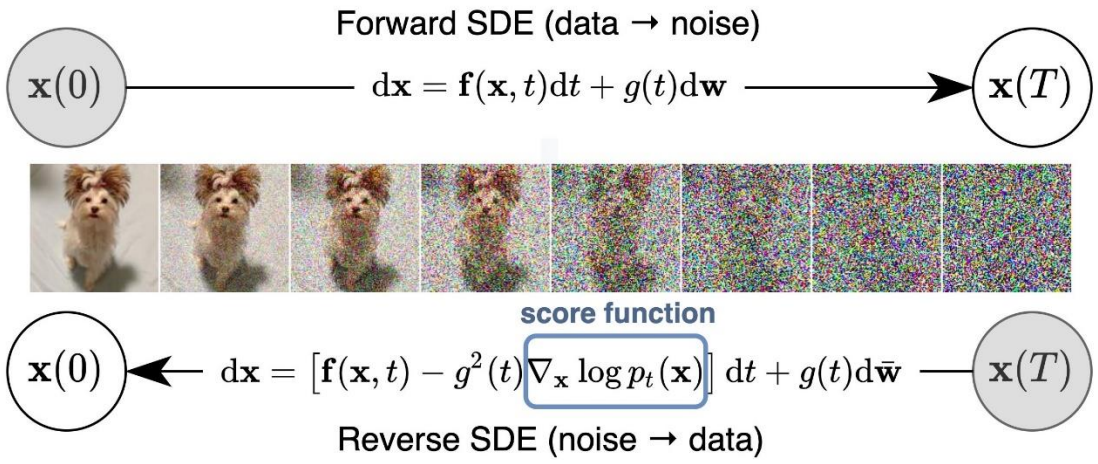
Po zamianie wartości wektor ten każdorazowo dodawany jest w warstwie U-Net przy zmianie rozdzielczości. Gdy obraz osiągnie minimalną rozdzielczość oraz największą ilość kanałów w sieci U-Net przechodzi w zwiększanie rozdzielczości zdjęć w którym również zachodzi proces zmniejszania szumu. Wzór używany w tym procesie odwracania szumu SDE[5]:

$$dx = [f(x,t) - g^2(t)\nabla_x log p_t(x)]dt + g(t)dw$$

dt – reprezentuje nieskończenie mały krok czasu, ponieważ SDE trzeba rozwiązywać wstecz w czasie.

Proces dodawania szumu do obrazu oraz jego odejmowanie przy użyciu powyższych wzorów zostało przedstawione na rysunku 3.

[5] https://yang-song.net/blog/2021/score/ (dostęp: 13.06.2024) źródło oryginalne: Reverse-time diffusion equation models B.D. Anderson. Stochastic Processes and their Applications, str. 313-326. Elsevier. 1982.
[6] https://arxiv.org/pdf/1706.03762 (dostęp: 13.06.2023)



Rysunek 3. Proces dodawania szumu i odszumiania w Stable Diffusion
Źródło: https://yang-song.net/blog/2021/score/ (dostęp 13.06.2024)

Aby rozpoznawać tekst wprowadzony przez użytkownika wykorzystuje się kodowanie pozycyjne wraz z wektorami przechowującymi wartości słów. Kodowanie pozycyjne stosuje się w celu uniknięcia zamiany słów przez model, co mogło by skutkować błędnym rezultatem. Jednym z najpopularniejszych rozwiązań jest „word2vec". Metoda ta posiada dwie identyczne listy z każdym słowem w języku angielskim oraz przynależnym mu wektorem. Między listami tworzone są relacje, które przyporządkowują słowa na bazie tego jak często są ze sobą stosowane (jeżeli wektor w pierwszej liście jest podobny do wektora w drugiej to słowa są ze sobą stosowane w innym przypadku nie są). Wielkości wektorów w danej liście oznaczają jak bardzo słowa są ze sobą spokrewnione, co jest zobrazowane na stronie https://projector.tensorflow.org/ (dostęp: 13.06.2024). Przykładową wizualizacją użycia wektorów może być równanie wektorów dla słów: „London" – „England" + „Japan" = „Tokyo". Metoda ta wykorzystywana jest w warstwie określanej jako „Self Attention Layer". Wyciąga ona cechy z wprowadzonych słów z wykorzystaniem relacji między słowami zależnej od ich wektorów. Warstwa ta tworzy połączenia między wektorami słów za pomocą mnożenia macierzowego określając wpływ danego słowa na inne, czy odwrotnie. Sprawia to, iż sieć może odczytywać negację danego słowa bądź np. okolicznika stopnia i miary.

4. Możliwość modyfikacji

Ze względu na otwarto-źródłowość Stable Diffusion użytkownicy korzystają z różnych metod, które mają na celu wzbogacenie i poprawienie działania modelu.

Embedding – metoda ta polega na osadzaniu udostępnianiu przez użytkowników małych reprezentacji wektorowych obrazów, które są dołączane do kodera tekstu modelu. Gdy dana nazwa zostanie wpisana w tekście, który generuje obraz, powoduje to zniekształcenie obrazu

w celu dopasowania do stylu wizualnego. Metoda ta umożliwia zastąpienie domyślnych biasów sieci, otrzymując unikalne style graficzne.

Hypernetwork – dla tej metody tworzy się małe sieci neuronowe w pełni połączone, które sterują modelem generowania tekstu, dostosowując podstawowe parametry modelu Stable Diffusion przez tworzenie nowych wag w sieci. Celem tej metody jest wpływ na styl generowanych obrazów (np. rysowanie postaci jako postacie z wody). W wyniku tej modyfikacji model naśladuje charakterystyczne style, których nie ma oryginalnie w danych treningowych.

Checkpoint model (Dreambooth) – modyfikacja ta polega na dostarczeniu przez użytkownika zestawu obrazów, które przedstawiają konkretną osobę lub koncepcję. Celem jest dostrojenie modelu i przeszkolenie na dodanych obrazach. Wynikowo model może bardziej dokładnie prezentować obiekty, które zostały dostarczone przez użytkownika.

5. Zastosowanie Stable Diffusion

Sztuka

Model ten ma zastosowanie w tworzeniu unikalnych dzieł sztuki na podstawie zadanych wzorców czy stylów, co może być wykorzystywane przez artystów i designerów w celu czerpania inspiracji.

Gry komputerowe i filmy

Użycie modelu może również dotyczyć generowania realistycznych tekstur, postaci czy środowisk, co znacznie redukuje czas i koszty produkcji w branży rozrywkowej. Może również przyspieszyć proces twórczy UI/UX.

Reklama i marketing

Tworzenie atrakcyjnych wizualizacji produktów, kampanii reklamowych czy materiałów promocyjnych, które są bardziej angażujące dla odbiorców. Obrazy wygenerowane przez Stable Diffusion mogą być wyidealizowane co przyciągnie potencjalnych nabywców.

Medycyna i biotechnologia

Przetwarzanie i analiza obrazów medycznych, takich jak MRI (rezonans magnetyczny) czy CT (tomografia komputerowa), co może pomóc w diagnozowaniu chorób i planowaniu leczenia. Ze względu na to, iż często zdjęcia z dziedziny medycyny były w ograniczonych ilościach, stworzono sieć U-Net, która miała umożliwić klasyfikację nawet na mniejszych zbiorach.

Edukacja i szkolenia

Tworzenie materiałów edukacyjnych, symulacji i wizualizacji, które pomagają w lepszym zrozumieniu skomplikowanych zagadnień.

Rekonstrukcja historyczna

Model może również odtwarzać historyczne miejsca, artefakty i sceny na podstawie dostępnych danych, co może być wykorzystywane w muzeach i wirtualnych wycieczkach.

6. Problemy oraz zagrożenia związane z Stable Diffusion

Jakość generowanych obrazów

Jednym z głównych wyzwań technicznych jest jakość generowanych obrazów. Mimo znacznego postępu w dziedzinie algorytmów generatywnych, obrazy wytworzone przez Stable Diffusion mogą czasami zawierać artefakty, które obniżają ich estetyczną wartość. Niedoskonałości mogą wynikać z niskiej jakości danych szkoleniowych, co może przedstawiać się w miernym odwzorowaniu tła, światła czy tekstur. Obrazy ze względu na proces uczenia w głównej mierze na rozdzielczości 512px x 512px mogą odbiegać jakościowo w innych rozdzielczościach, chociaż jest to mało prawdopodobne. Cechą charakterystyczną słabo wyuczonych modeli jest problem z odwzorowaniem palców człowieka.

Wydajność obliczeniowa

Stable Diffusion wymaga znacznej mocy obliczeniowej, co stanowi istotną barierę dla wielu użytkowników. Trening modeli oraz generowanie wysokiej jakości obrazów angażuje zasoby sprzętowe, które są kosztowne i trudno dostępne dla małych przedsiębiorstw oraz indywidualnych twórców. Obecnie, aby lokalnie włączyć model należy mieć co najmniej 4GB pamięci VRAM dostępnej na karcie graficznej.

Prawa autorskie i plagiat

Generowanie obrazów za pomocą Stable Diffusion może prowadzić do naruszenia praw autorskich. Modele często uczą się na ogromnych zbiorach danych, które mogą zawierać chronione prawem materiały. Przykładem mogą być podpisy autorów generowane przez sieci, które korzystały z próbek objętych prawami autorskimi. Wykorzystywanie takich obrazów bez odpowiednich licencji i zgód budzi poważne kontrowersje prawne i etyczne.

Generowanie nieodpowiednich treści

Technologia ta może być wykorzystywana do tworzenia treści nieodpowiednich, takich jak obrazy przemocy, nienawiści czy pornografii. Brak mechanizmów kontroli i moderacji treści

generowanych przez Stable Diffusion stanowi istotne zagrożenie dla społecznej odpowiedzialności technologii.

Wpływ na rynek pracy

Automatyzacja procesów kreatywnych za pomocą Stable Diffusion może prowadzić do zmian na rynku pracy. Zastępowanie pracy ludzkiej przez algorytmy może wpłynąć na zatrudnienie w sektorach związanych z grafiką, designem czy sztuką, co budzi obawy o przyszłość zawodów kreatywnych.

7. Podsumowanie

Artykuł przedstawia technologię Stable Diffusion, koncentrując się na jej działaniu, zastosowaniach oraz wyzwaniach. Stable Diffusion, model generatywny specjalizujący się w tworzeniu obrazów na podstawie opisów tekstowych, bazuje na zaawansowanych algorytmach, takich jak CNN. Dzięki temu modelowi możliwe jest generowanie realistycznych i szczegółowych grafik, co znajduje szerokie zastosowanie w różnych branżach, od sztuki po marketing. Pomimo licznych zalet, takich jak automatyzacja procesów twórczych i wsparcie dla artystów, technologia ta napotyka na istotne problemy techniczne, etyczne i społeczne. Wyzwania techniczne obejmują jakość generowanych obrazów i wysokie wymagania obliczeniowe. Z kolei problemy etyczne dotyczą ochrony praw autorskich i potencjalnego generowania nieodpowiednich treści. Społecznie, Stable Diffusion może wpływać na rynek pracy, zastępując tradycyjne role kreatywne. Artykuł podkreśla znaczenie odpowiedzialnego wykorzystania tej technologii, zwracając uwagę na konieczność świadomości i przestrzegania zasad etycznych przez użytkowników. Należy zaznaczyć, iż firmy jak OpenAI pracują nad ujednoliceniem zamiany języka naturalnego oraz obrazów w dane wektorowe odzwierciedlające obydwa pojęcia jednakowo[7]. Oznacza to, że dany wektor może jednoznacznie określać tekst oraz jego wynik w postaci graficznej. Projekt ten można śledzić na repozytorium github[8]. Podsumowując wprowadzenie Stable Diffusion do procesów twórczych otwiera nowe możliwości, ale jednocześnie wymaga rozwagi i odpowiedzialności w jej zastosowaniu.

[7] https://openai.com/index/clip/ (dostęp: 12.06.2024).
[8] https://github.com/openai/CLIP (dostęp: 12.06.2024).

Źródła internetowe

https://openaccess.thecvf.com/content/CVPR2022/papers/Rombach_High-Resolution_Image_Synthesis_With_Latent_Diffusion_Models_CVPR_2022_paper.pdf (dostęp: 12.06.2024).

https://arxiv.org/pdf/1706.03762 (dostęp: 12.06.2024).

https://yang-song.net/blog/2021/score/ (dostęp: 12.06.2024).

Szymon Jabłoński, Mateusz Skali, Karol Michoński, Łukasz Michnik, Maciej Nabożny
SKNI KOD

dr inż. Bartosz Trybus
Opiekun Koła Naukowego

Implementacja interpretera z użyciem języka C++

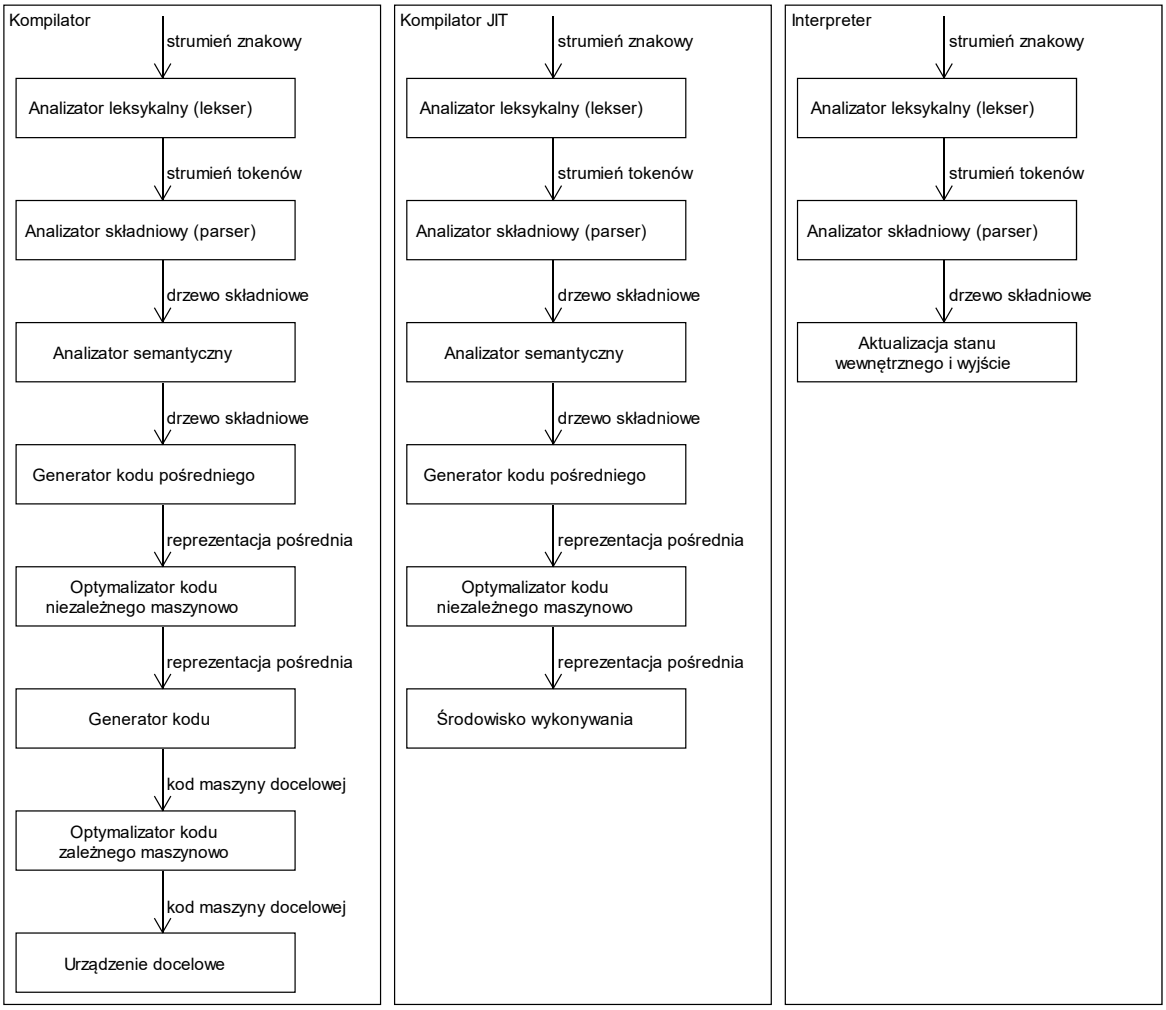
Streszczenie

Artykuł opisuje problematykę oraz implementację przykładowego interpretera skryptów z wykorzystaniem biblioteki standardowej C++. Autorzy omawiają podział języków ze względu na sposób wykonywania na urządzeniu użytkownika, wraz z ich wadami i zaletami. Następnie analizie poddano wymagania i wybory, które trzeba wziąć pod uwagę przy tworzeniu interpretera. Kolejne jest omówienie specyfiki implementacji takiego oprogramowania w wybranym języku, z uwzględnieniem wielu rodzajów struktur danych dostępnych w bibliotece standardowej, paradygmatów programowania, jak i funkcjonalności najnowszych standardów, mogących usprawnić i uprościć kod. Po tym poddany analizie zostaje proces decyzyjny podczas pisania, na podstawie projektu powstałego w ramach zajęć dydaktycznych z użycia C++. Posiada on własny język, stworzony od podstaw, z pewnymi niespotykanymi rozwiązaniami.

Słowa kluczowe: C++, interpreter, programowanie, polimorfizm, AST, kontenery, wskaźniki.

1. Wprowadzenie

W miarę rozwoju informatyki wyłoniło się wiele podejść do rozwiązywania problemu przetłumaczenia intencji programisty do formy zrozumiałej dla komputera. Przez lata powstało wiele języków o różnych składniach, funkcjonalnościach, poziomach abstrakcji, jak i wiele podejść do ich logiki. Można wśród nich wyróżnić różne kategorie, dzieląc ze względu na sposób działania, przede wszystkim: kompilowane do kodu maszynowego, kompilowane do kodu pośredniego oraz interpretowane.



Rysunek 1. Porównanie procesu przetwarzania kodu w kompilatorach i interpreterach.
Źródło: opracowanie z modyfikacjami na podstawie diagramu z „Kompilatory Reguły, Metody, Narzędzia”[1].

Pierwsze, takie jak C++ używają kompilatora do bezpośredniego wygenerowania kodu maszynowego, natywnego dla danego środowiska systemowego i sprzętowego. Pozwala to na tworzenie wysoko zoptymalizowanego oprogramowania, zapewniającego szybkie wykonywanie. Problemem jednak jest dystrybucja takiego oprogramowania, które wymaga osobnej kompilacji pod każdy system operacyjny i architekturę sprzętu, tak jak widać na rysunku 1. Programy takie nie pozwalają również na prostą modyfikację ich zachowania już po dystrybucji. Można również zaobserwować, że różne sposoby wykonywania różnią się kompleksowością rozwiązań używanych do obsługi kodu.

Drugie wytwarzają specyficzny dla języka kod pośredni, który przy uruchomieniu przez użytkownika poprzez użycie dedykowanego zestawu oprogramowania jest zamieniany do formatu odpowiedniego dla danego przypadku. Tu za przykładami są przede wszystkim C# i

[1] Aho V. A., Lam M. S., Sethi R., Ullman, J. D., „Kompilatory Reguły, Metody, Narzędzia”, Wydawnictwo Naukowe PWN, Warszawa 2019, strona 5.

Java. Osiągają bardzo podobne poziomy wydajności do poprzedniej kategorii, za cenę powolnego początku pracy. Wymagane jest uruchomienie ich kompilatora i przygotowanie kodu maszynowego. Dodatkowo wydajność działania często zwiększa się z czasem wykonywania, ponieważ wiele implementacji wdraża dokładniejsze optymalizacje dopiero w trakcie działania programu, szczególnie dla fragmentów o największej częstości wykonywania.

Ostatnie nie generują kodu, zamiast tego używają dedykowanego programu, który interpretuje bezpośrednio kod źródłowy. Pozwala to na olbrzymią elastyczność w użyciu za cenę obniżonej wydajności, szczególnie nadają się one do zarządzania wykonywaniem funkcji i programów o większej wydajności. Opisują to autorzy książki „Kompilatory Reguły, Metody, Narzędzia" w na stępującym cytacie:

„Interpretery często używane są do wykonywania języków poleceń, gdyż każdy operator w takim języku jest z reguły wykonaniem złożonej procedury, jak edytor lub kompilator. Podobnie niektóre „bardzo wysokopoziomowe" języki, jak APL, są zwykle interpretowane, ponieważ wiele informacji na temat danych, jak np. Rozmiar i struktura tablic, nie mogą być wyznaczone w czasie kompilacji"[2].

W tym wypadku najbardziej znanymi przypadkami są Python – używany do pisania programów, które co do zasady wykorzystują wywołania funkcji skompilowanych, napisanych w wydajniejszych językach, oraz Bash – służący do obsługi terminala systemowego w systemach opartych o jądro Linux.

Celem artykułu jest przeanalizowanie zasad działania oraz sposobu implementacji przykładowego[3] interpretera skryptów za pomocą języka C++.

2. Architektura programu

Interpreter ma przede wszystkim za zadanie przekonwertowanie wejścia od użytkownika albo z pliku na zrozumiały dla siebie format i jego wykonanie. Jest to główna funkcjonalność, na której opiera się reszta programu. Należy wyróżnić kilka najważniejszych punktów, które musi wykonywać dla każdej instrukcji:

- wczytać fragment z wejścia,
- podzielić rezultat na odpowiednie tokeny,
- zinterpretować z tokenów dane wyrażenie,
- wykonać odczytane polecenie.

[2] Aho V. A., Lam M. S., Sethi R., Ullman, J. D., „Kompilatory Reguły, Metody, Narzędzia", Wydawnictwo Naukowe PWN, Warszawa 2019, strona 3.
[3] https://github.com/FifthZoner/python-- (dostęp: 11.06.2024).

Początkowe wczytywanie z zasady odbywa się poprzez pobranie wejścia bezpośrednio od użytkownika poprzez interfejs interpretera lub przez plik zawierający kod do wykonania.

Dzielenie (lexing) jest o wiele bardziej skomplikowanym krokiem. Jego zadaniem jest podział wejścia na poszczególne tokeny według określonych warunków. Najprostszym sposobem jest użycie znaku spacji i zbudowanie tablicy z ciągami znaków. Jednak trzeba wziąć pod uwagę wiele innych aspektów. Niektóre języki, takie jak Python, używają indentacji do określenia miejsca, w którym kończy się dane wyrażenie, trzeba zapisać tę informację. Innym problemem jest usuwanie niepotrzebnych znaków i komentarzy, które programista mógł intencjonalnie lub nie pozostawić w kodzie. Do obsługi błędów może być również przydatna informacja o wykonywanym numerze linii i pliku, w którym się znajduje. Pozwala to na pokazywanie dokładnego miejsca, w którym wystąpił problem. Lexing staje się bardziej skomplikowany, w przypadku gdy trzeba wziąć pod uwagę przykładowo tekst w cudzysłowie, szczególnie gdy pewne znaki są zakazane i trzeba je zapisywać w odpowiedni sposób. W niektórych językach można również w tym miejscu wstawiać pewne zmienne do tekstu, co wymaga odpowiedniego przygotowania. Trzeba też sprawdzać, czy wartości liczbowe nie są ujemne i czy nie trzeba zapisać ich razem ze znakiem minus.

Parsowanie jest najbardziej skomplikowaną funkcjonalnością takiego programu. Ma za zadanie zrozumienie polecenia i przygotowanie go do wykonania. Można do tego podejść na wiele sposobów, przykładowo: działać na zasadzie swoistego drzewa decyzyjnego, w którym po danym słowie kluczowym oczekuje się danego konkretnego typu tokenu lub wartości, a pojawienie się nieoczekiwanego oznacza błąd. Innym podejściem jest zdefiniowanie konkretnych szablonów wyrażeń, do których próbuje się przypasować sekwencję tokenów i wartości.

Listing 1. Przykładowa gałąź drzewa AST w języku C.

```
// Sample AST Node Structure
typedef struct AstNode {
    NodeType type;
    union {
        int intval;
        char* strval;
        struct AstNode* child;
    } value;
    struct AstNode* next;
} AstNode;
```

Źródło: opracowanie na postawie listingu z książki „Compiler Construction with C Crafting Efficient Interpreters and Compilers”[4]

[4] Edet T., „Compiler Construction with C Crafting Efficient Interpreters and Compilers”, Wydawnictwo CompreQuest, 2024, strona 50.

Wyniki takiego działania z zasady zapisuje się w tzw. AST (Abstract Syntax Tree), czyli pewnej strukturze, która definiuje strukturę logiczną danego wyrażenia. Jej przykładowa implementacja w języku C została przedstawiona na listingu 1. Zawiera dane o danym tokenie, jego typie, kolejnej gałęzi oraz dodatkowe dane specyficzne dla danego typu tokenu. W języku C++ można ją zmodyfikować za pomocą polimorfizmu, nadając każdemu typowi tokenu osobny zestaw wartości, bez możliwości dostępu do tych odpowiednich dla innych.

Alternatywnie możliwe jest wykonywanie z jego pominięciem z użyciem funkcji rekurencyjnych bądź odpowiedników iteracyjnych, można wtedy od razu przejąć kompetencje kolejnego etapu.

Ostatnie jest wykonywanie, dobrą praktyką jest wykonanie go dopiero po parsowaniu, co pozwala na uniknięcie zmian w wypadku błędu w innej części wyrażenia. Aktualizowany jest stan wewnętrzny interpretera.

Są to najbardziej podstawowe funkcjonalności wymagane do działania interpretera. Jednak aby był on bardziej funkcjonalny, trzeba wprowadzić również możliwość zapisywania w pewnym formacie poprzednio podanych linii, na poczet wyrażeń takich jak pętle, czy definiowane przez programistę funkcje.

Do zapisywania można podejść na dwa sposoby. Pierwszym jest wstępne parsowanie, pozwalające na szybkie konwertowanie danych w wykonywalne polecenia po uzupełnieniu wartości i innych elementów dynamicznych. Alternatywnie można też zostawić wejście w stanie początkowym. Daje to potencjalnie możliwość uzależnienia działania funkcji od kontekstu wywołania Jest to również podejście łatwiejsze w implementacji, ponieważ może korzystać z tych samych funkcji, co wejście oryginalne.

3. Specyfika implementacji w C++

Ze względu na wybrany język programowania do wykonania projektu trzeba zastanowić się nad praktycznym sposobem realizacji logiki. Ze względu na użyteczne funkcje biblioteki standardowej, przykładowo do sprawdzenia, czy ciąg znaków zaczyna się danym ciągiem, użyty został standard C++20. Zastosowane zostały tylko funkcje zawarte w bibliotece standardowej, co pozwala na kompilacje bez zmian w kodzie w dowolnym kontekście.

Wybrany język dyktuje również sposób realizacji struktur danych, przede wszystkim z powodu występowania polimorfizmu, który jest alternatywą dla struktur ze zmienną określającą typ danego obiektu. Obydwa podejścia są poprawne, a drugie jest potencjalnie wydajniejsze, jednak ma również potencjał do bycia mniej czytelnym za względu bardziej skomplikowane funkcje wymagane do ich obsługi.

Kolejna jest kwestia sposobu dodawania kolejnych zagnieżdżonych obiektów, które można umieszczać poprzez użycie wskaźników. Jednym podejściem jest użycie „tradycyjnych" wskaźników z manualnym zarządzaniem pamięcią poprzez alokację kolejnych. Kolejne to stworzenie pewnej struktury danych, w której zapisywane będą obiekty, najlepiej w taki sposób, aby ich adres nie zmieniał się przy dodawaniu kolejnych. Tu wskaźnik zawiera adres zapisanego elementu. Ostatnim o użytym w projekcie rozwiązaniem jest wprowadzony niedawno szablon w języku C++: std::unique_ptr<T> i jego inne warianty. Pozwala on na przydzielenie obiektu przy tworzeniu, a następnie automatyczne zwolnienie wykorzystanej pamięci przy wyjściu z kontekstu. Odbywa się to dzięki zastosowaniu idiomu RAII (Resource Acquisition Is Initialization), który zakłada wywołanie konstuktora w momencie tworzenia i destruktora przy końcu użycia. Zapewnia to, że nie będzie sytuacji, w której obiekt rodzic jest zwalniany, a obiekt dziecko zostaje przydzielony, powodując wyciek pamięci. Takie podejście znacznie zmniejsza ryzyko problemów z działaniem programu po długim czasie wykonywania i zapewnia wydajne zarządzanie pamięcią.

Jeszcze inną kwestią jest sposób przechowywania danych o zmiennych i funkcjach. Najprostszym rozwiązaniem jest użycie tablic, jednak trzeba tu pamiętać, że wydajność spada w nich mocno w miarę dodawania większej ilości elementów. Problem dodawania mogłaby rozwiązać lista, jednak dostęp odbywa się poprzez przejście po kolei przez indeksy i uniemożliwia optymalizacje dostępu takie jak przeszukiwanie binarne posortowanej tablicy.

Alternatywą są kontenery std::map<T> i std::unordered_map<T>. Pierwszy jest implementacją czarno-czerwonego drzewa, czyli samo balansującego się drzewa binarnego. Zapewnia czas dostępu O(log n), z dodatkowym kosztem rebalansu przy zmianach. Takie rozwiązanie sprawdza się najlepiej przy dostępie sekwencyjnym do kolejnych elementów, a więc można założyć, że nie jest to idealna struktura dla dostępu do zmiennych i funkcji.

W przeciwieństwie do pierwszego, drugi nie opiera się na drzewie. Zamiast tego używa hash table, co powoduje, że losowość elementów nie jest problemem. Dostęp odbywa się zazwyczaj w czasie O(1), w najgorszym przypadku O(n). Z zasady stały poziom czasu dostępu powoduje, że działanie opartego na nim interpretera powinno być przede wszystkim bardziej przewidywalne pod względem prędkości, a nawet wyróżniać się dobrą wydajnością przy bardzo skomplikowanych programach.

4. Struktura projektu

W projekcie, na podstawie którego omawiany jest temat, zaimplementowanych zostało wiele funkcjonalności, między innymi: zmienne numeryczne i tekstowe, wyrażenia logiczne,

pętle, funkcje wbudowane i definiowane samodzielnie, parser działań matematycznych z zachowaniem kolejności wykonywania i lokalność zmiennych.

Działanie rozpoczyna się jak w każdym programie w języku C++, poprzez funkcję główną. Poprzez argumenty argc i argv ustalane jest czy do wykonania został podany plik z kodem, czy też nie. Następnie wywoływana jest funkcja do czytania wejścia z podaniem odpowiedniego wariantu polimorficznej struktury zwracającej kolejne linie. Dzięki temu nie ma z perspektywy reszty programu żadnej różnicy w formacie danych.

Wykonywanie odbywa się poprzez wczytywanie kolejnych linii, wywoływanie funkcji do usuwania znaków białych z obu stron linii i następnie odbywa się sama interpretacja. Jeżeli w danym przypadku zakończy się sukcesem, to linia jest dodawana do tablicy dynamicznej z wszystkimi innymi. Pozwala to na odwoływanie się do nich w pętlach i funkcjach.

W tej implementacji kolejne jest wykrywanie czy dane wejście jest jednym z kilku słów kluczowych, wyrażeniem warunkowym albo pętlą. Jeżeli wykryte zostanie wyrażenie albo funkcja, interpreter wchodzi w stan zapisywania linii. Zaczyna się od zapamiętania linii początkowej i kończy dopiero po napotkaniu odpowiedniej ilości słów „end”, równej ilości wykrytych. Same wyrażenia wykonywane są dopiero po wczytaniu całych, a w przypadku funkcji po zakończeniu definicji i wywołaniu. Brane pod uwagę jest również czy obecnie wykonywana jest funkcja. Do przechowywania danych o funkcji używany jest stos. Każdy element zawiera osobne zmienne lokalne i dodatkowo tablicę „poziomów”, które reprezentują kolejne poziomy indentacji, pozwalając na tworzenie zmiennych, które przestają istnieć po opuszczeniu danego wyrażenia.

Nareszcie, po wywołaniu funkcji do leksowania odbywa się rozbicie linii na tokeny. Na początku rozdzielane od reszty są znaki specjalne, z uwzględnieniem cudzysłowów. Są tu też wykrywane podwójne operatory, takie jak „+=”. Następne jest wyszukiwanie liczb ujemnych, odbywające się poprzez sprawdzenie miejsc, w których występuje kombinacja operatora, po którym występuje znak minus, a po nim wartość możliwa do przekonwertowania w liczbę. W takim przypadku przerwa po znaku jest usuwana. Przedostatnim krokiem jest rozbicie na tablicę tokenów, gdzie spacja jest znakiem rozdzielającym. Omijane są pozycje pomiędzy cudzysłowami, z wyłączeniem tych poprzedzonych znakiem „\”. Zakańczającym krokiem jest modyfikacja tablicy w przypadku operatorów podwójnych. Powoduje to zmiany takie jak przykładowo: n += 1 → n = n + 1. Pozwala to zmniejszyć poziom skomplikowania późniejszych komponentów interpretera i nie zmienia zachowania programu.

Samo interpretowanie kodu odbywa się poprzez uporządkowanie danych w polimorficznym drzewie, z różną zawartością w zależności od danego wariantu. Ustalanie, który wykorzystać

odbywa się poprzez znalezienie pasującego wzorca całego wyrażenia, a następnie przekazywanie reszty tokenów do konstruktorów kolejnych obiektów. W przypadku zestawów zwracających wartości używana jest dedykowana funkcja do obsługi matematyki. Bierze ona pod uwagę nawiasy, kolejność działań i funkcje.

Na tym etapie odbywa się również wyłapywanie błędów. W przypadku gdy pojawi się nieoczekiwany symbol, wywoływana jest funkcja z wyjątkiem i wiadomością w konsoli opisującą rodzaj problemu. Wyjątki powodują przerwanie wykonywania danej linii i w większości przypadków pozwalają na uniknięcie modyfikacji stanu wewnętrznego programu przez niewłaściwą instrukcję.

Ostateczne zatwierdzenie zmian odbywa się, jeżeli nie pojawiły się wcześniej żadne błędy, a może jeszcze zostać anulowane w przypadku znalezienia innego problemu i na tym etapie. Odbywa się poprzez przejście przez całe drzewo z użyciem funkcji do wykonywania, które zwracają różne dane w zależności od wariantu.

Dodatkowym niuansem jest obsługa pętli i funkcji. Wymagają one zapisywania kodu do późniejszego wykonania. W tym wypadku jest to realizowane poprzez zapisanie numeru linii, w której zaczyna się dane wyrażenie i wykonywanie kodu aż do polecenia kończącego. Kod nie jest zapisywany w postaci parsowanej, ponieważ sama struktura jest w tu zależna od obecnego stanu wewnętrznego interpretera, mianowicie od zadeklarowanych zmiennych i funkcji.

5. Składnia obsługiwanego języka

Na potrzeby projektu powstał prosty język skryptowy. W miarę postępu był sukcesywnie poszerzany o dodatkowe funkcjonalności. Przede wszystkim, aby umożliwić programiście stworzenie jakiegokolwiek programu, trzeba było wprowadzić obsługę zmiennych i wejścia/wyjścia. Zmienne posiadają dwa typy: tekstowe i numeryczne, których nie można bezpośrednio ze sobą mieszać.

Tabela 1. Składnia poszczególnych poleceń stworzonego języka

Składnia poleceń	
typ	forma
Deklaracja zmiennej	(implicit) string/num ([]) <nazwa> = <stała/zmienna/działanie>
Zmiana wartości zmiennej	<nazwa> ([<indeks>]) <=/+=/-=/*=/^=//=> <stała/zmienna/działanie>
Wywołanie funkcji	<nazwa> (<argumenty po przecinkach>)*
Definicja pętli	while <warunek> ... end
Definicja if	if <warunek> ... (else ...) end
Definicja switch	switch <wartość> case <wartość> ... end
Definicja funkcji	void/string/num <nazwa> (<argumenty po przecinkach>) ... end*

Deklaracja argumentu	(implicit) string/num <nazwa>
Zwracanie wartości	return <wartość>
Wielkość tablicy	<nazwa>:size
Modyfikacja tablicy	<nazwa>:clear()/push(<wartość>)/pop()/resize(<wartość>)*

Źródło: opracowanie własne na podstawie zawartości omawianego projektu, zawartość nawiasów nie jest wymagana poza przypadkami oznaczonymi *

Tabela 1. przedstawia wszystkie możliwe wzorce poleceń, możliwe do wywołania w tym języku. Pozwalają one na rozwiązanie szerokiej gamy problemów z użyciem tylko takich skryptów. Dodatkowo istnieje nieduża biblioteka standardowa, używana tak samo, jak definiowane funkcje, dzięki polimorfizmowi. Zawiera zestawy funkcji do:

- wypisywania do konsoli,
- wczytywania od użytkownika,
- sprawdzania typów,
- zmiany typów,
- operacji matematycznych,
- wykonywania poleceń systemu operacyjnego,
- obsługi plików tekstowych.

Taka biblioteka znacznie rozszerza możliwości programisty. Szczególnie ważna jest tu interakcja z konsolą systemu, co pozwala na wykonywanie wielu działań nieuwzględnionych w funkcjach. Można przykładowo wykonywać skrypty instalacyjne programów, czy też skrypty terminalowe z uwzględnieniem wyborów użytkownika.

6. Najtrudniejsze zagadnienia

Co oczywiste, tworzenie interpretera od podstaw jest złożonym procesem. Zapewnienie poprawnej współpracy między wszystkimi elementami, wraz ze spójnym sposobem użycia jest dużym wyzwaniem. Szczególnie trudne jest napisanie logiki odpowiedzialnej za obsługę funkcji i wyrażeń logicznych. Wymagają one uwzględnienia stanu wewnętrznego, w przeciwieństwie do innych, prostych poleceń. Konieczna jest implementacja stosu wywołań funkcji, jak i możliwości zapisywania pewnych danych pozwalających na wykonanie wielokrotnie pewnego fragmentu w pętli.

Problematyczne jest również zaimplementowanie całości w taki sposób, aby dodawanie kolejnych funkcjonalności nie stanowiło problemu i nie wymagało głębokich zmian w kodzie źródłowym. Szczególnie przemyślane musi być typowanie. W wypadku omawianego wyżej przypadku zakodowane statycznie zostały dwa typy danych. Implementacja kolejnego wymagałaby bardzo dużej ilości pracy. Dobrym pomysłem jest tu zaimplementowanie

otwartego systemu obsługi typów, jednak wymaga to większego wysiłku, w porównaniu do niewielu, predefiniowanych.

Kolejną kwestią jest sprawdzanie błędów. Wymaga to bardzo dużej ilości sprawdzeń poprawności tokenów i bardzo możliwe jest pominięcie pewnych przypadków. Może to powodować modyfikację stanu wewnętrznego przez niepoprawne polecenia.

7. Podsumowanie

Przedstawione podejście do wykonywania programów daje wielką elastyczność w użyciu. Pozwala na bardzo szybkie zmiany w programie, nie ma potrzeby przygotowywania specjalnego środowiska programistycznego, aby dokonać jakichkolwiek zmian, i nie ma potrzeby dystrybucji dodatkowego pliku wykonywalnego razem z programem. Interpretery idealnie sprawdzają się jako wygodny sposób zarządzania programami niższego poziomu.

Implementacja własnego rozwiązania tego typu jest dobrym sposobem na zdobycie doświadczenia w pisaniu skomplikowanych programów. Podczas rozwiązywania problemów związanych z tematem można również lepiej zrozumieć proces działania kompilatorów, które działają w dużej części na podobnych zasadach.

Literatura

1. Aho V. A., Lam M. S., Sethi R., Ullman, J. D., „Kompilatory Reguły, Metody, Narzędzia”, Wydawnictwo Naukowe PWN, Warszawa 2019.
2. Edet T., „Compiler Construction with C Crafting Efficient Interpreters and Compilers”, Wydawnictwo CompreQuest, 2024.

Źródła internetowe

1. https://github.com/FifthZoner/python-- (dostęp: 11.06.2024).

Łukasz Michnik, Maciej Nabożny, Karol Michoński, Szymon Jabłoński, Mateusz Skali
Studenckie Koło Naukowe Informatyków „KOD”

dr inż. Bartosz Trybus
Opiekun Koła Naukowego

Znaczenie języka Typescript w nowoczesnych aplikacjach webowych

Artykuł opisuje znaczenie języka TypeScript w nowoczesnych aplikacjach webowych na przykładzie biblioteki frontendowej React, podkreślając jego znaczącą rolę w tym środowisku. Szczegółowo omawia główne funkcje tego języka, takie jak statyczne typowanie zmiennych i funkcji, klasy, interfejsy, typowanie generyczne oraz usprawnione programowanie obiektowe. Przeanalizowane również zostały wady TypeScriptu, który przede wszystkim dodaje złożoności kodu, a takim działaniem skutecznie zniechęca nowicjuszy dopiero zaczynających swoją przygodę z tym językiem. Na koniec autor jasno przedstawia swoje stanowisko, które zostało wywnioskowane na podstawie wcześniejszej analizy, jednoznacznie określając znaczenie tego języka, który przede wszystkim zrewolucjonizował reużywalność komponentów frontendowych .

Słowa kluczowe: TypeScript, JavaScript, nowoczesne aplikacje webowe.

1. Wprowadzenie do języka Typescript

Typescript to potężne narzędzie dla programistów aplikacji webowych, jest językiem programowania opracowanym przez Microsoft w 2012 roku, jego twórcą jest Anders Hejlsberg. Stanowi nadzbiór (ang. superset) dla języka Javascript, określnie to oznacza rozwinięcie, wiąże się z tym szereg nowych możliwości oferowanych przez ten język, których nie ma w czystym JS. Każdy poprawny kod napisany w języku JS, jest równocześnie poprawnym kodem TS. W praktyce Typescript rozszerza możliwości o opcjonalne statyczne typowanie, nowe struktury danych, takie jak Enum i Klasy, oraz inne funkcje wymienione w dalszej części artykułu. Jest on językiem ciągle rozwijanym i posiadającym ogromne grono społeczności, która aktywnie rozwija możliwości o dodatkowe biblioteki ułatwiające pracę i efektywność pisania kodu. Typescript szybko zdobył popularność wśród deweloperów, którzy szukali sposobu tworzenia bardziej niezawodnego i łatwiejszego w pisaniu kodu JavaScript. Było to głównym celem powstania tego języka, ponieważ JS nie posiada systemu typów, co za tym idzie pisanie kodu większych projektów oraz w dużym gronie programistów bywało problematyczne. Język ten stał się popularny wśród deweloperów głównie dzięki swojej elastyczności, możliwościom oferowanym przez statyczne typowanie oraz łatwości integracji z istniejącymi projektami JS. Dzięki wspomnianemu statycznemu typowaniu, programiści mogą definiować typy zmiennych, funkcji,

parametrów i wartości zwracanych, co prowadzi do wczesnego wykrywania błędów i lepszego zarządzania kodem. W przeciwieństwie do JavaScriptu, gdzie typy są określane dynamicznie w czasie wykonywania programu, TypeScript pozwala na ich zdefiniowanie i sprawdzenie już na etapie kompilacji, co znacznie zmniejsza ryzyko błędów i ułatwia proces debugowania. Jedną z ważniejszych funkcji wprowadzonych przez TypeScript jest możliwość definiowania interfejsów, które pomagają w określeniu struktur obiektów. Interfejsy te są niezwykle przydatne w dużych projektach, gdzie współpraca między różnymi zespołami deweloperskimi jest kluczowa. Dzięki jasno określonym interfejsom, deweloperzy mogą z łatwością zrozumieć, jakie dane są wymagane przez różne części aplikacji, co ułatwia komunikację i zmniejsza ryzyko błędów. Ponadto, TypeScript wprowadza nowoczesne podejście do programowania obiektowego. Klasy, dziedziczenie, moduły i przestrzenie nazw są zaimplementowane w sposób, który jest bardziej zgodny z zasadami programowania obiektowego niż w czystym JavaScripcie. Modyfikatory dostępu, takie jak public, private i protected, pozwalają na lepszą enkapsulację danych i metod, co zwiększa bezpieczeństwo i organizację kodu. Wprowadzenie typów generycznych (ang. generic types) umożliwia tworzenie bardziej elastycznych komponentów przygotowanych pod ponowne wielokrotne użycie. Generyczność pozwala również na pisanie funkcji i klas, które mogą działać na różnych typach danych, co prowadzi do bardziej uniwersalnych rozwiązań. Silne wsparcie narzędzi, które TypeScript oferuje, jest kolejnym czynnikiem, który przyczynił się do jego popularności. Edytory kodu, takie jak Visual Studio Code dzięki dodatkowym rozszerzeniom, oferują zaawansowane funkcje autouzupełniania, podpowiedzi kodu, refaktoryzacji i nawigacji po kodzie, które są możliwe z pomocą informacji o dostępnych typach. To sprawia, że programowanie jest bardziej wydajne i mniej podatne na błędy. TypeScript jest również doskonale zintegrowany z nowoczesnymi frameworkami i bibliotekami, takimi jak Angular, React i Vue.js. Angular, na przykład, został stworzony z myślą o tym języku, co zapewnia pełne wykorzystanie jego możliwości typowania i narzędzi. React i Vue.js również oferują doskonałe wsparcie, umożliwiając deweloperom pisanie komponentów i aplikacji z większą pewnością co do poprawności kodu. Dzięki swojej zgodności z JavaScriptem, TypeScript pozwala na stopniowe przyjęcie w istniejących projektach, bez konieczności przepisywania całej bazy kodu od podstaw. To sprawia, że migracja do TypeScript jest łatwiejsza i mniej ryzykowna, co zachęca wielu deweloperów do korzystania z tego języka. W miarę jak TypeScript zdobywał coraz większą popularność, rozwinął się również jego ekosystem. Wiele popularnych bibliotek JavaScriptowych oferuje definicje typów dla TypeScriptu, co pozwala na pełne wykorzystanie możliwości typowania nawet przy użyciu zewnętrznych zależności. Aktywna społeczność deweloperów TypeScriptu dostarcza cennych

zasobów, takich jak dokumentacja, przykłady kodu i wsparcie techniczne, co dodatkowo ułatwia naukę i korzystanie z tego języka.

2. Możliwości języka Typescript

TypeScript wprowadza wiele dodatkowych funkcji i możliwości ponad standardowy JavaScript. W tej części artykułu zostaną omówione oraz przedstawione na przykładach najważniejsze z nich, przedstawione przykłady mają największe znaczenie podczas programowania aplikacji webowych. Poprawiają one bezpieczeństwo oraz wydajność kodu, oferują rówież niezastąpioną pomoc przy pracy nad dużymi projektami, gdzie pracuje na raz wielu programistów. Architektura omówionych przykładów będzie opierała się o bibliotekę frontendową – React.

2.1. Typowanie zmiennych

Główną cechą supersetu jest właśnie możliwość typowania wszystkich zmiennych, złożonych struktur danych takich jak obiekty, czy funkcji. Podstawowe statyczne typowanie zmiennych odbywa się za pomocą operatora „ : ”, listing 1 zawiera przykład typowania przykładowych zmiennych na różne typy.

```
const car: string = „audi”;
const engine: number = 4.2;
const accessories: boolean = true;
```

Listing 1. Przykład statycznego typowania zmiennych.

Źródło: Opracowanie własne

Typowanie struktur złożonych odbywa się za pomocą interfejsów, umożliwiają one definiowanie kształtów obiektów, co ułatwia pracę z bardziej złożonymi strukturami. Dodatkowo dzięki rozszerzeniom w edytorach, możliwe jest autouzupełnianie zmiennych pochodzących z obiektów zadeklarowanych z pomocą interfejsów. Przyśpiesza to znacząco pracę z tego typu zmiennymi.

```
interface Car {
   name: string;
   model: string;
   engine: number;
   accessories: boolean;
}
```

Listing 2. Przykład typowania obiektów z pomocą interfejsu.

Źródło: Opracowanie własne

TypeScript dodatkowo zawiera możliwość nadania kilku typów dla jednej zmniennej poprzez wykorzystanie operatora „ | ”. Jest to szczególnie pomocne w przypadku zmiennej, która nie jest dokładnie określona i może przyjać inny typ w zależności od konkretnego warunku. Wykorzystanie tej funkcjonalności zostało przedstawione na listingu 3.

```
const value: string | number
value = “Hello”;
value = 4;
```

Listing 3. Przykład zmiennej posiadające dwa typy.

Źródło: Opracowanie własne

Istnieje możliwość nadania wielu typów, wiecej niż dwa. Dodatkowo TypeScript posiada możliwość stworzenie własnego typu np. w przypadku elementu *select* na stronie internetowej można ustawić własny typ, który w tym przypadku określa wartość jaką może przyjąc zmienna. W przypadku wystąpienia innej wartości niż określona, kompilator zwraca błąd. Na listingu 4 widnieje przykład dla zmiennej obsługującej wiele wartości jakie może przyjąć.

```
const selectValues = “sortByName” | “sortByAge” | “sortByEngine”
```

Listing 4. Przykład zmiennej posiadającej własny typ.

Źródło: Opracowanie własne

Typowanie tablic polega na dokładnie tym samym co wcześniej opisane przykłady z jednym wyjątkiem, że w tym przypadku na koniec konieczne jest dopisanie nawiasów kwadratowych określających zmienną jako tablicę. Takie wykorzystanie przedstawione jest na listingu 5.

```
const numbers: number[] = [1, 2, 3, 4, 5];
const names: string[] = [“audi”, “mercedes”, “bmw”];
```

Listing 5. Przykład typowania tablic.

Źródło: Opracowanie własne

2.2. Typowanie funkcji

Typowanie funkcji w TypeScript umożliwia definiowanie zarówno typów parametrów, jak i typów zwracanych wartości. Dzięki temu można bardziej precyzyjnie określić, jakie dane mogą być przekazywane do funkcji i co funkcja powinna zwracać, to z kolei pozwala na wczesne wykrywanie błędów i zwiększa czytelność oraz utrzymanie kodu. Podczas przekazywania argumentów do funkcji, w nawiasach okrągłych zostaje określony typ wymaganych argumentów.

Po zamykającym nawiasie okrągłym znajduje się operator „ : ”, który określa zwracany typ z funkcji, w przypadku braku zwracanych wartości, należy w tym miejscy wpisać *void*. Określanie typów funkcji zostało przedstawione na listingu 6.

```
function add(a: number, b: number): number {
   return a + b;
}

console.log(add(2, 2)); //wypisane zostanie 4
console.log(add(2, "2")); //wystąpi błąd typów: Argument of type
„string” is not assignable to parameter of type "number"
```

Listing 6. Przykład wykorzystania typów dla funkcji.

Źródło: Opracowanie własne

Komponenty w bibliotece React opierają się na funkcjach, w związku z tym również wymagane jest ich typowanie. Jest to bardzo pomocna strona wykorzystania języka TypeScript w programowaniu aplikacji webowych, ponieważ w przypadku przekazania złego typu zmiennej do komponentu, mogą wystąpić w nim problemy, natomiast określenie typu od razu zwróci błąd. Określanie typów komponentów jest połączeniem wcześniej opisanych metod typowania. Na listingu 7 przedstawione zostało typowanie dla komponentu React. Komponenty przyjmują *props* jako obiekty, więc wymagane jest określenie jego typu za pomocą *interface*, w którym zostały zdefiniowane typy dla poszczególnych wartości. Operator „ ? ” jest dodatkowym operatorem dla języka TypeScript, który określa, że ta zmienna nie jest wymagana w kompnencie, co za tym idzie nie musi zostać użyta lub jej wartość domyślna jest już określona, w przykładzie widzimy watość domyślną – 15.

```
interface SpinnerProps {
   color: string;
   size?: number;
}

export default function Spinner({ color, size = 15 }:
SpinnerProps){
   return (
      <div>
         <PulseLoader color={color} size={size} />
      </div>
   );
}
```

Listing 7. Przykład wykorzystania typów dla komponentu React.

Źródło: Opracowanie własne

Wykorzystanie typowania dla komponentów biblioteki React jest niezwykle pomocne, usprawnia ono wykrywanie błędów na etapie kompilacji. Przy dużych projektach i wielu zespołach programistów ta właściwość jest niezastąpiona.

2.3. Typy generyczne (ang. generic types)

Typy generyczne w TypeScript to potężne narzędzie, które umożliwia tworzenie komponentów, funkcji, klas i interfejsów, które mogą działać z wieloma różnymi typami danych. Dzięki nim można pisać bardziej elastyczny i wielokrotnego użytku kod, który jest jednocześnie bezpieczny pod względem typów.

```
function car<T>(arg: T): T {
   return arg;
}

let output1 = car<string>("Audi"); // typ string
let output2 = car<number>(4.2); // typ number
```

Listing 8. Przykład wykorzystania typów generycznych.

Źródło: Opracowanie własne

Widoczny na listingu 8 przykład funkcji z wykorzystaniem typów generycznych, przyjmuje dwa typy, typ string lub typ number. Funkcja oczywiście może przyjąć dowolny, wybrany typ. Operator „ <> ” zawiera literę T, która według konwencji określa typ funkcji, zostaje on wpisany jako typ argumentu oraz typ zmiennej zwracanej przez tą funkcję. Istnieje również możliwość wykorzystania wielu typów generycznych w jednej funkcji.

2.4. Typ any, unknown oraz undefined

Każdy z tych typów ma swoje specyficzne zastosowania. Typ *any* jest używany, gdy programista chce całkowicie wyłączyć sprawdzanie typów dla danej zmiennej. Jest to przydatne przy migracji kodu lub integracji z nieznanymi typami, lub po prostu aby wyłączyć przy tej jednej konkretnej zmiennej właściwość TypeScriptu do typowania. *Unknown* jest bezpieczniejszą alternatywą dla *any*, wymuszającą sprawdzenie typu przed wykonaniem operacji na zmiennej. Jest używane, gdy otrzymane dane pochodzą z nieznanych źródeł. Przed użyciem wartości z typem *unknown* należy postawić warunek sprawdzający typ, bez tego mogą wystąpić błędy programu. *Undefined* jest używane, aby oznaczyć zmienne, które zostały zadeklarowane, ale jeszcze nie zainicjowane, lub w przypadku funkcji, które nie zwracają

wyraźnej wartości. Stosowanie odpowiednich typów w odpowiednich kontekstach pozwala na pisanie bardziej bezpiecznego i czytelnego kodu w TypeScript.

2.5. Klasy i modyfikatory dostępu

TypeScript wspiera nowoczesne podejście do programowania obiektowego, oferując klasy, dziedziczenie i modyfikatory dostępu, które nie są dostępne w JavaScripcie. Wprowadzone usprawnienia znacznie ulepszają sposób, w jaki możemy organizować i zarządzać naszym kodem podczas programowania obiektowego. Opisana sekcja nie jest zbyt powiązana z programowaniem aplikacji webowych, natomiast warto wspomnieć o tych możliwościach, ponieważ rewolucjonizują one programowanie obiektowe w języku JavaScript, któremu brakowało tych funkcji. TypeScript wprowadza pełną obsługę klas, zgodną z nowoczesnym JavaScriptem (ES6+), ale dodatkowo rozszerza ich możliwości dzięki typowaniu i innym zaawansowanym funkcjom. Dodatkowo wprowadzone modyfikatory dostępu, takie jak public, private, protected, readonly pomagają kontrolować widoczność i dostęp do właściwości i metod klasy. Tworzenie klas abstrakcyjnych oraz dziedziczenie, również znalazło miejsce w funkcjach TypeScriptu. Listing 9 przedstawia przykładową klasę abstrakcyją z której dziedziczy klasa Circle, widoczne jest wykorzystanie elementów wprowadzonych przez TypeScript.

```
abstract class Shape {
   abstract getArea(): number;
}

class Circle extends Shape {
   readonly radius: number;
   private color: string;
   public border: string;

   constructor(radius: number, color: string, border: string) {
      super();
      this.radius = radius;
      this.color = color;
      this.border = border;
   }

   public getArea(): number {
      return Math.PI * this.radius * this.radius;
   }

   public getColor(): string {
      return this.color;
   }
}
```

Listing 9. Przykład wykorzystania programowania obiektowego usprawnionego przez elementy TypeScriptu.

Źródło: Opracowanie własne

3. Zastosowanie w aplikacjach webowych

Zastosowanie TypeScript w aplikacjach webowych przynosi liczne korzyści, które wspierają rozwój, skalowalność i jakość kodu. Omawianym obszarem, w którym TypeScript znajduje szerokie zastosowanie, jest rozwój interfejsu użytkownika, czyli część frontendowa, natomiast jest również wykorzystywany we wszystkich obszarach, gdzie JavaScript, np backend – node.js i inne frameworki. Dzięki typowaniu statycznemu, programiści korzystający z popularnych bibliotek i frameworków frontendowych, takich jak React, Angular, Vue.js czy Svelte, mogą szybciej wykrywać błędy i zapewniać lepszą reużywalność dla komponentów. W miarę rosnącej złożoności aplikacji webowych, TypeScript pomaga w zarządzaniu kodem poprzez typowanie statyczne, co przekłada się na szybsze wykrywanie błędów. Bezpieczeństwo typów, czyli wprowadzenie typowania statycznego do JavaScriptu, jest kolejną zaletą tego języka. Dzięki niemu aplikacje webowe zbudowane w TypeScript są mniej podatne na błędy w czasie działania. TypeScript oferuje także bogate wsparcie dla narzędzi deweloperskich, co znacznie zwiększa produktywność programistów. Funkcje takie jak autouzupełnianie kodu, nawigacja po projekcie, refaktoryzacja i debugowanie są znacznie

bardziej zaawansowane w środowisku TypeScript niż w przypadku zwykłego JavaScriptu. Dodatkowo, typowanie statyczne ułatwia testowanie jednostkowe i zapewnienie wysokiej jakości kodu poprzez wykrywanie nieprawidłowości i potencjalnych problemów już na etapie pisania kodu. To również ułatwia współpracę w zespole deweloperskim poprzez jednoznaczną dokumentację struktury danych i interfejsów. Wszystkie te czynniki sprawiają, że TypeScript stał się nieodłącznym elementem procesu tworzenia nowoczesnych aplikacji webowych, przyczyniając się do ich szybszego, bardziej niezawodnego i łatwiejszego w utrzymaniu rozwoju.

4. Wady TypeScriptu

TypeScript zdobył ogromną popularność wśród programistów i jest szeroko stosowany w tworzeniu nowoczesnych aplikacji webowych, ale tak jak każdy język, nie jest on pozbawiony wad. Wprowadza on dodatkowy poziom złożoności w porównaniu do JavaScriptu. Nowicjusze mogą napotkać początkowe trudności w nauce TypeScriptu, zwłaszcza jeśli nie mają doświadczenia z innymi językami statycznie typowanymi. Muszą oni nauczyć się nowej składni oraz zrozumieć koncepcje takie jak typowanie statyczne, generyki, interfejsy, klasy oraz modyfikatory dostępu. Dodatkowo w przypadku osób, które początkowo pisały w podstawowym JavaScripcie mogą zastanawiać się nad celem takiego typowania, jednak takie myślenie jest błędne i każda osoba w trakcie nauki i wykorzystywania tego języka zrozumie swoją pomyłkę. Kolejną wadą jest fakt, że TypeScript w przeciwieństwie do JavaScriptu, który jest językiem interpretowanym, musi zostać skompilowany przed wykonaniem. Proces ten może znacząco wydłużyć czas budowania aplikacji, szczególnie w dużych projektach. Początkujący programiści mogą również mieć problemy z poprawnym skonfigurowaniem TypeScriptu w sposób optymalny i dostosowany do potrzeb projektu, proces ten zazwyczaj jest dosyć skomplikowany i problematyczny. Następną wadą jest brak wsparcia dla wszystkich narzędzi i bibliotek, które mogą być niekompatybilne. Chociaż wsparcie dla tego języka ciągle rośnie, czasem pojawiają się problemy z integracją w niektórych środowiskach, może to wymagać dodatkowej konfiguracji lub poszukiwania alternatywnych rozwiązań. Typowanie w TypeScripcie również czasami prowadzi do nadmiernego kodu, poprzez definiowanie typów dla skomplikowanych struktur danych tworzy zazwyczaj bardzo długie i złożone deklaracje. Sprawia to, że kod staje się trudniejszy do czytania i zrozumienia. Podsumowując TypeScript nie jest pozbawiony wad na które należy zwracać uwagę, początkujący programiści przechodzący z JavaScriptu na omawiany język nie mają łatwo na początku procesu nauki.

5. Podsumowanie

Podsumowując, TypeScript to nie tylko rozszerzenie JavaScriptu, ale pełnoprawny język programowania, który przynosi wiele korzyści w tworzeniu nowoczesnych aplikacji webowych. Jego rola w przyszłości będzie prawdopodobnie rosła, stając się standardem dla dużych i złożonych projektów webowych. W dzisiejszym świecie tworzenia aplikacji webowych, TypeScript oferuje znaczące korzyści, które przewyższają jego wady. Jego zdolność do wprowadzenia większej niezawodności kodu oraz wsparcia dla nowoczesnych funkcji językowych czyni go niezastąpionym narzędziem dla deweloperów. TypeScript nadal rozwija się dynamicznie, zyskując coraz większą popularność wśród programistów i organizacji. Microsoft oraz szeroka społeczność deweloperów nieustannie pracują nad jego udoskonalaniem, wprowadzając nowe funkcje i usprawnienia. Dzięki temu pozostaje jednym z najbardziej innowacyjnych i efektywnych narzędzi do tworzenia nowoczesnych aplikacji webowych. Pomimo początkowych trudności w procesie nauki tej technologii, warto zadać sobie trud i nauczyć się tego języka, ponieważ podczas tworzenia dużych i nowoczesnych aplikacji webowych język ten jest niezastąpiony.

Literatura

Vanderkam D., *Effective TypeScript. 62 Specific Ways to Improve Your TypeScript*, O'Reilly', Październik 2019.

Cherny B., *Programming TypeScript. Making Your JavaScript Applications Scale*, O'Reilly', Kwiecień 2019.

Źródła internetowe

https://www.typescriptlang.org/ (dostęp: 09.06.2024).

Karol Michoński, Mateusz Skali, Łukasz Michnik, Szymon Jabłoński, Maciej Nabożny
Studenckie Koło Naukowe Informatyków „Kod”

Dr inż. Bartosz Trybus
Opiekun Koła Naukowego

Optymalizacja procesu tworzenia stron internetowych przy pomocy technologii TailwindCSS

Streszczenie

Celem artykułu jest zbadanie wpływu biblioteki TailwindCSS na tworzenie stron internetowych. W artykule skupiono się na zbadaniu efektywności, responsywności, optymalizacji oraz prędkości z jaką można stworzyć aplikacje przy jego użyciu. Poruszono również jego historię oraz główną filozofie „Utility-First” technologii i jej wpływ na projektowanie UI. Następnie analizie poddano jak Tailwind skraca czas tworzenia stron internetowych, co uzyskano poprzez użycie reużywalnych klas. Sprawdzono też wpływ biblioteki na usprawnienie kodu oraz optymalizacje stron pod względem wyszukiwarek internetowych. Główną charakterystyczną cechą TailwindCSS jest prostota oraz łatwa implementacja responsywności poprzez budowanie strony zaczynając od widoku na telefony powoli dodając nowe style dla większych urządzeń.

Słowa kluczowe: TailwindCSS, UI, Utility-First, responsywność, SEO, Bootstrap

1. Wprowadzenie

Strona internetowa jest niesamowicie ważnym elementem prowadzenia swojej marki w dzisiejszych czasach, przez co ludzie tworzący witryny starają się usprawnić proces ich tworzenia. W artykule poddano analizie bibliotekę TailwindCSS oraz jak jej implementacja wpływa na proces budowania strony. Tailwind charakteryzuję się podejściem „Utility-First”, którego celem jest używanie gotowych klas odpowiadających za pojedyncze style. Technologia ta zapewnia wiele innowacyjnych rozwiązań, które powodują przyśpieszenie tworzenia strony internetowej oraz poprawia responsywność, przy czym usprawnia kod.

TailwindCSS jest frameworkiem CSS stworzonym przez dwóch pasjonatów programowania Steve’a Schogera i Adama Wathana, którzy na własne potrzeby eksperymentowali z stylowaniem CSS. W taki sposób stworzyli oprogramowanie by ułatwić i przyśpieszyć proces tworzenia aplikacji webowych. Razem z Steve’em Schogerem stworzyli narzędzie, dzięki któremu deweloperzy mogą szybko stylować swoje dzieła w prosty sposób, bez nadmiernego kodu CSS. Technologia ta wyróżnia się od innych przede wszystkim swoją konfiguralnością i możliwościa dopasowania jej do swoich potrzeb.

TailwindCSS konkuruje z innymi bibliotekami takimi jak Bootstrap, poprzez użycie wspomnianych wcześniej klas użyteczności, dzięki którym można dostosować stronę, co do najmniejszych szczegółów. W porównaniu do innych bibliotek Tailwind nie operuje na gotowych komponentach, lecz na modalnych stylach przygotowanych przez autorów. Przez takie podejście Framework używany jest przez deweloperów do projektów, które wymagają indywidualnego podejścia i spersonalizowanych komponentów.

Warto również wspomnieć o tym z jaką łatwością implementowana jest responsywność na stronach dzięki klasom responsywnym, które stosują style w zależności od szerokości urządzenia. Klasy te używane są poprzez prefixy takie jak „sm" lub „md". Tailwind zaprojektowany został w stylu mobile-first, co oznacza, że style początkowo ustawiane są dla urządzeń mobilnych i przez użycie wspomnianych wyżej prefixów, style aktualizowane są do wartości podanych przy prefixie dla odpowiadającej mu szerokości ekranu.

W artykule oceniono również wpływ frameworka na bezpieczeństwo oraz SEO strony, które jest nierozłącznym elementem tworzenia stron internetowych w obecnych czasach.

2. Instalacja i architektura TailwindCSS

Instalacja frameworka jest niesamowicie prosta, aby zainstalować go do swojego projektu wystarczy wykonanie kilku kroków, takich jak wpisanie komend „npm install -D tailwindcss" i „npx tailwindcss init" w konsoli, po czym wystarczy zaktualizować nowo powstałe pliki poprzez określenie ścieżek do szablonów i dopisanie dyrektyw „@tailwind".

TailwindCSS to framework CSS, którego głównym założeniem jest filozofia „utility-first", czyli pisanie predefiniowanych stylów w miejscu w którym inicjalizowane są komponenty HTML. Poprzez użycie takich klas pomocniczych tworzenie stron internetowych zostało niesamowicie uproszczone oraz przyśpieszone. Klasy te są niesamowicie proste i intuicyjne, przez co próg wejścia jest bardzo niski i nawet początkujący programista może zająć się tworzenie stylów z użyciem stylów w znacznikach HTML.

Podstawą TailwindCSS są klasy stylizujące, w których każda pojedyncza klasa odpowiada jednej właściwości, dzięki temu kod jest czytelny i łatwo jest w nim znaleźć element, który deweloper chce zmienić. Klasy obejmują szeroki zakres stylów zdefiniowanych przez autorów technologii, które można edytować. Technologia wspiera wiele klas, którymi można edytować stronę od marginesów, przez wysokość, po rodzaje i kolory czcionek i teł. Wszystkie predefiniowane klasy używane są w atrybucie „className".

```
<input type="text" />
<input type="text" className="m-4 bg-gray-300 h-12"/>
```

Rysunek 1. Kod obejmujący dwa inputy, gdzie jeden jest bez stylów, a drugi posiada klasy predefiniowane TailwindCSS.

Rysunek 2. Ukazane na stronie dwa inputy, z których jeden został wystylizowany frameworkiem TailwindCSS.

Rysunek 1 przedstawia kod, na którym widać z jaką prostotą implementuje się style dla elementów strony. Pierwszy input nie posiada żadnych stylów, za to na drugi nałożono margines 16 pikseli, tło o kolorze jasnoszarym oraz ustawiono wysokość inputa na wartość 48 pikseli. Jednostki używane w predefiniowanych klasach frameworka TailwindCSS to jednostki skalowania TailwindCSS, gdzie domyślnie jedna jednostka przedstawia 0.25 rem lub 4 piksele.. Powodów wybrania takiej jednostki jest wiele, lecz głównymi jest skalowalność, responsywność oraz łatwość utrzymania. Dzięki używaniu jednostek relatywnych, takich jak rem, łatwiej utrzymać projekt, który z czasem rośnie. Na rysunku 2 zaprezentowano wywołany kod na stronie. Jeśli programista chce ustawić jakiś styl na niestandardowe wartości to autorzy również o tym pomyśleli. W tailwindCSS, by ustawić własną wartość dla danego stylu, należy wpisać ją w nawiasach kwadratowych wraz z jednostkami wielkości. Pokazuje to jak proste jest stylizowanie komponentów przy pomocy tego frameworka.

3. Efektywność tworzenia projektów

Przez tempo w jakim rozwija się świat oraz technologia, procesy tworzenia stron internetowych muszą nadążać przez co prędkość ich tworzenia stale rośnie. TailwindCSS, dzięki swojemu podejściu “Utility-First”, oferuje narzędzia, które mogą znacząco przyspieszyć proces ich tworzenia, podnosząc przy tym jakość.

```
colors: {
  main: "#ff3000",
  mainHover: "#fb8c23",
  bgWhite: "#fcfcfc",
  bgWhiteHover: "#f4f2f0",
  bgDark: "#1f1b1a",
  bgDarkHover: "#282322",
  bgDarkHover2: "#383130",
},
```

Rysunek 3. Definicja własnych schematów kolorów znajdująca się w pliku tailwind.config.js.

Głównymi cechami technologii Tailwind są moduły i klasy reużywalne, które tworzone są przez połączenie wielu styli, które deweloper chce złączyć w jedną klasę przy pomocy dyrektywy „@apply”. Dodatkowo, Tailwind CSS umożliwia tworzenie własnych schematów kolorów w pliku „tailwind.config.js”, co widoczne jest na rysunku 3. przedstawia on kolory, które można zaaplikować w modułach w taki sposób, by po wpisaniu odpowiedniej klasy odwoływała się ona do kolorystyki, która nie była standardowo ustawiona w projekcie. Użytkownicy mogą również definiować własne klasy pomocnicze, które są specyficzne dla ich projektu, a następnie aplikować je do komponentów za pomocą tej samej dyrektywy „@apply”. Takie złączenie pozwala na wielokrotne używanie tych samych stylów na różnych komponentach, co znacząco przyśpiesza proces tworzenia strony i zachowuje spójność projektu.

Prędkość z tworzone są projekty jest związana z łatwością z jaką programista jest w stanie zaimplementować zmiany na stronie. Dzięki wspomnianym wcześniej klasom pomocniczym, deweloper znacznie przyśpiesza swoją pracę, poprzez brak potrzebny przepisywania dużych bloków kodu. Analiza czasu potrzebnego, by stworzyć stronę przy użyciu TailwindCSS, pokazuje, że z użyciem frameworka można znacząco skrócić czas potrzebny do stylizacji aplikacji webowej.

Tailwind sprzyja również optymalizacji kodu. Dzięki narzędziom takim jak PurgeCSS, które są częścią ekosystemu TailwindCSS, nadmiarowe style nałożone na komponenty są usuwane. Przez takie podejście zmniejszony jest rozmiar plików CSS, a strona ładuję się szybciej. Jest to niesamowicie ważne w kontekście wyda jności jak i optymalizacji pod kątem wyszukiwarek internetowych.

4. Optymalizacja i responsywność

Ten rozdział w artykule skupia się na responsywności oraz na możliwości optymalizacji kodu stron internetowych dzięki frameworkowi TailwindCSS. Dzięki swojej elastyczności i modułom, system oferuje wiele możliwości w tworzeniu stron automatycznie dopasowujących się do szerokości konkretnego urządzenia.

Optymalizacja kodu jest jedną z najważniejszych rzeczy w tworzeniu stron internetowych. TailwindCSS, dzięki użyciu klas pomocniczych, pozwala tworzyć zwięzły kod, który jest niesamowicie łatwy do pisania i przyjemny w sytuacjach, kiedy programista chce rozbudować stronę. Dzięki integracji w systemie wspomnianego w rozdziale o efektywności narzędzia PurgeCSS, projektant może na bieżąco usuwać niepotrzebne style, co wpływa na optymalizacje strony i jej wydajność.

Schemat działania PurgeCSS:

- Narzędzie przegląda pliki projektu takie jak dokumenty HTML oraz komponenty JSX i zapisuje używane selektory CSS.
- Następnym krokiem jest usunięcie z arkusza stylów wszystkich selektorów, które nie zostały oznaczone jako używane, przez co końcowy rozmiar pliku jest zmniejszony, a efektywność plików rośnie.

Efektem opisanych wyżej kroków jest finalny plik CSS, w którym znajdują się niezbędne deklaracje, co ma bezpośredni wpływ na poprawę wydajności strony internetowej.

Responsywność w dzisiejszych czasach jest podstawą dla tworzenia aplikacji webowych. Każda strona internetowa musi być funkcjonalna na szerokiej gamie urządzeń, od komputerów stacjonarnych po smartfony z bardzo małym wyświetlaczem. W TailwindCSS oferowany jest zestaw klas responsywnych, który ułatwia dopasowanie elementów strony w zależności od szerokości ekranu. Dzięki takiemu podejściu deweloperzy z łatwością mogą tworzyć estetyczne strony, w oparciu o układ interfejsu zaprojektowany przez projektanta, które są funkcjonalne na każdym urządzeniu.

Badanie możliwości optymalizacji i responsywności strony przy pomocy technologii TailwindCSS wykazało, że framework sprzyja optymalizacji strony internetowej poprzez tworzenie lżejszych i bardziej responsywnych stron, które szybciej się ładują i lepiej działają na urządzeniach o róznych rozdzielczościach, co jest korzystne dla użytkowników i programistów tworzących oprogramowanie.

5. Bezpieczeństwo i SEO

Bezpieczeństwo stron internetowych interpretować na wiele sposobów, zaczynając od kodu źródłowego, kończąc na interakcji z użytkownikiem. TailwindCSS to narzędzie służące do stylizacji aplikacji, przez co nie wpływa bezpośrednio na aspekty bezpieczeństwa takie jak autentykacja czy zarządzanie sesją, lecz przez usuwanie nadmiarowego kodu oraz używanie modułów, przyczynia się do zniwelowania miejsc w których może być przeprowadzony atak. Taki zabieg działa prewencyjnie dla przeprowadzenia ataków typu Cross-Site Scripting(XSS). Dodatkowo poprzez używanie przygotowanych wcześniej klas pomocniczych zamiast zagnieżdżonego kodu CSS ułatwia audyt bezpieczeństwa.

Optymalizacja dla wyszukiwarek internetowych (SEO) to proces w którym pod uwagę trzeba wziąć czynniki takie jak treści znajdujące się na stronie jak i szybkość z jaką strona się ładuję. Technologia TailwindCSS może mieć pozytywny pływ na wyszukiwarki internetowe, dzięki swojemu podejściu do wytwarzania semantycznego i czystego kodu HTML. Klasy pomocnicze pomagają programistą uniknąć powtarzania selektorów, dzięki czemu struktura plików jest dużo bardziej przejrzysta. Dodatkowo narzędzie TailwindCSS wspomaga tworzenie responsywnych stron internetowych, które są lepiej pozycjonowane w silniku wyszukiwarek internetowych, ponieważ są przyjazne dla smartfonów.

Dyrektywy, które można użyć do bezpieczeństwa jak i SEO, to „@layer” i „@apply”. Pierwsza z nich służy do organizacji klas pomocniczych w warstwy przez co łatwo utrzymać kod. Grupowanie w taki sposób klas, powoduję również szybsze ładowanie stron, co pozytywnie wpływa na ranking SEO. Natomiast wspomniana we wcześniejszych rozdziałach dyrektywa „@apply” służy do zwiększenia czytelności kodu i reużywalności komponentów, dzięki czemu zwiększana jest efektywność pracy programistów.

6. Wady TailwindCSS

Pomimo wielu zalet technologia TailwindCSS ma też swoje wady. Rewolucjonizująca technologia używania modułów w znacznikach HTML, może być również jej największym obciążeniem. Moduły odpowiadają za pojedyncze zmiany na stronie, tak więc bez używania klas pomocniczych w kodzie robi się straszny nieład i ciężko w nim coś znaleźć. By uniknąć nieczytelnego kodu framework wymusza na deweloperze pisania kodu w architekturze komponentowej, co w pewien sposób limituje swobodę tego narzędzia.

Tailwind nie jest również najprostszym frameworkiem służącym do tworzenia interfejsów graficznych dla początkujących. Technologia wymaga poznania wielu klas, co może być

przytłaczające dla nowych użytkowników. W przeciwieństwie do innych frameworków CSS w TailwindCSS nie używa się gotowych komponentów, co również może odrzucać programistów dopiero wybierających swoje środowisko.

Podsumowując, największymi wadami frameworka TailwindCSS są jego zalety. Technologia ta jest stworzona w taki sposób, by jak najbardziej przyśpieszyć tworzenie i ładowanie dedykowanych stron internetowych, bez bazowania na gotowych komponentach, lecz poprzez tworzenie własnych spersonalizowanych stylów.

7. Porównanie frameworków CSS

Porównanie TailwindCSS z innymi frameworkami pokazuję jak ważne jest podejście „Utility-first" stosowane przez Tailwind. Narzędzia, które zostaną porównane z technologią opisywaną w artykule to Bootstrap i Foundation, które są najbardziej rozpoznawalnymi narzędziami kaskadowych arkuszów stylów.

TailwindCSS stawia na skalowalność projektów i ich modularność, co wywodzi się z używania przygotowanych przez producenta klas. Jeden z największych konkurentów opisywanego w artykule frameworka to Bootstrap, który w przeciwieństwie do Tailwinda opiera się na gotowych komponentach i nadpisywaniu gotowych stylów. Gotowe komponenty Bootstrap wymagają dokładnego zrozumienia ich i zmiany istniejących już stylów, przez co biblioteka jest mniej efektywna i wymaga poświęcenia czasu, by opanować tworzenie stron przy jego pomocy. Przez takie podejście technologia dużo traci na personalizacji, ponieważ w TailwindCSS wszystkie style tworzone są przez dewelopera, który bazuję na zaprojektowanym wyglądzie strony przez projektanta i może dokładnie odwzorować zarys, bez czasochłonnego zagłębiania się w dokumentacje.

Filozofia „Utility-first" jest kluczowa dla prędkości rozwoju strony, przez możliwość wprowadzania szybkich zmian i tworzenia prototypów, dzięki predefiniowanym klasom. Pozwala to na zwiększenie swobody tworzenia projektu i wprowadzanie na bieżąco poprawek do designu. W frameworku Foundation style są dużo bardziej złożone, przez co tworzenie stron jest dużo bardziej czasochłonne.

Opisany w rozdziale 4 PurgeCSS służący do usuwania nieużywanych stylów, optymalizacji wydajności i zmniejszenia rozmiaru plików CSS wspierany przez TailwindCSS, niestety nie jest obsługiwany przez Bootrstrap. W przypadku używania kilku gotowych komponentów z biblioteki, powoduję załadowanie całego zestawu stylów, co znacząco wpływa na prędkość ładowania strony.

Tailwind pokazuje również swoją wyższość w tworzeniu responsywnych stron, gdzie do stworzenia stylizacji na mniejsze urządzenia potrzebne jest dużo mniej wysiłku dzięki stosowaniu stylów dla szerokości ekranu. W innych frameworkach tworzenie responsywności bazowane jest na tak zwanych siatkach dzielące ekran na kolumny.

8. Podsumowanie

TailwindCSS jest niesamowitym narzędziem stworzonym, by uprościć tworzenie kaskadowych arkuszów stylów. Jest to rewolucjonizujące narzędzie, które podeszło do tematu w inny sposób, niż swoi konkurenci. Wyróżnia się filozofią „Utility-First", która bazuję na używaniu zdefiniowanych przez autorów klas, które wpisywane są w atrybutach komponentów HTML. Takie podejście odróżnia to narzędzie od innych frameworków, ponieważ większość z nich bazuje na używaniu gotowych komponentów i nadpisywaniu ich stylów. Jak można się domyślić nadpisywanie użytego już kodu zmniejsza efektywność i optymalizacje strony, z czym Tailwind świetnie sobie radzi dzięki technologii PurgeCSS, która usuwa ze stylów nieużywany kod. Modularna charakterystyka frameworka nie przeszkadza w dokładnej stylizacji strony, dzięki możliwości ustawienia własnych wartości dla gotowych już klas. Jednocześnie Tailwind nie wymusza używania przygotowanych klas, lecz zachęca do tworzenia własnych reużywalnych klas dzięki dyrektywie „@apply", ay jeszcze bardziej przyśpieszyć i ujednolicić proces tworzenia aplikacji webowej.

Wielkim plusem tej technologii jest koncepcja z jaką podejmuje obsługiwanie responsywności na stronach. Inne frameworki do tworzenia interfejsu obsługiwane są poprzez dzielenie strony na kolumny i używanie modułów takich jak flexbox lub grid. Tailwind jest prekursorem ze swoim nowatorskim podejściem, w którym używa prefiksów połączonych z modułami, dzięki którym style z modułów używane są tylko wtedy kiedy szerokość urządzenia jest większa od tego podanego w prefiksie.

Największymi wadami frameworka TailwindCSS są jego zalety, ponieważ jest on stworzony w taki sposób, by ułatwić tworzenie spersonalizowanego interfejsu. Znaczniki HTML, w których używane są klasy stylizujące, mogą być bardzo nieczytelne, gdy nakładane jest wiele stylów, przez co programiści są niejako zobligowani do tworzenia reużywalnych komponentów i bazowaniu na nich swojego kodu. Technologia nie posiada również gotowych komponentów, co powoduję, że próg wejścia jest trochę większy niż w innych frameworkach, a początkujący deweloperzy dopiero wybierający swoje środowisko programistyczne mogą kierować się w strone innych narzędzi takich jak Bootstrap.

Literatura

Bhat K, *Ultimate TailwindCSS Handbook, Build sleek and modern websites with immersive UIs using TailwindCSS, 2023*

Źródła internetowe

https://tailwindcss.com/docs/utility-first (dostęp: 10.06.2024).

https://getbootstrap.com/docs/5.3/getting-started/introduction/ (dostęp: 12.06.2024).

https://get.foundation/sites/docs/index.html (dostęp: 13.06.2024).

Maciej Karczmarz, Jakub Jucha, Hubert Kraus, Adam Krawczyk, Sebastian Cwynar
SKNI „Kod”

dr inż. Bartosz TRYBUS
Opiekun Koła Naukowego

Techniki renderowania i optymalizacji. Ray tracing oraz technologie skalowania rozdzielczości

Celem niniejszego artykułu jest ogólne omówienie renderowania oraz techniki śledzenia promieni (ray tracing) w kontekście ich zastosowań w grafice komputerowej.

Renderowanie to proces generowania obrazów na podstawie modeli 3D, który odgrywa kluczową rolę w grafice komputerowej. Technika śledzenia promieni, szczególnie wspomagana metodami Monte Carlo, pozwala na tworzenie niezwykle realistycznych wizualizacji, modelując zjawiska optyczne takie jak odbicia, załamania światła i cienie.

Jednakże, nowoczesne algorytmy śledzenia promieni, wykazują dużą skuteczność w generowaniu realistycznych obrazów, choć kosztem wyższych wymagań obliczeniowych. Metody Monte Carlo, mimo swojej efektywności, wiążą się z problemem szumu w generowanych obrazach, który może być redukowany za pomocą technik denoising’u i optymalizacji próbkowania.

Słowa kluczowe: renderowanie, algorytm, Monte Carlo, ray tracing, model 3D, denoising.

1. Wprowadzenie

Grafika komputerowa jest dziedziną nauki i technologii, która dynamicznie rozwija się od kilku dekad, znacząco wpływając na różnorodne aspekty naszego życia, od rozrywki i edukacji, po projektowanie i badania naukowe. Jednym z kluczowych elementów grafiki komputerowej jest renderowanie, czyli proces generowania obrazów na podstawie modeli 3D. W tym kontekście technika śledzenia promieni (ray tracing) odgrywa szczególną rolę, oferując możliwości tworzenia niezwykle realistycznych wizualizacji.

Śledzenie promieni, wprowadzone po raz pierwszy w latach 60. XX wieku, stało się fundamentem wielu zaawansowanych technik renderowania. Algorytm ten polega na symulacji ścieżki, jaką pokonują promienie światła od źródła światła, przez obiekty w scenie, aż do obserwatora. Dzięki temu możliwe jest precyzyjne odwzorowanie zjawisk takich jak odbicia, załamania światła, cienie oraz efekty globalnego oświetlenia.

Metody Monte Carlo, szeroko stosowane w śledzeniu promieni, umożliwiają losowe próbkowanie promieni, co pozwala na bardziej realistyczne modelowanie złożonych interakcji świetlnych. Algorytmy takie jak path tracing czy bidirectional path tracing są w stanie generować obrazy o wysokiej jakości, choć kosztem znacznych zasobów obliczeniowych.

Zastosowania techniki śledzenia promieni oraz metod Monte Carlo są szerokie i różnorodne. W przemyśle filmowym i animacyjnym pozwalają na tworzenie spektakularnych efektów wizualnych i realistycznych środowisk wirtualnych. W grach komputerowych, mimo większych wymagań sprzętowych, techniki te są coraz częściej wykorzystywane dzięki postępom w technologii, umożliwiając generowanie wysokiej jakości grafiki w czasie rzeczywistym. W architekturze i projektowaniu wnętrz technika ta umożliwia tworzenie precyzyjnych wizualizacji, które wiernie oddają oświetlenie i materiały planowanych przestrzeni.

W niniejszym artykule dokonam analizy algorytmów renderowania, ze szczególnym uwzględnieniem techniki śledzenia promieni oraz metod Monte Carlo. Celem jest zrozumienie, w jaki sposób te zaawansowane techniki mogą być wykorzystane do tworzenia realistycznych obrazów oraz jakie wyzwania wiążą się z ich implementacją.

2. Definicja ray tracingu oraz jego historia

Ray tracing, jest techniką renderowania przy użyciu metod śledzenia promieni obrazów, która polega na symulacji ścieżki, jaką promienie światła pokonują w przestrzeni, aby dotrzeć do oka obserwatora. Technika ta jest znana z generowania bardzo realistycznych obrazów, ponieważ dokładnie odwzorowuje sposób, w jaki światło oddziałuje z obiektami w scenie.

Początki sięgają lat 60. XX wieku, kiedy to Arthur Appel w 1968 roku zaproponował metodę śledzenia promieni w celu generowania cieni w obrazach komputerowych. Jego praca była jednym z pierwszych kroków w kierunku realistycznego renderowania. W latach 70. technika ta była dalej rozwijana przez takich badaczy jak Turner Whitted, który w 1980 roku wprowadził koncepcję śledzenia promieni w kontekście globalnego oświetlenia. Whitted rozszerzył metodę o odbicia i załamania światła, co pozwoliło na tworzenie bardziej realistycznych obrazów. W latach 80. i 90. ray tracing zyskał na popularności dzięki rosnącej mocy obliczeniowej komputerów. W tym okresie technika ta była intensywnie badana i rozwijana, co doprowadziło do powstania wielu algorytmów optymalizacyjnych, które umożliwiły szybsze i bardziej efektywne renderowanie. Współcześnie ray tracing jest szeroko stosowany w przemyśle filmowym, gier komputerowych oraz w wizualizacjach architektonicznych.

Rodzaje ray tracingu:

Whitted-Style Ray Tracing:

Klasyczny ray tracing, wprowadzony przez Turnera Whitteda w 1980 roku, symuluje podstawowe efekty świetlne takie jak odbicia, załamania i cienie. Algorytm śledzi

promienie światła od kamery do sceny, a następnie generuje nowe promienie w punktach interakcji. Używany głównie do generowania obrazów o wysokim stopniu realizmu w zastosowaniach, gdzie precyzyjne odwzorowanie odbić i załamań jest kluczowe.

Path Tracing:

To metoda Monte Carlo, która śledzi promienie od kamery i losowo próbuje wiele ścieżek światła, aby dokładnie symulować efekty globalnego oświetlenia. Promienie są śledzone do momentu, aż dotrą do źródła światła lub zostaną pochłonięte. Idealny do generowania fotorealistycznych obrazów, szczególnie w filmach i animacjach, gdzie realizm oświetlenia jest kluczowy.

Bidirectional Path Tracing:

Rozszerzenie path tracing'u, które śledzi promienie zarówno od kamery, jak i od źródeł światła. Łączy ścieżki promieni, aby poprawić efektywność i dokładność symulacji światła, szczególnie w scenach z trudnym oświetleniem. Używany w renderingu scen o złożonym oświetleniu, takich jak wnętrza z ograniczonym dostępem światła.

Metropolis Light Transport (MLT):

Technika oparta na metodzie Metropolis-Hastings, która generuje ścieżki światła bardziej efektywnie poprzez próbkowanie ścieżek o wysokiej ważności. MLT zwiększa konwergencję (zbieżność) i redukuje szum w porównaniu do tradycyjnego path tracing'u. Efektywny w scenach z trudnym do symulowania oświetleniem, takich jak scenerie z dużą ilością refleksów lub małych źródeł światła.

Photon Mapping:

Dwufazowa technika, która najpierw emituje fotony ze źródeł światła i śledzi ich interakcje z powierzchniami (faza mapowania fotonów), a następnie wykorzystuje zebrane informacje do dokładnego obliczania oświetlenia (faza renderingu). Szczególnie skuteczny w symulacji efektów globalnego oświetlenia, *caustics* (obwiednia wiązki promieni świetlnych) i innych złożonych efektów świetlnych.

Real-Time Ray Tracing:

Technika ray tracingu zoptymalizowana do działania w czasie rzeczywistym, możliwa dzięki nowoczesnym kartom graficznym (np. NVIDIA RTX) wyposażonym w dedykowane rdzenie RT (Ray Tracing). Wykorzystuje zaawansowane algorytmy optymalizacji i przyspieszenia sprzętowego. Stosowany głównie w grach komputerowych i aplikacjach VR/AR, gdzie wymagana jest wysoka jakość grafiki w czasie rzeczywistym.

3. Śledzenie promieni od strony technicznej

1. **Kamera** w procesie śledzenia promieni odgrywa ważną rolę, ponieważ to od niej rozpoczyna się cały proces tworzenia sceny. Kamera jest umieszczona w określonym punkcie w przestrzeni 3D, a jej kierunek patrzenia oraz inne parametry, takie jak pole widzenia, są zdefiniowane przez użytkownika lub plik sceny. Dla każdego piksela na ekranie, z pozycji kamery generowany jest promień, który przechodzi przez dany piksel i jest przedłużany w głąb sceny. Promienie te są śledzone w celu określenia, jakie obiekty znajdują się na ich drodze. Gdy promień napotka obiekt, algorytm oblicza punkt przecięcia oraz właściwości powierzchni w tym punkcie, takie jak kolor, normalna powierzchni, oraz właściwości materiału (np. przezroczystość, odbicie). Na podstawie punktu przecięcia i właściwości powierzchni, algorytm oblicza, jak światło oddziałuje z tym punktem. Może to obejmować bezpośrednie oświetlenie od źródeł światła, odbicia od innych powierzchni oraz załamania światła. Jeśli powierzchnia jest odbijająca lub przezroczysta, generowane są dodatkowe promienie (odbijające się lub załamujące), które są śledzone w celu dalszego obliczenia interakcji światła. Proces ten może być powtarzany wielokrotnie, aż do osiągnięcia określonego limitu odbić lub gdy promień nie napotka więcej obiektów. Po obliczeniu wszystkich interakcji promieni z obiektami, końcowy kolor i jasność każdego piksela są ustalane na podstawie zebranych informacji. W ten sposób powstaje realistyczny obraz sceny.
2. Gdy **promień światła** pada na powierzchnię, może być odbity, załamany lub pochłonięty. W ray tracingu śledzone są promienie pierwotne, które wychodzą z kamery, oraz promienie wtórne, które powstają w wyniku interakcji promieni pierwotnych z powierzchniami. Padanie światła jest modelowane przy użyciu różnych praw fizyki, takich jak prawo odbicia i prawo załamania. Zetknięcie promienia z obiektem daje nam punkt do zacienienia i pewne informacje o lokalnej geometrii w tym punkcie. Ostatecznym celem jest znalezienie ilości światła wychodzącego z tego punktu w kierunku kamery. Aby to zrobić, musimy wiedzieć, ile światła dociera do tego punktu.

Dotyczy to zarówno geometrycznego, jak i radiometrycznego rozkładu światła w scenie. W przypadku bardzo prostych źródeł światła (np. świateł punktowych) geometryczny rozkład oświetlenia polega po prostu na znajomości położenia świateł. Jednakże światła punktowe nie istnieją w prawdziwym świecie, dlatego oświetlenie fizyczne często opiera się na źródłach światła obszarowego. Oznacza to, że źródło światła jest skojarzone z obiektem geometrycznym, który emituje światło z jego powierzchni.

Przykładowe wzory matematyczne stosowane w procesie śledzenia promieni:

- Załamanie (refrakcja)

Załamanie światła jest opisane przez prawo Snelliusa (prawo załamania), które określa zmianę kierunku promienia świetlnego przy przejściu między różnymi ośrodkami o różnych współczynnikach załamania. Wzór na kierunek załamania jest następujący:

$$\frac{sin\theta_i}{sin\theta_t} = \frac{n_t}{n_i}$$

gdzie:

- θi - to kąt padania,
- θt - to kąt załamania,
- ni - to współczynnik załamania ośrodka, z którego światło wychodzi,
- nt - to współczynnik załamania ośrodka, do którego światło wchodzi.

- Równanie Renderowania (Rendering Equation):

Równanie renderowania (ang. Rendering Equation) to fundamentalne równanie w grafice komputerowej, które opisuje sposób, w jaki światło odbija się od powierzchni i dociera do obserwatora. Zostało ono wprowadzone przez Jamesa Kajiya w 1986 roku i stanowi matematyczną podstawę dla wielu technik renderowania, w tym ray tracing'u. Równanie renderowania wyraża ilość światła wychodzącego z punktu na powierzchni w określonym kierunku jako sumę światła emitowanego przez powierzchnię oraz światła odbitego od innych źródeł. Równanie ma następującą postać:

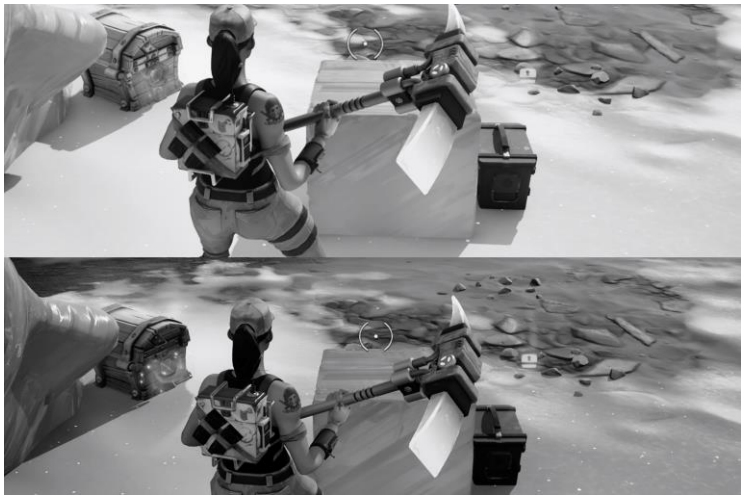
$$L_o(p,\, \omega_o) = L_e(p,\, \omega_o) + \int_\Omega f(p,\, \omega_0,\, \omega_i)\; L_i(p,\, \omega_i)\; |\cos\theta_i|\; d\omega_i$$

gdzie:

- $L_o(p, \omega_o)$ - Radiancja (emisja) wychodząca: światło wychodzące z punktu p w kierunku ωo,
- $L_e(p, \omega_o)$ - Radiancja emitowana: światło emitowane przez powierzchnię w punkcie x w kierunku ωo,
- $\int_{S2}$ - Całkowanie po półsferze: uwzględnia wszystkie kierunki padania światła ωi,
- $f(p, \omega_0, \omega_i)$ - Funkcja BRDF: opisuje, jak światło jest odbijane przez powierzchnię w punkcie $\mathcal{x}$ w kierunku ωo,
- $L_i(p, \omega_i)$ - Radiancja padająca: światło docierające do punktu x z kierunku ωi,
- $cos\theta_i$ - Cosinus kąta: iloczyn skalarny wektora normalnego do powierzchni i kierunku padającego światła θ_i, ważny dla efektywności oświetlenia.

4. Ray tracing w praktyce

W trakcie testowania technologii ray tracing'u, chciałbym przedstawić jej działanie w grze Fortnite.



Rysunek 1. Różnica przy włączonym i wyłączonym ray tracing'u w grze Fortnite.
Źródło: opracowanie własne.

Zdjęcie zostało podzielone na dwie części: górna z włączonym ray tracing'iem oraz dolna bez. Co można zauważyć, funkcja śledzenia cieni spowodowała, że obraz staje się bardziej realistyczny. Znacząco poprawiła się jakość wizualna w tej grze, dodając realistyczne efekty świetlne i cieniowanie. Dzięki temu gra staje się bardziej immersyjna. Technologia ta jest szczególnie przydatna w trybach kreatywnych, podczas robienia zrzutów ekranu oraz jest generalnie przyjemna dla oka, gdzie maksymalna jakość obrazu jest kluczowa.

5. Metody Monte Carlo

Techniki Monte Carl to metody numeryczne używane do rozwiązywania problemów matematycznych za pomocą losowego próbkowania. W kontekście ray tracing'u i path

tracing'u, metody Monte Carlo są wykorzystywane do symulacji rozpraszania światła w scenie, co pozwala na uzyskanie realistycznych efektów oświetleniowych.

Polegają na użyciu losowych próbek do oszacowania wartości funkcji lub rozwiązania problemu. W ray tracing'u, te metody są używane do symulacji różnych efektów świetlnych, takich jak odbicia, załamania, cienie i globalne oświetlenie. Kluczowym elementem tych metod jest generowanie dużej liczby losowych próbek i użycie ich do oszacowania średniej wartości, co pozwala na uzyskanie dokładnych wyników.

6. Techniki denosing'u

Obrazy wygenerowane przy użyciu ray tracing'u często zawierają szum, który jest efektem ubocznym metod Monte Carlo wykorzystywanych do próbkowania ścieżek światła. Aby zredukować ten szum i uzyskać gładsze, bardziej realistyczne obrazy, stosuje się różne techniki denoising'u:

1. **Filtry bilateralne**: To popularna technika filtracji obrazu, która redukuje szum, jednocześnie zachowując ostre krawędzie i detale. Działa ona poprzez zastępowanie wartości piksela średnią ważoną z sąsiednich pikseli, gdzie wagi są oparte na podobieństwie kolorów i odległości przestrzennej.
2. **Techniki oparte na uczeniu maszynowym**: W ostatnich latach pojawiły się zaawansowane techniki denoising'u oparte na głębokim uczeniu maszynowym. Wykorzystują one sieci neuronowe, takie jak autoencodery lub sieci resztkowe, do przewidywania i usuwania szumu z obrazów wygenerowanych ray tracing'iem. Sieci te są trenowane na dużych zbiorach danych zawierających pary obrazów z szumem i bez szumu, co pozwala im uczyć się efektywnego usuwania szumu, zachowując jednocześnie istotne detale.
3. **Techniki adaptacyjne**: Adaptacyjne techniki denoising'u dostosowują siłę filtracji w zależności od lokalnych właściwości obrazu, takich jak kontrast lub obecność krawędzi. Pozwala to na skuteczniejsze usuwanie szumu w gładkich obszarach, jednocześnie zachowując ostre krawędzie i detale.
4. **Techniki oparte na rekonstrukcji**: Te techniki wykorzystują informacje o ścieżkach promieni i właściwościach materiałów w scenie do rekonstrukcji obrazu bez szumu. Przykładem jest technika rekonstrukcji filtrowanej, która wykorzystuje filtry rekonstrukcyjne do odtworzenia gładkich powierzchni na podstawie próbek ścieżek promieni.

Temat technik opartych na rekonstrukcji jest całkiem świeżym tematem, jeżeli chodzi o rynek gier. Techniki oparte na rekonstrukcji, takie jak:

- NVIDIA DLSS,
- AMD FSR,
- Intel XeSS.

są stosunkowo nowymi rozwiązaniami, które zrewolucjonizowały sposób tworzenia gier i aplikacji w wysokiej rozdzielczości. Czym każda z nich się wyróżnia?

1. NVIDIA DLSS (Deep Learning Super Sampling):

DLSS wykorzystuje uczenie maszynowe do renderowania gier w niższej rozdzielczości, a następnie rekonstrukcji obrazu do wyższej rozdzielczości przy użyciu specjalnej sieci neuronowej. Rezultatem jest obraz o jakości zbliżonej do natywnej wysokiej rozdzielczości, ale z mniejszym obciążeniem obliczeniowym. DLSS jest ekskluzywną technologią NVIDII dostępną tylko na kartach graficznych RTX serii 20 i nowszych takich jak seria 30 i 40.

2. AMD FSR (FidelityFX Super Resolution):

FSR to odpowiednik DLSS od AMD, który również wykorzystuje techniki rekonstrukcji obrazu do renderowania gier w niższej rozdzielczości, a następnie skalowania do wyższej rozdzielczości. W przeciwieństwie do DLSS, FSR nie wykorzystuje uczenia maszynowego, ale opiera się na bardziej tradycyjnych algorytmach skalowania i rekonstrukcji obrazu. FSR jest kompatybilne z szeroką gamą kart graficznych AMD i NVIDIA, gdzie NVIDIA kompatybilności z kartami AMD.

3. Intel XeSS (Xe Super Sampling):

XeSS to technologia rekonstrukcji obrazu opracowana przez Intela, która podobnie jak DLSS, wykorzystuje uczenie maszynowe do renderowania gier w niższej rozdzielczości, a następnie rekonstrukcji do wyższej rozdzielczości. XeSS będzie dostępne na procesorach Intel ze zintegrowaną grafiką Xe oraz na dedykowanych kartach graficznych Intel Arc.

7. Dążenie do realizmu w grach? - Ray reconstruction

Ray Reconstruction to technologia wprowadzona w ramach NVIDIA DLSS 3.5, która wykorzystuje sztuczną inteligencję do poprawy jakości obrazu w efektach ray tracingu. Działa na wszystkich kartach graficznych GeForce RTX i ma na celu zastąpienie tradycyjnych denoiserów, które wymagają ręcznego dostrajania dla każdej sceny, jednym modelem AI.

Dzięki temu możliwe jest uzyskanie lepszej jakości obrazu, szczególnie w dynamicznych scenach, gdzie tradycyjne denoisery mogą powodować artefakty takie jak ghosting.

Ray Reconstruction działa poprzez integrację dodatkowych danych z silnika gry oraz nowego modelu AI, który jednocześnie wykonuje super rozdzielczość i rekonstrukcję promieni. Technologia ta poprawia ostrość odbić i dynamiczne oświetlenie, co jest szczególnie widoczne w ruchu. Przykłady zastosowania Ray Reconstruction można zobaczyć w grach takich jak Cyberpunk 2077, gdzie technologia ta znacząco poprawia jakość wizualną scen.

8. Podsumowanie

W artykule omówiono ray tracing jako zaawansowaną technikę renderowania, która symuluje fizyczne interakcje światła z obiektami, umożliwiając generowanie fotorealistycznych obrazów. Przedstawiono różne rodzaje ray tracingu, takie jak Whitted-Style Ray Tracing, Path Tracing, Bidirectional Path Tracing, Metropolis Light Transport, Photon Mapping oraz Real-Time Ray Tracing. Opisano podstawowe zasady działania, w tym modelowanie odbić i załamań światła, oraz kluczowe wzory matematyczne. Zastosowania ray tracingu obejmują przemysł filmowy, gry komputerowe, architekturę, symulacje naukowe i reklamę, co podkreśla jego wszechstronność i rosnące znaczenie w różnych branżach.

Jednakże, jedynie poruszono zamysł i idee ray tracing'u, nie wyczerpując tematu, ponieważ jest on bardzo obszerny i obejmuje wiele zaawansowanych zagadnień technicznych oraz praktycznych zastosowań.

Literatura

Matthew Pharr, Wenzel Jakob, *Physically Based Rendering, fourth edition*, Wydawnictwo MIT Press Ltd, marzec 2023

Peter Shirley, Trevor David Black, Steve Hollasch, *Ray Tracing in One Weekend Series*

Kevin Suffern, *Ray Tracing from the Ground Up*, Wydawnictwo A K Peters/CRC Press, Wrzesień 2007

Źródła internetowe

https://www.pbr-book.org/ (dostęp 27.05.2024)

https://arxiv.org/pdf/1504.01896 (dostęp 29.05.2024)

Hubert Kraus, Adam Krawczyk, Jakub Jucha, Sebastian Cwynar, Maciej Karczmarz
SKNI „KOD”

dr inż. Bartosz Trybus
Opiekun Koła Naukowego

Zastosowanie ORM w tworzeniu aplikacji webowych w języku JavaScript na przykładzie Node.js

Streszczenie

Artykuł przedstawia implementację ORM podczas tworzenia aplikacji webowych. Autorzy opisują czym jest mapowanie obiektowo-relacyjne i jakie ma cechy. Wyjaśniają jakie są jego najważniejsze funkcje oraz gdzie znajduje zastosowanie. Porównują również oba podejścia do pracy z relacyjną bazą danych. Zaprezentowane są zarówno wady i zalety korzystania z ORM jak i pisania surowych zapytań w języku SQL. Artykuł przedstawia także dostępne biblioteki ORM dla języka JavaScript, prezentuje ich dostępność oraz popularność, porównuje różnice jak i podobieństwa pomiędzy nimi. Każda z bibliotek ma swój opis oraz analizę zalet i wad. Autorzy zawarli także porównanie kodu i implementacji połączenia z bazą danych podczas tworzenia projektu. Artykuł obejmuje także stworzenie przykładowej aplikacji wykorzystując Node.js realizującą implementację ORM używając do tego biblioteki Sequelize oraz PostgreSQL jako bazy danych. Artykuł pokazuje proces oraz wymagania stworzenia projektu, charakteryzacje dokładnie modele w ORM i opisuje ich zastosowanie oraz prezentuje możliwości jakie ma ich tworzenie. Omówiono a także zaprezentowano operację na danych zawartych w bazie danych takie jak tworzenie i wyszukiwanie obiektu.

Słowa kluczowe: backend, postgres, sequelize, model, sql, baza danych.

1. Wprowadzenie

Współczesne aplikacje webowe wymagają operowania na zewnętrznym źródle danym. W większości przypadków dane przechowywane są w relacyjnych bazach danych które zapewniają ich trwałość. Zadaniem programisty jest zintegrowanie razem dwóch różnych modeli operowania na danych – obiektowego i relacyjnego. Jest to przyczyną wyzywania jakim jest złożoność dostosowania kodu programowania do struktur bazy danych. Mapowanie obiektowo-relacyjne (ORM) to technika, która tworzy warstwę pomiędzy językiem a bazą danych, pomagając tym samym programistom pracować na danych w niej zawartych.

Problemem dla programistów jest konieczność dobrego zrozumienia i programowania w strukturalnym języku zapytań (SQL), który pozwala na połączenie aplikacji webowej z bazą danych. Pisanie surowych zapytań SQL może być bardzo czasochłonne i nieefektywne z powodu potrzeby wyodrębnienia każdego elementu z kodu.

ORM można postrzegać jako warstwę łączącą programowanie obiektowe z relacyjnymi bazami danych. Umożliwia pracę z danymi zawartymi w bazie w sposób, który bardziej zgadza się z obiektowym podejściem do programowania. Dzięki ORM operacje takie jak dodawanie nowych danych czy aktualizacja już tych istniejących, jest realizowane przez operacje na obiektach w kodzie, które następnie mapowane są na rekordy w bazie danych. ORM-y są jednocześnie wzbudzają popularność jak i kontrowersyjne. Argumentami za ORM jest niewątpliwie zwiększenie produktywności, poprawienia projektowania aplikacji oraz ponownego wykorzystania kodu, sprawiają również, że utrzymują aplikację aktualną. Negatywnym aspektem ORM jest przede wszystkim wydajność. Celem artykułu jest porównanie różnych ORM z wykorzystaniem Node.js, sposób ich implementacji wraz z przykładami zastosowań praktycznych jaki i również porównanie wad i zalet tych rozwiązań.

2. Czym jest ORM?

Mapowanie obiektowo-relacyjne, w skrócie (ORM) to nowoczesne podejście do zagadnienia współpracy z bazą danych. Jego cechą charaktery-styczną jest wykorzystanie filozofii programowania obiektowego[1]. ORM zapewnia warstwę obiektową między relacyjnymi bazami danych a obiektowym podejściem do programowania bez konieczności pisania surowych zapytań SQL. Standaryzuje on interfejsy, redukując ilość szablonów i przyspiesza czas programowania.

ORM charakteryzuje się kilkoma kluczowymi cechami. Pierwszą z nich jest abstrakcja, mapowanie obiektowo-relacyjne umożliwia programistom skupienie się na logice biznesowej, a nie na szczegółach implementacyjnych bazy danych. Operacje na bazie danych stają się niezależne od konkretnego silnika bazy danych, co otwiera drogę do większej elastyczności i łatwiejszej migracji pomiędzy różnymi systemami baz danych[2]. Kolejną cechą jest bezpieczeństwo, większość dostępnych narzędzi ORM zapewnia mechanizmy które zabezpieczają przed atakami, takimi jak wstrzykiwanie SQL. Jest to kluczowy aspekt bezpieczeństwa danych. Omawiane podejście umożliwia także wygodne zarządzanie danymi. Łatwe tworzenie i modyfikacja istniejących struktur baz danych a także zarządzanie danymi za pomocą prostych i przejrzystych zapytań obiektowych to realizacja tej cechy.

[1] *Co to jest ORM oraz czym się różni od SQL*, https://mindboxgroup.com/pl/co-to-jest-orm-oraz-czym-sie-rozni-od-sql-object-relational-mapping/, [dostęp: 18.05.2024].
[2] T. Kozon, *ORM: Co to jest i jak działa?*, https://boringowl.io/blog/orm-co-to-jest-i-jak-dziala, [dostęp: 18.05.2024].

Mapowanie obiektowo-relacyjne oferuje wiele funkcji, które ułatwiają interakcję pomiędzy programami napisanymi obiektowo a relacyjnymi bazami danych.

Do najważniejszych funkcji ORM można zaliczyć:

- tworzenie zapytań SQL,
- tworzenie modeli danych w kodzie, przedstawiających tabele z baz danych,
- obsługa relacji pomiędzy tabelami.

Mechanizm, który przekształca dane w ORM działa na podstawie mapowania obiektów stworzonych w aplikacji na odpowiadające tabele w relacyjnej bazie danych. Każda instancja obiektu reprezentuje rzędy, a atrybuty reprezentują kolumny w tabeli[3].

ORM znajduje zastosowanie w wielu dziedzinach programowania. Przykładem jest tworzenie zaawansowanych systemów zarządzania bazami danych, które pozwalają na intuicyjne i efektywne manipulowanie danymi z bazy, nie zależnie od użytego języka programowania. Użycie ORM jest także praktyczne, jeśli chcemy poprawić bezpieczeństwo naszych aplikacji, większość bibliotek ma wbudowane mechanizmy zapobiegające atakom typu wstrzykiwanie SQL. praktycznych jaki i również porównanie wad i zalet tych rozwiązań.

3. Różnice pomiędzy ORM a SQL

SQL to język zapytań, który jest bezpośrednio skoncentrowany na manipulacji danymi w relacyjnych bazach danych. Pozwala na precyzyjne definiowanie zapytań, co pozwala na osiągnięcie optymalizacji i wydajności. Praca bezpośrednio z SQL daje pełny nadzór nad tym, jak dane są manipulowane[4]. SQL to uniwersalny język zapytań, bardzo dobrze udokumentowany i obsługiwany przez wiele narzędzi oraz posiadający ogromną społeczność. Zapytania SQL mogą jednak stać się bardzo skomplikowane, zwiększa to ryzyko pojawienia się błędów i utrudnia utrzymanie kodu. Brak abstrakcji od szczegółów może wymagać od programistów dobrego zrozumienia technologii bazy danych. Stosowanie SQL stwarza także duża zależność od używanej bazy danych przez silnie powiązanie z konkretnymi systemami zarządzania bazami danych, może to ograniczyć elastyczność w zmianie technologii.

Mapowanie obiektowo-relacyjne (ORM) natomiast otwiera nowe możliwości na współpracę z bazami danych, pozwalając na korzystanie z różnych języków programowania bez konieczności bezpośredniej pracy z SQL. Dzięki ORM, dane są prezentowane w łatwiejszym dla człowieka formacie, co ułatwia zrozumienie struktury bazy danych.

---

[3] T. Kozon, *ORM: Co to jest i jak działa?*, https://boringowl.io/blog/orm-co-to-jest-i-jak-dziala, [dostęp: 18.05.2024].

[4] R. Awati, *object-relational mapping (ORM)*, https://www.theserverside.com/definition/object-relational-mapping-ORM, [dostęp: 19.05.2024].

Z czasem, rosnąca komplikacja operacji jest przyczyną dłuższych i bardziej skomplikowanych zapytań SQL, zwiększa to ryzyko popełnienia błędów. Wieloznaczność i różnorodność sposobów pisania zapytań dodatkowo komplikuje proces. ORM działa jako mediator, przekształca kod z jednego języka na drugi, tworząc abstrakcję nad bazą danych. Pozwala to na pracę z danymi poprzez operacje na obiektach, w ten sposób pozbywamy się bezpośredniej manipulacji kodem SQL. ORM może być jednak wolniejszy niż bezpośredni SQL ze względu na dodatkowe warstwy abstrakcji, które mogą wpływać na wydajność. Wadą rozwiązania jest także złożoność, niektóre ORM mogą być zbyt skomplikowane w działaniu lub składni albo mają ograniczone funkcje, co może powodować naukę w celu zrozumienia ich działania.

Każde rozwiązanie, zarówno ORM jak i SQL mają swoje unikalne zalety i wady, które powinny decydować o wyborze odpowiedniego podejścia w zależności od projektu. ORM oferuje abstrakcję, szybkość rozwoju i łatwość integracji, ale może kosztować wydajność działania aplikacji. SQL zapewnia pełną kontrolę i wydajność, ale może być skomplikowany i wymagać głębokiego zrozumienia technologii bazy danych oraz większej uwagi podczas zabezpieczania bazy. Wybór między nimi zależy od priorytetów projektu, wymagań dotyczących wydajności i preferencji zespołu programistów.

4. Porównanie dostępnych bibliotek

ORM możemy zaimplementować przy użyciu wielu bibliotek. Każda z nich różni się w sposobie działania, jakości i trudności w wdrożeniu. Każdy z wymienionych ORM ma swoje mocne i słabe strony, dlatego warto przeprowadzić analizę i testowanie, aby znaleźć najlepsze rozwiązanie dla tworzonego projektu.

Bezpośrednie porównanie dostępnych bibliotek:

Tabela 1. Porównanie bibliotek ORM dla języka Javascript

ORM	**Sequelize**	**Knex.js**	**Mongoose**	**TypeORM**	**Prisma**
Popularność	**Duża**	**Mała**	**Duża**	**Średnia**	**Średnia**
Dostępne bazy danych	Postgres MySQL MariaDB SQLite MSSQL	Postgres MySQL MariaDB SQLite3 Oracle MSSQL Amazon Redshift	MongoDB	Postgres MySQL MariaDB MongoDB SQLite MSSQL Oracle	Postgres MySQL SQLite

Źródło: opracowanie własne

Sequelize jest dojrzałą i popularną biblioteką z doskonałą dokumentacją i świetnie wytłumaczonym kodem. Posiada także własny konstruktor zapytań. Obsługuje wiele funkcji

warstwy danych, wspiera większość popularnych baz danych. Sequelize oferuje pełne API dla operacji na danych, obsługę transakcji, migracje.

Knex.js jest obecnie najbardziej dojrzałym kreatorem zapytań SQL w JavaScript, który może działać zarówno w Node.js, jak i w przeglądarce. Jest w stanie generować wysoce wydajne zapytania SQL, które są na równi z ręcznie napisanymi instrukcjami SQL. Należy pamiętać, że wiele bibliotek ORM korzysta z Knex.js. Należą do nich między innymi Bookshelf, Objection.js czy MikroORM.

Moongoose jest ORM idealnym, jeśli w projekcie używana będzie baza danych MongoDB. Jest to obecnie na chwilę obecną najpopularniejsza biblioteka ORM w świecie JavaScript. Posiada takie funkcje jak wbudowane rzutowanie typów, walidacja, budowanie zapytań oraz haki poprzez oprogramowanie pośredniczące.

TypeORM to rozwiązanie, które działa zarówno w Node.js, przeglądarce jak i platformach React Native, NativeScript. Może być implementowane zarówno z TypeScript jak i JavaScript co jest dla tej biblioteki ogromnym plusem. TypeORM skupia się na wspieraniu wszystkich aktualizacji oraz stara się pomagać w rozwoju nawet najmniejszych aplikacji.

Prisma to nowoczesne podejście i rozwiązanie ORM, oferuje ono zintegrowane narzędzia do generowania schematów bazy danych, migracji i obsługi relacji. Prisma jest znane z dobrej dokumentacji i łatwości implementacji w kodzie, powoduje to, że jest atrakcyjnym wyborem przy tworzeniu nowych aplikacji. Rozwiązanie to różni się od wszystkich innych omówionych ORM. Nie używa modeli obiektowych, ale pliku schematu, który służy do mapowania wszystkich tabel i kolumn. Plik jest wykorzystywany przez narzędzie migracji do generowania pliku migracji SQL.

5. Porównanie operacji na bazie danych dla każdej z bibliotek

Każda z bibliotek wymaga inicjalizacji i połączenia z bazą danych. Jest to realizowane w różny sposób. Poniżej znajdują się fragmenty kodu odpowiedzialne za ustawienie połączenia z bazą.

Dla dodatkowego porównania zostanie przeprowadzona operacja pobierania wartości z bazy. Dla każdej biblioteki zrealizowane zostanie zapytanie polegające na wyszukaniu wszystkich postów dla danego użytkownika na podstawie podanego adresu e-mail.

Przykład kodu dla biblioteki Knex.js:

```
const knex = require('knex')({
    client: 'mysql',
    connection: {
      host : '127.0.0.1',
      user : 'username',
      password : 'password',
      database : 'database_name'
    }
  });

  const posts = await knex('posts').join('users',{
      'users.id': 'posts.user_id',
      'users.email': 'przyklado-wy@knex.js',
      })
      .select('*')
```

Rysunek 1. Połączenie z bazą oraz zapytanie SELECT dla biblioteki Knex.js
Źródło: opracowanie własne.

Przykład kodu inicjalizacji dla biblioteki Sequelize:

```
const sequelize = new Sequelize('database', 'username', 'password', {
    host: 'localhost',
    dialect: 'postgres'
  });

  const posts = await User.findOne({
      where:{
      email: "przykladowy@sequelize.pl
      },
      include: Post,
  })
```

Rysunek 2. Połączenie z bazą oraz zapytanie SELECT dla biblioteki Sequelize
Źródło: opracowanie własne.

Przykład kodu inicjalizacji dla biblioteki Mongoose:

```
const mongoose = require('mongoose');
mongoose.connect('mongodb://localhost/test', {useNewUrlParser: true, useUnifiedTopology: true});

const posts = await User.findOne({
    email: 'przykladowy@mongoose.pl',
}).populate('posts')
```

Rysunek 3. Połączenie z bazą oraz zapytanie SELECT dla biblioteki Mongoose
Źródło: opracowanie własne.

Przykład kodu inicjalizacji dla biblioteki TypeORM:

```
const AppDataSource = new DataSource({
    type: "postgres",
    host: "localhost",
    port: 5432,
    username: "test",
    password: "test",
    database: "test",
    synchronize: true,
    logging: true,
    entities: [Post, Category],
    subscribers: [],
    migrations: [],
})

const posts = await userRepository.findOne(id, {
    relations: ['posts'],
})
```

Rysunek 4. Połączenie z bazą oraz zapytanie SELECT dla biblioteki TypeORM
Źródło: opracowanie własne.

Inicjalizacja połączenia w ORM Prisma wygląda inaczej niż w przypadku poprzednich bibliotek. Polega ona na zmianach w pliku schema.prisma ustawień dotyczących bazy danych.

```
datasource db {
    provider = "postgresql"
    url      = env("DATABASE_URL”)
}

const posts = await prisma.user.findOne({
    where:{
        {email: 'przykladowy@prisma.io'}
    }
}).posts()
```

Rysunek 5. Połączenie z bazą oraz zapytanie SELECT dla biblioteki Prisma
Źródło: opracowanie własne.

Schemat inicjalizacji połączenia różni się w mniejszym lub większym stopniu w zależności od użytej biblioteki. Zasada działania jest jednak taka sama, połączenie wymaga zdefiniowania adresu, użytkownika i jego hasła oraz nazwy używanej bazy danych.

Składnia zapytania realizującego wyszukiwanie postów użytkownika dla każdej z bibliotek jest inna, jednak nie różni się w znacznym stopniu co pozwala na szybką naukę innej biblioteki w razie potrzeby. Użyty kod z każdej z bibliotek jest elegancki i prosty do pisania jak i analizowania.

Będąc świadom różnic oraz podobieństw, deweloper aplikacji powinien wybrać ORM w oparciu o cechy i cele realizowanego projektu.

6. Stworzenie projektu

Do implementacji ORM użyta zostanie biblioteka Sequelize ze względu dobrej, szczegółowej dokumentacji oraz prostoty implementacji. Projekt będzie stworzony w środowisku uruchomieniowym Node.Js używającym dodatkowo Express.Js. Jako baza danych zastosowana będzie PostgreSQL.

Do zaczęcia projektu wymagana jest instalacja potrzebnych pakietów. Użyty zostanie do tego domyślny manager pakietów npm. Polecenia do realizacji projektu:

```
npm install pg
npm install pg-hstore
npm install sequelize
```

Rysunek 6. Polecenia npm do instalacji pakietów
Źródło: opracowanie własne.

Zastosowanie w projekcie mapowania obiektowo-relacyjnego należy roz-począć od połączenia z bazą danych. Atrybuty wymagane do połączenia zostały zdefiniowane w pliku db.config.js.

```
1  module.exports = {
2      HOST: "localhost",
3      USER: "postgres",
4      PASSWORD: "123",
5      DB: "rfid",
6      dialect: "postgres",
7    };
```

Rysunek 7. Plik db.config.js, atrybuty połączenia z bazą danych
Źródło: opracowanie własne.

Powyższe wartości są następnie wykorzystywane w tworzeniu połączenia. W ten sposób zwiększa się czytelność kodu i możliwość ponownego użycia.

W pliku index.js zdefiniowane jest połączenie z stworzoną bazą w PostgreSQL zgodnie z przykładem pokazanym w porównaniu bibliotek.

```
const config = require("../config/db.config.js");

const Sequelize = require("sequelize");
const sequelize = new Sequelize(
  config.DB,
  config.USER,
  config.PASSWORD,
  {
    host: config.HOST,
    dialect: config.dialect,
    pool: {
      max: config.pool.max,
      min: config.pool.min,
      acquire: config.pool.acquire,
      idle: config.pool.idle
    }
  }
);
```

Rysunek 8. Fragment pliku index.js z implementacją połączenia z bazą danych
Źródło: opracowanie własne.

W celu synchronizacji z bazą danych użyty zostanie model.sync biblioteki Sequelize. Automatycznie tworzy on tabelę w bazie danych, jeśli nie istnieje. Jeśli istnieje nie modyfikuje jej. Synchronizacja zdefiniowana została w pliku server.js.

```
const db = require("./app/models");
db.sequelize.sync().then(() => {
    console.log('Database sequelized');
});
```

Rysunek 9. Fragment pliku server.js synchronizujący modele z bazą danych
Źródło: opracowanie własne.

Powyższy fragment kodu jednocześnie synchronizuje wszystkie zdefiniowane w projekcie modele.

7. Modele

Model w mapowaniu obiektowo-relacyjnym jest abstrakcyjną reprezentacją tabeli w bazie danych. W bibliotece Sequelize jest to klasa, która dziedziczy po klasie Model. Każdy stworzony model odpowiada jednej tabeli w relacyjnej bazie danych natomiast każda instancja modelu odpowiada re-kordowi w danej tabeli.

Kluczowymi elementami modelu jest między innymi klasa modelu, jej nazwa zwykle odpowiada nazwie tabeli, lecz jest to w pełni konfigurowalne.

Kolejnym elementem są atrybuty, czyli pola klasy modelu odpowiadające kolumnom w tabeli. Typy tych atrybutów z klasy zostaną zmapowane na odpowiadające im typy danych w poszczególnej bazie danych. Przykładowo typ String zostanie zamieniony na VARCHAR a typ INTEGER na INT. Jest to kolejną zaletą wykorzystywania ORM, zdefiniowane modele będą automatycznie tworzone bez względu na to w jakim dialekcie SQL będą tworzone. Trzecim elementem modelu są metadane, zawierają one dodatkowe informację i konfigurację. Są to między innymi: typy pól, klucze główne i obce, ograniczenia takie jak UNIQUE czy NOT NULL oraz definiowanie relacji pomiędzy tabelami.

Przykładowy model jednej z tabel w projekcie:

```
module.exports = (sequelize, Sequelize) => {
    const inventory_item = sequelize.define("inventory_item", {
        ID: {
            type: Sequelize.INTEGER,
            primaryKey: true,
            autoIncrement: true,
            unique: true,
        },
        sap_item_id: {
            type: Sequelize.STRING,
        },
        vm_item_id: {
            type: Sequelize.STRING,
        },
        inventory_id: {
            type: Sequelize.INTEGER,
        },
    },
    {
    timestamps: false,
    freezeTableName: true
    });

    inventory_item.associate = function(models) {
        inventory_item.hasOne(models.inventory_item_outcome, {
            foreignKey: 'inventory_item_id',
            as: 'inventoryitemid'
        });
    };
    return inventory_item;
};
```

Rysunek 10. Plik inventory_item.model.js definiujący model
Źródło: opracowanie własne.

Jest to model, który implementuje tabele inventory_item służącą jako połączenie pomiędzy dwoma innymi tabelami.

Zawiera ona każdy kluczowy element wymieniony wcześniej. Pola zdefiniowane mają typy danych które w kodzie są reprezentowane przez bibliotekę Sequelize, zostaną one zamienione podczas synchronizacji na odpowiadające im typy w bazie danych. Dla pola ID zostały także

zdefiniowane meta-parametry ustalające klucz główny, auto inkrementację oraz unikalność pola.

W dolnej części modelu ustalona została relacja jeden do wielu z tabelą inventory_item_outcome. Tak stworzony moduł może być używany w wielu miejscach po odwołaniu się do niego. Model jest przejrzysty, możliwy do ponownego użycia oraz ma jasno określoną strukturę. Stworzone w ten sposób modele używając ORM pozwalają na efektywniejszą pracę dewelopera.

8. Operacje na danych z użyciem ORM

Zaimplementowane mapowanie obiektowo-relacyjne ułatwia również pracę na danych podczas tworzenia aplikacji. Pozwala na łatwe tworzenie nowego obiektu oraz automatyczne stworzenie go.

```
await Promise.all(body.map(async item => {
    const assetId = `${item.id}${item.rev}`;
    await SapItem.create({
        asset_no: item.id,
        inventory_id: latestInvent,
        description: item.description,
        capitalized_date: item.capDate,
        room: item.room,
        sub_no: item.rev,
        asset_id: assetId
    });
}));
```

Rysunek 11. Fragment kodu realizujący tworzenie nowego rekordu w tabeli sapitem
Źródło: opracowanie własne.

Funkcja biblioteki Sequelize automatycznie przeprowadza wszystkie operację dodania obiektu do bazy danych. Kod jest czytelny i nie wymaga dodatkowych definicji tak jak w przypadku używania czystego SQL.

Przykładowa realizacja zapytania SELECT do bazy danych:

```
const vmdata = await VmItem.findAll({
    where: {
        inventory_id: inventories.ID,
        location: req.body.location
    },
    attributes: ['asset_id'],
})
```

Rysunek 12. Fragment kodu realizujący zapytanie SELECT na tabeli VmItem
Źródło: opracowanie własne.

Powyższy kod w prosty sposób realizuje zapytanie SELECT i wyszukuje wszystkie id przedmioty w danej inwentaryzacji i lokalizacji. Kod jest wyraźny i pozwala na łatwą interpretację bez znajomości języka SQL.

Używanie ORM ma jednak swoje wady, w przypadku pisania bardziej złożonych zapytań do bazy kod staje się mniej czytelny i może powodować większe trudności w analizie niż zapytanie w języku SQL. Ma to również wpływ na efektywność, ORM poprzez swoją złożoność często generuje o wiele mniej wydajne zapytania od tych które są możliwe do osiągnięcia.

Nie wszystkie zapytania można obsłużyć ORM, w szczególności są to za-pytania, które zawierają w sobie podzapytania. Uzyskanie takiego efektu w ORM jest ciężkie, dla niektórych bibliotek nie możliwe. Systemy ORM w takim przypadku często oferują wprowadzanie czystego SQL jednak w takim przypadku oznacza to, że deweloper aplikacji musi stworzyć zapyta-nie pisząc je w osobnym języku co powoduje rozbieżności w kodzie.

9. Podsumowanie

Zastosowanie ORM w tworzeniu aplikacji webowych jest niewątpliwie rozwiązaniem, które wymaga indywidualnej analizy dotyczącej każdego projektu. Pozwala ono z pewnością na efektywniejszą pracę dewelopera poprzez możliwość tworzenia całych modeli reprezentujących tabele oraz wielokrotne wykorzystywanie jej. Taki tryb zwiększa także czytelność oraz spójność całego kodu. Może mieć jednak negatywny wpływ na efektywność całej aplikacji co w pewnych przypadkach odgrywa kluczową rolę. ORM pozwala przede wszystkim na osiągnięcie kodu, który jest ustandaryzowany, bezpieczny oraz łatwy w utrzymaniu. Pomaga w łatwiejszy sposób pracować z bazą danych i oferuje duże możliwości.

Literatura

1. Niegowski M., *Generyczne mapowanie obiektowo-relacyjne z wykorzystaniem dedykowanego oprogramowania*, Polsko-Japońska Wyższa Szkoła Technik Komputerowych, Warszawa 2009.

Źródła internetowe

1. https://enterthecode.pl/specjalizacja/orm-mapowanie-obiektowo-relacyjne-zasady-zalety/ (dostęp: 18.05.2024)
2. https://mindboxgroup.com/pl/co-to-jest-orm-oraz-czym-sie-rozni-od-sql-object-relational-mapping/ (dostęp: 19.05.2024)
3. https://boringowl.io/blog/orm-co-to-jest-i-jak-dziala (dostęp: 19.05.2024)
4. https://www.altexsoft.com/blog/object-relational-mapping/ (dostęp: 20.05.2024)
5. www.theserverside.com/definition/object-relational-mapping-ORM (dostęp: 20.05.2024)

Oskar Niedziałek, Krystian Kiełbasa, Hubert Futoma
Studenckie Koło Naukowe Informatyków „KOD”

Dr. inż. Bartosz Trybus
Opiekun Koła Naukowego

Jetpack Compose w Android: nowoczesne podejście do tworzenia interfejsów użytkownika

Streszczenie

Celem artykułu jest dostarczenie kompleksowego przeglądu Jetpack Compose, nowoczesnej biblioteki UI dla platformy Android opracowanej przez Google. Biblioteka ta wprowadza deklaratywne podejście do tworzenia interfejsów użytkownika, integrując się z językiem programowania Kotlin. Artykuł omawia teoretyczne podstawy Jetpack Compose, praktyczne aspekty jego użycia oraz zarządzanie stanem i animacjami w aplikacjach Android. Metody badawcze obejmują analizę literatury przedmiotu oraz praktyczne eksperymenty programistyczne. Najważniejsze wyniki wskazują, że Jetpack Compose zwiększa efektywność programistyczną, poprawia wydajność aplikacji i zapewnia wsparcie dla najnowszych widoków tworzonych przez Google. Wnioski potwierdzają, że Jetpack Compose jest kluczowym narzędziem w nowoczesnym ekosystemie rozwoju Android.

Słowa kluczowe: Jetpack Compose, Android, deklaratywne programowanie, zarządzanie stanem, animacje, Kotlin.

1. Wprowadzenie

Celem artykułu jest dostarczenie kompleksowego przeglądu Jetpack Compose, podkreślając jego znaczenie i wpływ na rozwój aplikacji na platformę Android. Artykuł ma za zadanie zaprezentować zarówno teoretyczne podstawy, jak i praktyczne aspekty użycia Jetpack Compose, aby programiści mogli efektywnie wykorzystać tę technologię w swoich projektach.

Obiektem badań artykułu jest Jetpack Compose – nowoczesna biblioteka UI dla platformy Android, opracowana przez Google. Biblioteka ta wprowadza deklaratywne podejście do budowy UI, w pełni zintegrowane z językiem programowania Kotlin.

Metody badawcze obejmują analizę literatury przedmiotu oraz praktyczne eksperymenty programistyczne. Analiza literatury opiera się na publikacjach takich jak "Jetpack Compose 1.5 Essentials" Neila Smytha oraz "Android UI Development with Jetpack Compose" Thomasa Künnetha. Eksperymenty obejmują tworzenie komponentów UI, zarządzanie stanem oraz integrację z istniejącymi aplikacjami Android.

Jetpack Compose wprowadza nowoczesne, deklaratywne podejście do tworzenia interfejsów użytkownika, co zwiększa efektywność programistyczną, poprawia wydajność

aplikacji i zapewnia pełne wsparcie dla Material Design 3. Biblioteka ta, zintegrowana z Kotlin, ułatwia tworzenie responsywnych i atrakcyjnych wizualnie interfejsów, dostosowanych do różnych urządzeń i form faktorów, co czyni ją kluczowym narzędziem w nowoczesnym ekosystemie rozwoju Android.

2. Deklaratywne Tworzenie Interfejsów Użytkownika

Paradygmat deklaratywny w programowaniu polega na opisywaniu pożądanych wyników bez konieczności szczegółowego definiowania kroków potrzebnych do ich osiągnięcia. W kontekście tworzenia interfejsów użytkownika oznacza to, że programista definiuje, jak interfejs ma wyglądać w danym stanie, a nie jak ten stan osiągnąć. Jetpack Compose wykorzystuje to podejście, umożliwiając programistom definiowanie interfejsów użytkownika za pomocą funkcji, które opisują strukturę i wygląd UI w sposób deklaratywny. Dzięki temu kod staje się bardziej przejrzysty, łatwiejszy do zrozumienia i utrzymania.

Tradycyjne metody tworzenia interfejsów użytkownika w Androidzie opierały się na XML i imperatywnym programowaniu. Programiści musieli definiować każdy element interfejsu w plikach XML, a następnie manipulować nimi w kodzie Java lub Kotlin. Proces ten często prowadził do skomplikowanego i trudnego do utrzymania kodu, zwłaszcza w dużych projektach. W przeciwieństwie do tego, Jetpack Compose umożliwia tworzenie UI bezpośrednio w Kotlinie, co eliminuje potrzebę rozdzielania logiki od struktury interfejsu. W Compose deklarujemy, jak UI powinno wyglądać w danym stanie, a framework sam zajmuje się aktualizacją interfejsu, gdy stan się zmienia. To podejście pozwala na bardziej intuicyjne i efektywne tworzenie dynamicznych interfejsów użytkownika.

Tabela 1. Porównanie XML oraz Jetpack Compose

Cecha	**Stare Podejście (XML)**	**Jetpack Compose**
Paradygmat	Imperatywne programowanie	Deklaratywne programowanie
Definiowanie UI	Elementy UI definiowane w plikach XML	UI definiowane za pomocą funkcji composable w Kotlinie
Oddzielenie logiki od UI	Wymagane oddzielne pliki XML i kod Java/Kotlin	UI i logika zintegrowane w kodzie Kotlin
Aktualizacja UI	Ręczna aktualizacja i zarządzanie stanem	Automatyczna aktualizacja w odpowiedzi na zmiany stanu
Złożoność kodu	Skomplikowany i trudny do utrzymania w dużych projektach	Bardziej przejrzysty i łatwiejszy do utrzymania
Reaktywność	Ograniczona reakcja na zmiany stanu bez dodatkowego kodu	Naturalnie reaktywny dzięki zarządzaniu stanem w Compose
Wsparcie dla Kotlin	Kotlin używany głównie do logiki, XML do UI	Pełne wsparcie dla Kotlin, brak potrzeby używania XML
Modularność	Mniej elastyczne, trudniejsze do tworzenia wielokrotnie używanych komponentów	Łatwe tworzenie i ponowne używanie funkcji composable

Wydajność	Większe obciążenie przy dynamicznych zmianach UI	Lepsza wydajność dzięki mechanizmom recomposition i optymalizacji
Animacje	Ręczna implementacja animacji	Wbudowane wsparcie dla animacji i przejść
Nauka i adaptacja	Dłuższy czas nauki ze względu na rozdzielenie logiki i UI	Szybsza adaptacja dzięki zintegrowanemu podejściu

Źródło: opracowanie własne

3. Tworzenie i Układanie Elementów UI

Funkcja composable jest podstawowym elementem w Jetpack Compose, który umożliwia definiowanie interfejsu użytkownika w sposób deklaratywny. Każda funkcja composable jest oznaczona adnotacją @Composable, co pozwala na jej użycie w ramach hierarchii Compose. Wewnątrz tych funkcji definiujemy, jakie komponenty UI mają być wyświetlane oraz jak mają reagować na zmiany stanu i interakcje użytkownika.

Podstawowe zarządzanie komponentami w Jetpack Compose opiera się na przekazywaniu parametrów do funkcji composable. Parametry te mogą kontrolować różne aspekty komponentów UI, takie jak tekst wyświetlany na przycisku, kolor tekstu, czy też źródło obrazu[1].

Jetpack Compose udostępnia kilka podstawowych układów do organizowania komponentów UI. Kolumny (Column), wiersze (Row) i pudełka (Box) są fundamentalnymi elementami służącymi do układania komponentów w interfejsie użytkownika.

Kolumny pozwalają na umieszczanie komponentów jeden pod drugim, tworząc układ pionowy. Można definiować parametry takie jak wyrównanie elementów i odstępy między nimi. Wiersze umożliwiają ułożenie komponentów obok siebie w układzie poziomym, z podobnymi opcjami konfiguracji. Pudełka oferują bardziej elastyczne podejście do pozycjonowania elementów, pozwalając na nakładanie komponentów jeden na drugi i precyzyjne ustawianie ich pozycji[2].

Jetpack Compose zapewnia również zaawansowane układy, które pozwalają na tworzenie bardziej złożonych i dynamicznych interfejsów użytkownika.

Przepływy (FlowRow, FlowColumn) umożliwiają dynamiczne rozkładanie komponentów w zależności od dostępnej przestrzeni. Pagery (Pager) są używane do tworzenia interfejsów z przesuwanymi stronami, idealnych do implementacji karuzel obrazów lub kart z informacjami. Listy (LazyColumn, LazyRow) to wydajne komponenty do wyświetlania

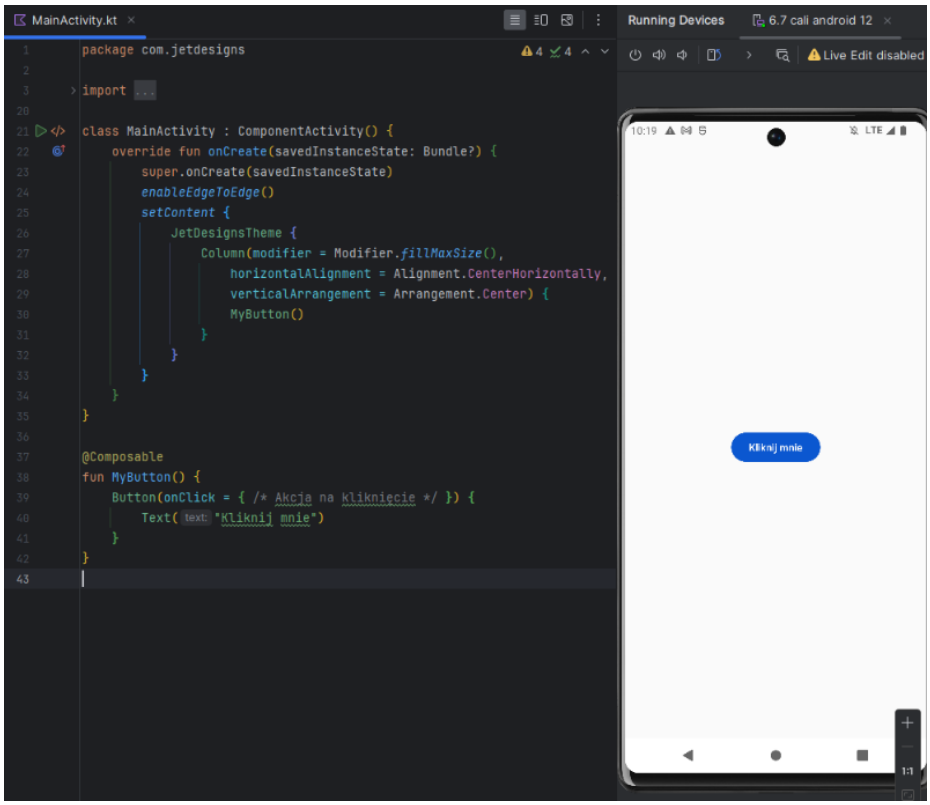
[1] Smyth Neil, Jetpack Compose 1.5 Essentials: Developing Android Apps with Jetpack Compose 1.5, Android Studio, and Kotlin, Payload Media, 5 stycznia 2024. s.30-32

[2] https://developer.android.com/develop/ui/compose/layouts/basics (dostęp 07.06.2024)

dużych zestawów danych w postaci list, które mogą być dynamicznie ładowane i aktualizowane[3].

Po wprowadzeniu do struktury możemy przyglądnąć się implementacji paru podstawowych elementów:

Przyciski w Compose są deklarowane przy użyciu funkcji Button. Ta funkcja umożliwia definiowanie akcji na kliknięcie przycisku oraz dostosowywanie jego wyglądu poprzez przekazywanie parametrów. Na przykład, parametr onClick określa, co ma się stać po kliknięciu przycisku, natomiast inne parametry mogą kontrolować jego kolor, rozmiar i inne właściwości wizualne. Funkcja MyButton może być stworzona w celu wyświetlenia przycisku, który wykonuje określoną akcję po kliknięciu. Parametry tej funkcji mogą obejmować tekst wyświetlany na przycisku oraz akcję, która ma być wykonana:

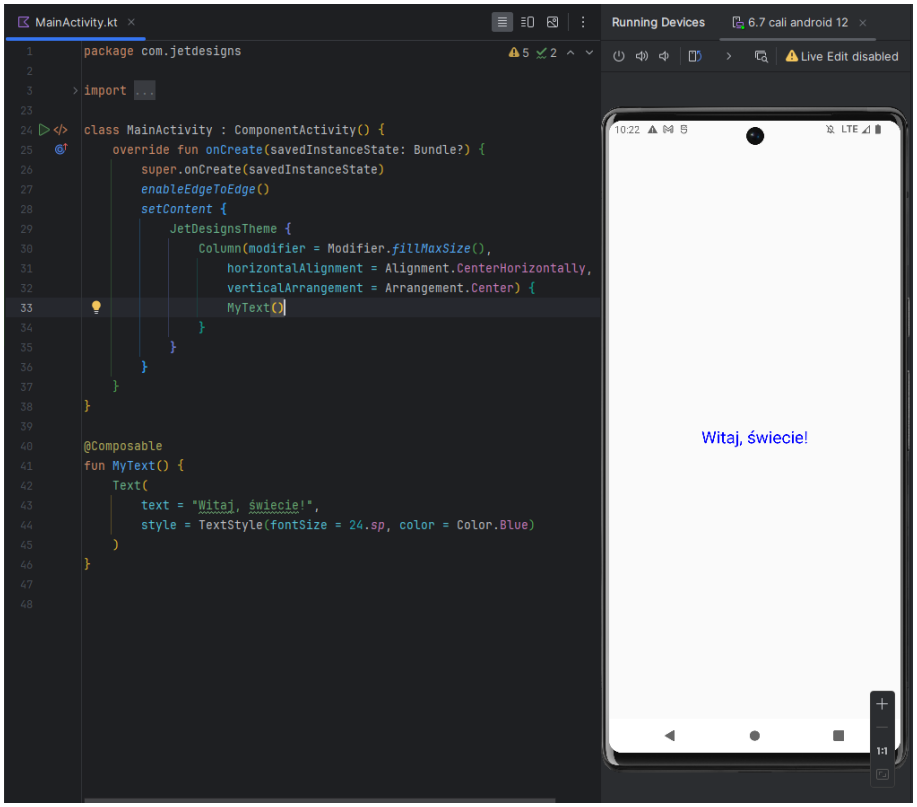


Rysunek 1. Funkcja MyButton stworzona dla wyświetlenia guzika
Źródło: opracowanie własne.

Teksty są zarządzane za pomocą funkcji Text, która umożliwia wyświetlanie różnorodnych treści tekstowych. Funkcja Text przyjmuje parametry, takie jak text, który określa wyświetlany tekst, oraz style, który definiuje styl tekstu, w tym rozmiar czcionki, kolor i inne właściwości. Funkcja MyText może być stworzona w celu wyświetlenia tekstu z określonym stylem, takim

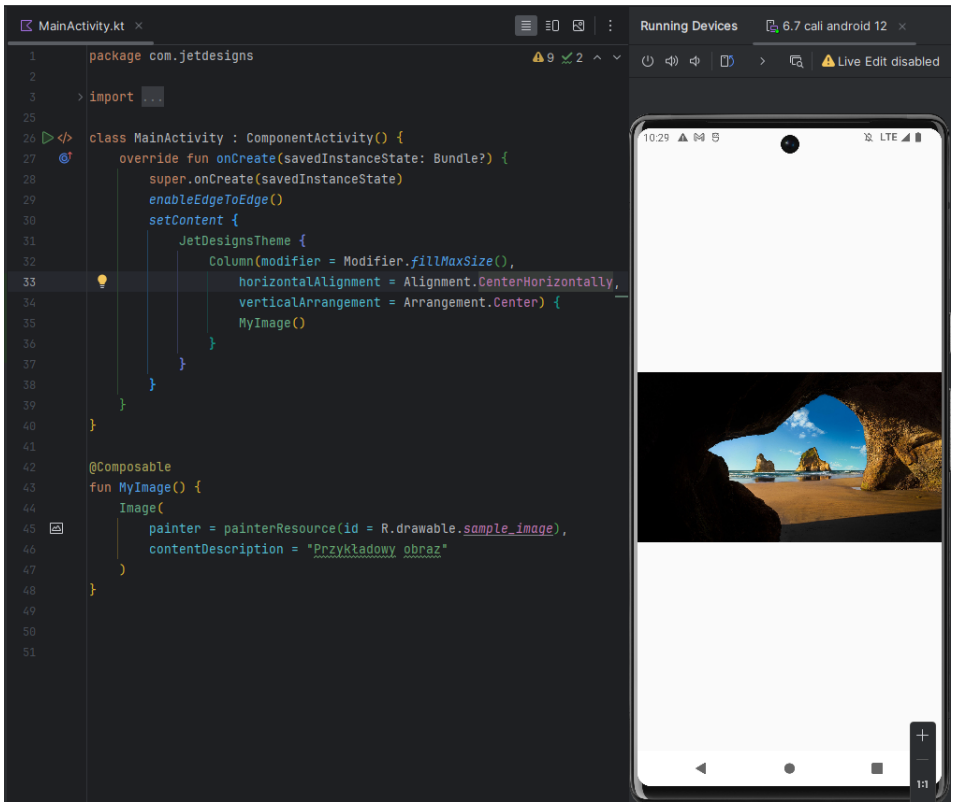
[3] https://developer.android.com/develop/ui/compose/layouts/flow (dostęp 07.06.2024)

jak kolor i rozmiar czcionki. Parametry tej funkcji pozwalają na dostosowanie wyglądu tekstu do potrzeb aplikacji.



Rysunek 2. Funkcja MyText stworzona dla wyświetlenia tekstu
Źródło: opracowanie własne.

Obrazy można dodawać do interfejsu przy użyciu funkcji Image, która pozwala na wstawianie i modyfikowanie obrazów z zasobów lokalnych lub z sieci. Parametry tej funkcji mogą obejmować źródło obrazu oraz jego właściwości, takie jak rozmiar i kształt. Funkcja MyImage może być stworzona w celu wyświetlenia obrazu z lokalnych zasobów. Parametry tej funkcji mogą obejmować identyfikator zasobu obrazu oraz właściwości modyfikujące jego wygląd, takie jak wymiary.



Rysunek 3. Funkcja MyImage stworzona dla wyświetlenia obrazu
Źródło: opracowanie własne.

4. Zarządzanie Stanem w Jetpack Compose

Zarządzanie stanem jest kluczowym elementem w tworzeniu dynamicznych i responsywnych interfejsów użytkownika. W Jetpack Compose, stan odnosi się do wszelkich danych, które wpływają na to, jak UI wygląda i jak reaguje na interakcje użytkownika. Przez deklaratywne podejście do programowania, Compose automatycznie aktualizuje UI w odpowiedzi na zmiany stanu, co upraszcza proces tworzenia aplikacji[4].

Jednym z podstawowych mechanizmów zarządzania stanem w Jetpack Compose jest funkcja State. Używa się jej do deklarowania zmiennych, które mogą dynamicznie wpływać na UI. Stan można zarządzać za pomocą funkcji remember i mutableStateOf, które utrzymują i reagują na zmiany stanu w komponencie.

```
@Composable
fun Counter() {
   var count by remember { mutableStateOf(0) }

   Button(onClick = { count++ }) {
      Text("Clicked $count times")
   }
```

[4] Smyth Neil, Jetpack Compose 1.5 Essentials: Developing Android Apps with Jetpack Compose 1.5, Android Studio, and Kotlin, Payload Media, 5 stycznia 2024. s.157-160

```
}
```

Listing 1. Przykład użycia State
Źródło: opracowanie własne.

LiveData jest częścią biblioteki Android Architecture Components i jest używana do obserwowania danych, które mogą się zmieniać w czasie. Jest często używana z ViewModel do zarządzania danymi aplikacji w sposób bardziej trwały. LiveData może być łatwo zintegrowana z Jetpack Compose za pomocą funkcji observeAsState.

```
@Composable
fun LiveDataCounter(viewModel: MyViewModel) {
   val count by viewModel.count.observeAsState(0)

   Button(onClick = { viewModel.incrementCount() }) {
      Text("Clicked $count times")
   }
}
```

Listing 2. Przykład użycia LiveData
Źródło: opracowanie własne.

Oto realny przykład zastosowania State:

```
@Composable
fun LoginForm() {
   var username by remember { mutableStateOf("") }
   var password by remember { mutableStateOf("") }
   var isLoggedIn by remember { mutableStateOf(false) }

   Column {
      TextField(
         value = username,
         onValueChange = { username = it },
         label = { Text("Username") }
      )
      TextField(
         value = password,
         onValueChange = { password = it },
         label = { Text("Password") },
         visualTransformation = PasswordVisualTransformation()
      )
      Button(onClick = { isLoggedIn = authenticate(username, password) }) {
         Text("Login")
      }
      if (isLoggedIn) {
         Text("Logged in successfully!")
      }
   }
}
```

Listing 3. Formularz logowania ze State
Źródło: opracowanie własne.

Oto realny przykład zastosowania LiveData:

```
@Composable
fun ItemList(viewModel: ItemViewModel) {
   val items by viewModel.items.observeAsState(listOf())

   Column {
      Button(onClick = { viewModel.addItem("New Item") }) {
         Text("Add Item")
      }
      LazyColumn {
         items(items.size) { index ->
            Text(items[index])
         }
      }
   }
}
```

Listing 4. Lista elementów z LiveData
Źródło: opracowanie własne.

5. Animacje i Przejścia

Zarządzanie stanem jest kluczowym elementem w tworzeniu dynamicznych i responsywnych interfejsów użytkownika. W Jetpack Compose, stan odnosi się do wszelkich danych, które wpływają na to, jak UI wygląda i jak reaguje na interakcje użytkownika. Przez deklaratywne podejście do programowania, Compose automatycznie aktualizuje UI w odpowiedzi na zmiany stanu, co upraszcza proces tworzenia aplikacji.

W Jetpack Compose tworzenie animacji jest niezwykle intuicyjne dzięki wbudowanym funkcjom, które pozwalają na animowanie właściwości takich jak pozycja, rozmiar, kolor czy przejrzystość. Na przykład, funkcja animateFloatAsState umożliwia płynne animowanie wartości zmiennoprzecinkowych. Funkcje takie jak animateColorAsState, animateDpAsState i animateSizeAsState pozwalają na animowanie innych typów danych. Dla bardziej zaawansowanych animacji, Compose oferuje narzędzia takie jak Transition i AnimatedVisibility, które umożliwiają animowanie przejść między stanami oraz dynamiczne włączanie i wyłączanie elementów UI[5].

Przykładem prostej animacji może być zmiana koloru przycisku w odpowiedzi na kliknięcie:

[5] https://developer.android.com/develop/ui/compose/animation/quick-guide (dostęp 07.06.2024)

```
@Composable
fun ColorChangingButton() {
    var isClicked by remember { mutableStateOf(false) }
    val backgroundColor by animateColorAsState(
        targetValue = if (isClicked) Color.Red else Color.Blue
    )

    Button(
        onClick = { isClicked = !isClicked },
        colors = ButtonDefaults.buttonColors(backgroundColor)
    ) {
        Text("Kliknij mnie")
    }
}
```

Listing 5. Zmiana koloru przycisku w odpowiedzi na kliknięcie
Źródło: opracowanie własne.

Innym przykładem jest użycie AnimatedVisibility do animowania widoczności elementu:

```
@Composable
fun AnimatedVisibilityDemo() {
    var isVisible by remember { mutableStateOf(true) }

    Column {
        Button(onClick = { isVisible = !isVisible }) {
            Text("Przełącz widoczność")
        }
        AnimatedVisibility(visible = isVisible) {
            Text("Widoczny tekst")
        }
    }
}
```

Listing 6. Animowanie widoczności elementu
Źródło: opracowanie własne.

W zakresie przejść między ekranami, Jetpack Compose oferuje AnimatedContent, które umożliwia animowanie zmian w zawartości komponentu w sposób płynny. Można na przykład animować przejście między dwoma różnymi widokami w aplikacji:

```
@Composable
fun AnimatedContentDemo() {
   var currentContent by remember { mutableStateOf("Content 1") }

   Column {
      Button(onClick = { currentContent = if (currentContent == "Content 1") "Content 2" else "Content 1" })
{
         Text("Zmień zawartość")
      }
      AnimatedContent(targetState = currentContent) { targetContent ->
         Text(targetContent)
      }
   }
}
```

Listing 7. Animowanie zmian w zawartości komponentu w sposób płynny
Źródło: opracowanie własne.

Najlepsze praktyki w zakresie animacji w Jetpack Compose obejmują używanie animacji do poprawienia interakcji użytkownika, a nie do ozdobników, które mogą rozpraszać uwagę. Animacje powinny być płynne i nieprzesadzone, aby nie obciążać zasobów urządzenia i nie spowalniać działania aplikacji. Kluczowym aspektem jest również testowanie animacji na różnych urządzeniach, aby upewnić się, że działają one poprawnie na wszystkich docelowych platformach.

Dobrą praktyką jest także utrzymywanie spójności w animacjach w całej aplikacji. Używanie standardowych wzorców i przejść, które są intuicyjne i przewidywalne dla użytkowników, pomaga stworzyć spójne i przyjazne doświadczenie użytkownika. Ponadto, warto korzystać z narzędzi do podglądu i debugowania animacji, które są dostępne w Android Studio, aby precyzyjnie dostosować animacje do potrzeb aplikacji.

6. Interoperacyjność z Istniejącymi Aplikacjami

Integracja Jetpack Compose z istniejącymi projektami XML to jeden z kluczowych aspektów, który ułatwia stopniową migrację aplikacji na nowe technologie. Compose pozwala na płynne włączenie nowoczesnych komponentów do aplikacji, które zostały zbudowane przy użyciu tradycyjnych metod opartych na XML i imperatywnym programowaniu. Dzięki temu, programiści mogą korzystać z zalet Compose bez konieczności całkowitej przebudowy swoich aplikacji.

Jednym z podstawowych sposobów integracji Jetpack Compose z istniejącymi projektami XML jest użycie komponentu ComposeView, który pozwala na wstawianie elementów Compose do układów XML. ComposeView może być dodany bezpośrednio do pliku XML lub

programowo w kodzie Java lub Kotlin. Oto przykład dodania ComposeView do istniejącego układu XML:

```
<androidx.compose.ui.platform.ComposeView
   android:id="@+id/composeView"
   android:layout_width="match_parent"
   android:layout_height="wrap_content"
/>
```

Listing 8. Użycie ComposeView
Źródło: opracowanie własne.

W kodzie Kotlin można następnie skonfigurować ComposeView i zdefiniować jego zawartość za pomocą funkcji setContent:

```
val composeView = findViewById<ComposeView>(R.id.composeView)
composeView.setContent {
   MyComposableFunction()
}
```

Listing 9. Definiowanie ComposeView za pomocą funkcji setContent
Źródło: opracowanie własne.

Dzięki tej metodzie, można łatwo wprowadzać elementy Compose do istniejących widoków, korzystając z nowoczesnych funkcji deklaratywnych bez konieczności całkowitego porzucenia starego kodu XML. Na przykład, w projekcie e-commerce, można wprowadzić nowoczesny interfejs koszyka zakupowego za pomocą Compose, podczas gdy reszta aplikacji pozostaje oparta na XML.

Ponadto, Jetpack Compose wspiera interoperacyjność z widokami opartymi na XML poprzez AndroidView, który pozwala na wstawianie tradycyjnych widoków w komponentach Compose. To dwukierunkowe podejście umożliwia korzystanie z najlepszych elementów obu podejść, co jest szczególnie przydatne podczas stopniowej migracji lub w projektach hybrydowych, gdzie niektóre części interfejsu wymagają zachowania istniejących technologii.

Przykładowe zastosowanie AndroidView w komponencie Compose wygląda następująco:

```
@Composable
fun LegacyView() {
   AndroidView(
      factory = { context ->
         LayoutInflater.from(context).inflate(R.layout.legacy_view, null, false)
      }
```

```
    )
}
```

Listing 10. Zastosowanie AndroidView w komponencie Compose
Źródło: opracowanie własne.

7. Podsumowanie

Jetpack Compose wprowadza rewolucyjne podejście do tworzenia interfejsów użytkownika na platformie Android, zastępując tradycyjne metody oparte na XML nowoczesnym, deklaratywnym programowaniem. Dzięki integracji z językiem Kotlin, Compose upraszcza tworzenie, zarządzanie i aktualizowanie interfejsów użytkownika, co znacząco zwiększa efektywność programistyczną i wydajność aplikacji. Artykuł szczegółowo omawia teoretyczne podstawy Jetpack Compose, praktyczne aspekty jego użycia, zarządzanie stanem oraz animacje, które przyczyniają się do tworzenia bardziej responsywnych i interaktywnych aplikacji. Dodatkowo, interoperacyjność Jetpack Compose z istniejącymi projektami XML umożliwia płynną migrację i integrację z tradycyjnymi metodami tworzenia UI. Wyniki badań potwierdzają, że Jetpack Compose jest nie tylko elastycznym i wydajnym narzędziem, ale także kluczowym elementem nowoczesnego ekosystemu rozwoju aplikacji na Android.

Literatura

1. Künneth Thomas, *Android UI Development with Jetpack Compose - Second Edition: Bring declarative and native UI to life quickly and easily on Android using Jetpack Compose and Kotlin*, Packt Publishing, 3 listopad 2023.
2. Smyth Neil, *Jetpack Compose 1.5 Essentials: Developing Android Apps with Jetpack Compose 1.5, Android Studio, and Kotlin*, Payload Media, 5 stycznia 2024.

Źródła internetowe

1. https://developer.android.com/codelabs/jetpack-compose-basics#0 (dostęp: 07.06.2024).
2. https://developer.android.com/develop/ui/compose/documentation (dostęp: 07.06.2024).

Hubert Futoma, Krystian Kiełbasa, Oskar Niedziałek
Studenckie Koło Naukowe Informatyków „KOD”

dr inż. Bartosz Trybus
Opiekun Koła Naukowego

Rozwój algorytmów kryptograficznych od starożytności do współczesności

Streszczenie

Celem niniejszego artykułu jest przedstawienie zachodzącej ewolucji algorytmów kryptograficznych. Skupia się on na 3 zasadniczych okresach: starożytności, średniowieczu i renesansie oraz czasach nowożytnych. Obserwowane jest przejście z prostych technik podstawieniowych do bardziej zaawansowanych szyfrów wieloalfabetycznych, aż po skomplikowane algorytmy blokowe wykorzystujące wielokrotne przekształcenia. Szerzej zostały omówione między innymi takie techniki jak: szyfr Cezara, szyfr Vigenère’a, AES, DES i RSA, a także różnice pomiędzy algorytmami symetrycznymi oraz asymetrycznymi. Techniki kryptograficzne powstałe na przestrzeni ostatnich 50 lat cechuje znaczny wzrost zaawansowania, co przekłada się na wyższy poziom bezpieczeństwa szyfrowanych informacji. Dodatkowo dla każdej metody zostały przedstawione wady i zalety rozwiązania.

Słowa kluczowe: kryptografia, algorytm, szyfrowanie, klucz, bezpieczeństwo.

1. Wprowadzenie

Komunikacja międzyludzka na przestrzeni dziejów odgrywała kluczową rolę w kształtowaniu cywilizacji. Wymiana informacji, niezależnie od jej postaci, była podstawą budowania relacji, handlu i edukacji, a także tworzenia społeczeństw. Jednakże wraz z rozwojem społeczności i zwiększeniem liczby interakcji pojawiła się potrzeba ochrony niektórych informacji przed niepożądanymi odbiorcami. W odpowiedzi na to wyzwanie powstała kryptografia – nauka umożliwiająca ukrywanie informacji dla zapewnienia bezpieczeństwa komunikacji.

Kryptografia, w swojej najprostszej formie, polega na przekształcaniu wiadomości w taki sposób, aby była ona nieczytelna dla nieuprawnionych osób. Pierwsze próby szyfrowania wiadomości miały miejsce już w starożytności, a ich celem była ochrona sekretów wojskowych i dyplomatycznych. Przykładem stosowanego algorytmu jest szyfr Cezara, który polegał na przesunięciu liter alfabetu o określoną liczbę miejsc.

Wraz z rozwojem społeczności i metod komunikacji pojawiła się potrzeba stosowanie coraz bardziej zaawansowanych technik kryptograficznych. Okres średniowiecza oraz renesansu cechował się rozkwitem nauk ścisłych, co przyczyniło się do powstania nowych, bardziej

skomplikowanych algorytmów szyfrowania, między innymi szyfru Vigenère'a. Technika ta zapewniała wyższy poziom bezpieczeństwa przekazywanych informacji, co było szczególnie istotne podczas prowadzonych wojen.

Znaczenie historii kryptografii jest nie do przecenienia, ponieważ pozwala zrozumieć, jak ewoluowały metody ochrony informacji oraz jak wcześniejsze techniki i odkrycia wpłynęły na współczesne rozwiązania. Celem artykułu jest przedstawienie najpopularniejszych algorytmów szyfrujących na przestrzeni różnych okresów historycznych, wraz z omówieniem ich wad i zalet.

2. Kryptografia w starożytności

Kryptografia, jako metoda ochrony informacji, ma swoje korzenie w starożytności. Już wówczas pojawiały się pierwsze próby zaszyfrowania wiadomości, które miały na celu uniemożliwienie ich odczytania przez osoby nieuprawnione. W tym rozdziale omówione zostaną szczegółowo niektóre z najwcześniejszych technik szyfrowania, w tym szyfr Cezara oraz inne metody stosowane w starożytnym Egipcie i Grecji.

Szyfr Cezara jest jednym z najstarszych i najprostszych szyfrów podstawieniowych. Nazwa pochodzi od Juliusza Cezara, który używał tej metody do komunikacji ze swoimi dowódcami wojskowymi. Mechanizm działania szyfru Cezara polega na przesunięciu każdej litery tekstu jawnego o stałą liczbę miejsc w alfabecie.

Najpopularniejszą odmianą jest przesunięcie o trzy miejsca, zatem litera 'A' staje się 'D', 'B' staje się 'E' i tak dalej. Na rysunku 1. został przedstawiony schemat działania szyfru Cezara z przesunięciem o 3 miejsca.

Litera jawna	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
Szyfr Cezara	**D**	**E**	**F**	**G**	**H**	**I**	**J**	**K**	**L**	**M**	**N**	**O**	**P**	**Q**	**R**	**S**	**T**	**U**	**V**	**W**	**X**	**Y**	**Z**	**A**	**B**	**C**

Rysunek 1. Szyfr Cezara
Źródło: opracowanie własne.

Przykładowo po zaszyfrowaniu tą techniką słowa „PIES" otrzymamy „SLHV", podczas gdy dla słowa „KOT" szyfrogramem będzie „NRW". Jednocześnie proces deszyfrowania będzie polegał na przesunięciu liter w przeciwnym kierunku o tę samą liczbę miejsc.

Choć szyfr Cezara jest stosunkowo prosty do złamania, jego rola w rozwoju kolejnych algorytmów szyfrujących jest niepodważalna. Jako jeden z pierwszych przykładów szyfrowania w praktyce, ukazuje on rosnące znaczenie ochrony informacji już w starożytności.

W starożytnym Egipcie stosowano odmienne techniki szyfrowania niż te znane z innych cywilizacji. Zamiast typowych algorytmów, Egipcjanie wykorzystywali skomplikowane hieroglify i rzadkie znaki do ukrywania informacji. Te szyfry miały różne zastosowania, m.in.:

- ochrona tekstów religijnych, stosowana do ukrywania świętych tekstów lub imion bogów, aby uchronić je przed profanacją,
- zapewnienie spokoju zmarłym poprzez umieszczenie szyfrów na grobowcach faraonów, aby zapewnić im spokój w zaświatach.

Metoda ta bazowała na skomplikowanych znakach, których znaczenie nie było oczywiste dla osób niepowołanych. Proces ten znacząco utrudniał rozszyfrowanie informacji. W porównaniu do wcześniej opisanego szyfru Cezara celem takiego zabiegu nie była komunikacja pomiędzy 2 osobami, lecz ochrona informacji o charakterze religijnym lub duchowym.

Również starożytna Grecja przyczyniła się do rozwoju kryptografii. Jednym z najpopularniejszych rozwiązań było skytale – przyrząd w kształcie podłużnego wałka z nawiniętym pergaminem.

Istotą jego działania jest wykorzystanie dwóch identycznych walców o określonej średnicy. Nadawca wiadomości zapisuje tekst na wąskim pasie materiału, owijając go wokół cylindra. Dzięki temu litery były zapisywane w określonym porządku. Następnie, tak zapieczętowany komunikat był wysyłany do adresata. Ten poprzez wykorzystanie identycznego cylindra mógł odczytać zaszyfrowaną informację, co umożliwiało prawidłowe ułożenie liter.

Szczególnie ceniona przez Spartan technika skytale odznaczała się użytecznością na polu bitwy. Dzięki swojej prostocie zapewniała szybką i skuteczną wymianę informacji, co miało kluczowe znaczenie dla koordynacji działań wojennych.

3. Kryptografia w średniowieczu i renesansie

W okresie średniowiecza i renesansu kryptografia zyskała na znaczeniu wraz z rosnącą potrzebą zabezpieczania informacji. Rozwój nauki umożliwił opracowanie bardziej zaawansowanych algorytmów szyfrowania. W tym rozdziale omówione zostaną niektóre z najważniejszych technik szyfrowania tego okresu, takie jak szyfr Vigenère’a oraz szyfr Albertiego.

Pierwszy z nich, wynaleziony przez francuskiego kryptografa Blaise’a de Vigenère’a, stanowi jedno z najważniejszych osiągnięć w dziedzinie kryptografii klasycznej. Jest on zaliczany do grupy wieloalfabetowych szyfrów podstawieniowych, który wykorzystuje klucz do zaszyfrowania wiadomości.

Jego działanie można opisać w dwóch krokach:

2. Przygotowanie tekstu jawnego i klucza szyfrującego. Tekst jawny to oryginalna wiadomość, której treść ma zostać ukryta przed nieuprawnionym dostępem. Klucz szyfrujący pełni funkcję algorytmu determinującego przekształcenie danych. Jego długość powinna odpowiadać długości danych wejściowych, a w celu zapewnienia spójności szyfrowania, klucz jest powtarzany odpowiednią ilość razy. Jego znajomość jest potrzebna zarówno w procesie kodowania i odczytywania wiadomości.
3. Szyfrowanie wiadomości z wykorzystaniem tablicy Vigenère'a (znana również jako kwadrat Vigenère'a). Składa się ona z alfabetu liter (wymiary 26 x 26), gdzie każdy wiersz jest przesunięciem alfabetu o jedną pozycję w prawo względem poprzedniego wiersza, z cyklicznym owijaniem się wokół litery 'Z'. Tablica ta została przedstawiona na rysunku 2. Dla każdej litery tekstu jawnego oraz odpowiadającej mu litery klucza szyfrującego odczytuje się wynikowy znak.

Przykładowo, dla wiadomości „ALAMAKOTA" oraz klucza szyfrującego „KLUCZ" otrzymamy „KWUOZUZNC". Traktując wiersz tablicy za litery tekstu jawnego, a kolumny tablicy za litery klucza (można również zastosować odwrotne oznaczenie) na przecięciu odpowiadających sobie liter otrzymujemy znak zakodowanej wiadomości. Dla znaków 'A' i 'K' jest to 'K', dla 'L' i 'L' jest to 'W' i tak dalej. Oczywiście należy pamiętać o zasadzie, zgodnie z którą długość klucza musi odpowiadać długości szyfrowanej wiadomości. Dla tego przypadku, ostateczna postać wykorzystanego klucza to „KLUCZKLUC".

Szyfr ten był przez długi czas uchodził za odporny na kryptoanalizę. Dopiero w XIX wieku podatność tego algorytmu została odkryta i udokumentowana przez Friedricha Kasiskiego. Metoda złamania szyfru Vigenère'a opierała się na wykorzystaniu analizy częstotliwości liter w zaszyfrowanym tekście, co wynikało z użycia krótkiego klucza i prowadziło do powstawania charakterystycznych sekwencji. Na podstawie zidentyfikowania tych powtarzających się wzorców można było odgadnąć klucz i w konsekwencji odszyfrować wiadomość. Odkrycie to skłoniło kryptografów do stosowania długich kluczy szyfrujących i zainicjowało intensywne poszukiwania bardziej odpornych algorytmów szyfrowania.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
A	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
B	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A
C	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B
D	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C
E	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D
F	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E
G	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F
H	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G
I	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H
J	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I
K	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J
L	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K
M	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L
N	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M
O	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N
P	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
Q	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
R	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
S	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
T	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
U	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
V	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
W	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
X	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
Y	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
Z	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y

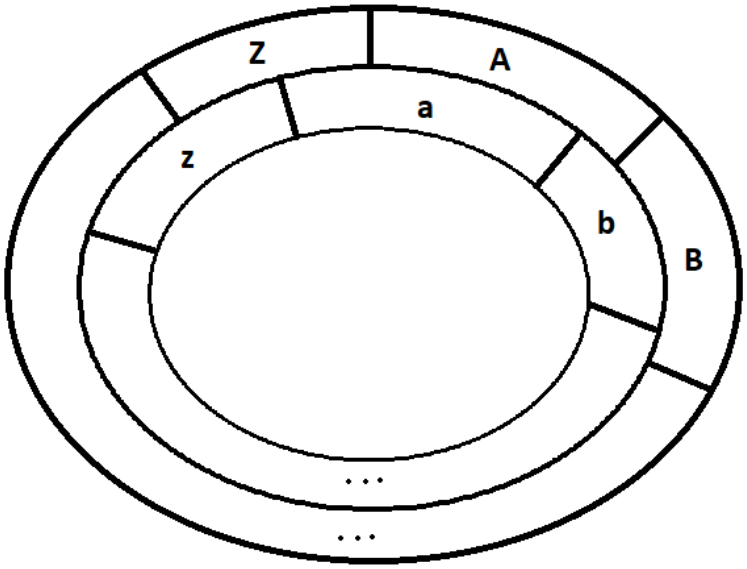
Rysunek 2. Tablica Vigenère'a
Źródło: https://en.wikipedia.org/wiki/Vigen%C3%A8re_cipher (dostęp 11.06.2024).

Kolejnym istotnym sposobem kodowania wiadomości z tamtego okresu jest szyfr Albertiego. Szyfr opracowany przez włoskiego architekta Leona Battistę Albertiego stanowi jeden z pierwszych przykładów wieloalfabetowych szyfrów podstawieniowych. W tego typu algorytmach ta sama litera w tekście jawnym może być zaszyfrowana na różne litery w zależności od użytego alfabetu.

Innowacyjnym elementem tego szyfru jest dysk Albertiego, składający się z dwóch współosiowych, obrotowych dysków. Każdy z nich zawiera pełny alfabet, a dysk wewnętrzny można obracać względem dysku zewnętrznego. Poglądowy schemat takiego dysku został zaprezentowany na rysunku 3. Możliwość obrotu wewnętrznego pozwala na zmianę alfabetu podstawienia w trakcie szyfrowania, co znacząco zwiększa poziom bezpieczeństwa. Proces ten jest kontrolowany przez klucz, określający momenty i sposoby obracania dysków.

Procedura szyfrowania przy użyciu dysku Albertiego jest następująca:

1. Oba dyski są ustawiane w położeniu początkowym w taki sposób, że litery tekstu jawnego odpowiadają literom zaszyfrowanym. Ta konfiguracja stanowi bazę dla kolejnych przekształceń.
2. W następnym kroku szyfrowana jest kolejna litera tekstu jawnego. Na przecięciu wiersza odpowiadającego literze tekstu jawnego na dysku zewnętrznym i kolumny odpowiadającej literze klucza na dysku wewnętrznym odczytywana jest litera zaszyfrowana. Ten proces przypomina działanie algorytmu Vigenère'a, z tą różnicą, że zamiast tabeli Vigenère'a, stosowane są obrotowe dyski.
3. Po zaszyfrowaniu każdej litery lub po określonej liczbie liter obracany jest dysk zgodnie z instrukcjami zawartymi w kluczu. Zmienia to wykorzystywany alfabet podstawienia.



Rysunek 3. Uproszczony schemat dysku Alberta
Źródło: opracowanie własne.

Istotną cechą wyróżniającą algorytm Albertiego jest możliwość wprowadzania dodatkowych obrotów dysków podczas szyfrowania. Oznacza to, że oprócz zmiany alfabetu podstawienia w zależności od klucza, możliwe jest również modyfikowanie go w ramach szyfrowania pojedynczej litery.

Istnieje kilka sposobów wprowadzania dodatkowych obrotów w algorytmie Albertiego:

- modyfikacja klucza – do klucza dodawane są symbole informujące o kierunku i ilości obrotów dla każdego znaku wiadomości. Przykładowo klucz mógłby być zbudowany w następujący sposób: „LKOLKO…", gdzie L oznacza litera, na którą ma zostać

przesunięty dysk wewnętrzny, K to kierunek obrotu (lewo lub prawo), a O stanowi ilość obrotów,

- umowne sterowanie – obroty dysków są regulowane według z góry ustalonych schematów, np. naprzemienne obroty w lewo i prawo lub obroty w parzystych i nieparzystych turach szyfrowania.

Wprowadzenie dodatkowych obrotów w algorytmie Albertiego zwiększa poziom bezpieczeństwa szyfrowania. Staje się ono bardziej odporne na złamanie, ponieważ kryptoanalityk musi uwzględnić nie tylko sekwencję alfabetów podstawieniowych, ale również schemat obrotów dysków. Należy jednak zaznaczyć, że ta modyfikacja niesie ze sobą również wzrost złożoności algorytmu. Generowanie szyfrogramu wymaga większej liczby operacji, a odszyfrowanie wiadomości staje się bardziej skomplikowane.

4. Kryptografia w erze nowożytnej

Era nowożytna w historii kryptografii naznaczona została rewolucją technologiczną, która przełożyła się na powstanie nowych, bardziej złożonych technik szyfrowania. Postęp w dziedzinie matematyki odegrał kluczową rolę w tym procesie, umożliwiając opracowanie rozwiązań o nieporównywalnie wyższym poziomie bezpieczeństwa w stosunku do metod stosowanych wcześniej. Zjawisko to zostało spotęgowane przez rozpowszechnienie komputerów, które stały się nieodzownym narzędziem zarówno dla kryptografów, jak i osób łamiących szyfry.

Przed przystąpieniem do analizy konkretnych algorytmów, niezbędne jest przedstawienie koncepcji technik szyfrowania z kluczem symetrycznym i niesymetrycznym.

Algorytmy symetryczne charakteryzują się wykorzystywaniem jednego tajnego klucza zarówno do szyfrowania, jak i deszyfrowania wiadomości. Oznacza to, że zarówno nadawca, jak i odbiorca muszą posiadać ten sam klucz, aby móc się ze sobą komunikować. Przykładami takiego rozwiązań są DES i AES.

Główną wadą technik szyfrowania symetrycznego jest ich wrażliwość na wyciek klucza. W przypadku jego ujawnienia, napastnik uzyskuje możliwość bezproblemowego odkodowania wiadomości, przy założeniu, że zna sposób działania algorytmu.

Aby zaradzić tej słabości, opracowywane są metody asymetryczne opierające się na parze kluczy: prywatnym (tajnym) i publicznym (jawnym). Nadawca używa klucza publicznego odbiorcy do zaszyfrowania wiadomości, a odbiorca korzysta ze swojego klucza prywatnego do jej odszyfrowania. Klucze te są ze sobą matematycznie powiązane, ale nie można ich łatwo

wywnioskować jeden z drugiego. Dzięki tej właściwości techniki asymetryczne eliminują problem dzielenia się tajnym kluczem, co znacząco zwiększa poziom bezpieczeństwa. Przykładem algorytmu niesymetrycznego jest RSA.

W dalszej części artykułu zostaną przedstawione 2 algorytmy korzystające z pojedynczego tajnego klucza – są to DES i AES oraz metoda asymetryczna RSA.

Data Encryption Standard (w skrócie DES) to standard opracowany przez IBM w latach siedemdziesiątych i zatwierdzony przez National Institute of Standards and Technology (NIST). DES reprezentuje algorytmy symetryczne, które operują na blokach danych. Oznacza to, że dane są dzielone na bloki o stałej długości (w tym przypadku to 64 bity) i każdy z bloków jest przetwarzany osobno. Dzięki temu jest on bardziej odporny na błędy transmisji danych, ponieważ uszkodzenie jednego z bloku nie wpływa na inne. Jednocześnie konsekwencją takiego podejścia jest zwiększenie opóźnienia transmisji danych, ponieważ każdy blok przed przesłaniem musi być zaszyfrowany.

Schemat działania algorytmu jest następujący:

- podział danych wejściowych na bloki o wielkości 64 bitów,
- poddanie każdego bloku permutacji początkowej przy pomocy tabeli IP,
- podział bloku na 2 równe części – prawy (P) i lewy (L) blok o długości 32 bitów,
- wybranie 56 bitów z 64-bitowego klucza głównego w wyniku permutacji PC-1 i rozdzielenie ich na dwie równe części po 28 bitów każda,
- 16-krotne wykonanie funkcji Feistela. Proces ten zostanie szerzej omówiony w dalszej części artykułu,
- po wykonaniu 16 rund, lewa i prawa połowa danych jest łączona za pomocą operacji XOR,
- poddanie otrzymanego bloku permutacji końcowej, która jest odwrotnością permutacji początkowej.

Funkcja Feistela składa się z następujących działań:

1. Wybranie 48-bitowego podklucza. Bity połówek są poddawane przesunięciu bitowemu w lewo o jeden bit (dla 1., 2., 9. i 16. rundy) lub dwa bity dla pozostałych przypadków. Następnie połówki są łączone w 56-bitowy blok, który jest przekształcany przez permutację PC-2, która prowadzi do odrzucenia 8 bitów.
2. Rozszerzenie prawej (P) połówki bloku danych. 32 bity prawego bloku są rozszerzane do 48 bitów za pomocą permutacji rozszerzającej.

3. Wykonanie sumowania na wcześniej rozszerzonym bloku prawej połówki i kluczu dla danej rundy. Wykorzystana operacja to XOR.
4. Podział wynikowego bloku na osiem 6-bitowych części. Następnie każda z nich jest podawana na odpowiedni S-Blok czego wynikiem są 4 bity wyjściowe.
5. Łączenie danych wyjściowych z S-Bloków w 32-bitową tablicę. Nowo utworzony blok jest poddawany permutacji w P-Bloku.
6. Dodanie otrzymanego bloku do lewego (L) bloku. Wynik tej operacji staje się nowym prawym (P) blokiem, podczas gdy poprzedni prawy (P) blok staje się lewym (L) bokiem.

Każda ze wspomnianych wcześniej tabel permutacji została przedstawiona pod tym adresem[1]. Warto zaznaczyć, że permutowanie bloków na początku i końcu algorytmu nie miało na celu zwiększenia bezpieczeństwa. Ich podstawowym zadaniem było ułatwienie wprowadzania danych do maszyn szyfrujących.

Pomimo 64-bitowej długości klucza głównego, jednie 56 bitów jest faktycznie wykorzystywanych do generowania kluczy rundowych. Pozostałych 8 służy do kontroli parzystości. Niewielka długość 56-bitowego klucza (w kontekście współczesnych standardów informatyki) czyni algorytm DES podatnym na ataki brute force, co uniemożliwia jego rekomendowanie jako bezpiecznego rozwiązania.

W odpowiedzi na niedoskonałości DES-a został wprowadzony w 2001 r. nowy standard szyfrowania – Advanced Encryption Standard (w skrócie AES). Operuje on na blokach danych o długości 128 bitów i umożliwia stosowanie kluczy o trzech różnych długościach: 128, 192 oraz 256 bitów.

Proces szyfrowania wiadomości przy użyciu AES składa się z następujących etapów[2]:

- podział danych wejściowych na bloki o wielkości 128 bitów, które następnie reprezentowane są jako macierze stanu o wymiarach 4 na 4 bajtów (gdzie każdy element macierzy stanowi jeden bajt),
- realizacja rundy inicjującej, w której każdy bajt bloku jest sumowany przy pomocy operacji XOR z odpowiadającym mu bajtem pierwszego podklucza,
- przeprowadzenie 9/11/13 rund głównych w zależności od długości stosowanego klucza głównego dla szyfrowania odpowiednio 128-bitowego, 192-bitowego i 256-bitowego. Przykładowa runda główna zostanie zaprezentowana w dalszej części artykułu,

[1] https://www.crypto-it.net/pl/symetryczne/des.html (dostęp: 11.06.2024).
[2] https://www.crypto-it.net/pl/symetryczne/aes.html (dostęp: 11.06.2024).

- wykonanie rundy kończącej, która jest podobna do rundy głównej, ale pomijana jest operacja mnożenia kolumn MC.

Aby móc przeprowadzić te operacje potrzebne jest wcześniejsze wygenerowanie kluczy dla poszczególnych rund poprzez rozszerzenie klucza 128-bitowego do 176 bajtów (11 rund, w tym runda początkowa, 9 głównych i końcowa, po 16 bajtów macierzy stanów), 192-bitowego do 208 bajtów (13 rund po 16 bajtów) i 256-bitowego do 240 bajtów (15 rund po 16 bajtów). Pierwsze bajty nowo otrzymanego klucza to po prostu kopia oryginalnego klucza głównego. Kolejne bajty rozszerzonego klucza generowane są iteracyjnie, aż do uzyskania pożądanej liczby bajtów. W dalszej części artykułu przedstawione zostaną kroki algorytmu generowania kluczy rundowych właśnie dla 128-bitowego (n-bitowego) klucza głównego, ponieważ redukuje to liczbę wykonywanych operacji.

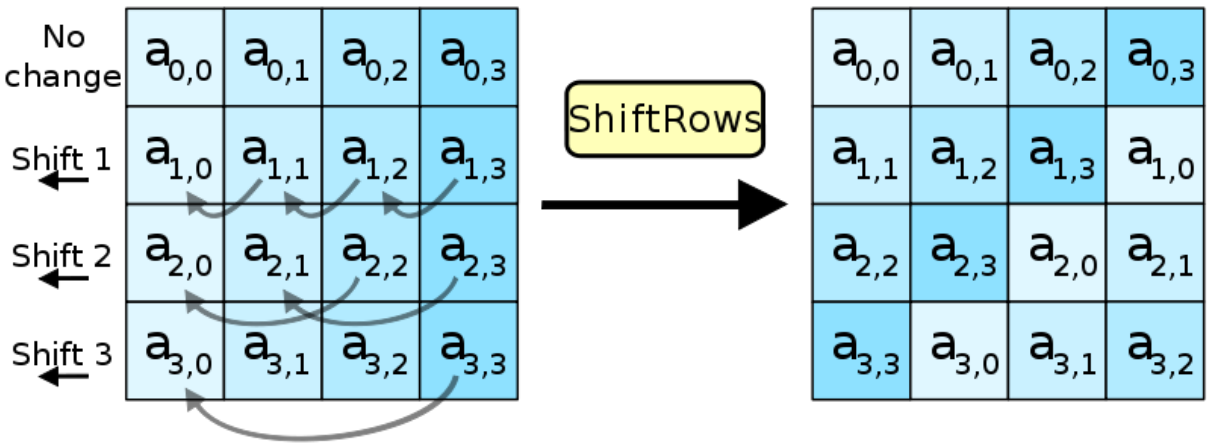
Pierwsza część pętli to utworzenie 4 kolejnych bajtów klucza poprzez:

1. Pobranie 4 ostatnich bajtów aktualnego rozszerzonego klucza do wektora. Te bajty stanowią podstawę dla generowania kolejnych 4 bajtów klucza rundowego.
2. Wykonanie rotacji bajtów w wektorze o jedną pozycję w lewo. Skrajnie lewy bajt zostaje przesunięty na skrajnie prawą pozycję, a pozostałe bajty są przesunięte o jedno miejsce w lewo.
3. Zastąpienie każdego bajtu w wektorze innym bajtem z tablicy S-Box Rijndaela. Każdy bajt jest zamieniany na inny bajt zgodnie z tabelą S-Box, która wprowadza nielinearność do procesu generowania klucza.
4. Operacja Rcon. Do najbardziej lewego bajtu w wektorze dodawany jest XOR z wartością 2 do potęgi (numer aktualnej iteracji - 1).
5. Sumowanie XOR otrzymanego 4-bajtowego wektora z 4-bajtowym blokiem danych. Blok danych pobierany jest z n bajtów przed aktualnym końcem rozszerzonego klucza. Otrzymany wynik zostaje dodany do rozszerzonego klucza.

Druga część pętli generuje pozostałe 12 bajtów klucza. W tym celu wykonywana jest operacja XOR na 4 ostatnich bajtach tworzonego klucza z 4-bajtowym blokiem danych pobranym z n bajtów przed aktualnym końcem rozszerzonego klucza. Otrzymany wynik jest następnie dołączany do rozszerzonego klucza. Uzyskany w ten sposób klucz rozszerzony jest wykorzystywany w kolejnych rundach szyfrowania AES.

Jednocześnie wcześniej wspomniana runda główna składa się z następujących operacji wykonywanych w przedstawionej kolejności:

1. Zamiana bajtów (Substitute Bytes – SB). Każdy bajt bloku danych jest zastępowany innym bajtem zdefiniowanym w tablicy S-box.
2. Przesunięcie wierszy (Shift Rows - SR). Bajty w trzech ostatnich wierszach macierzy stanu są przesuwane cyklicznie w lewo. Bajty z pierwszego wiersza pozostają na swoich pozycjach, a bajty z kolejnych wierszy są przesuwane odpowiednio o 1, 2 i 3 pozycje. Operacja ta zapewnia rozpraszanie danych w obrębie macierzy stanu. Wizualizacja tego działania została przedstawiona na rysunku 4.
3. Mieszanie kolumn (Mix Columns - MC). Każda kolumna macierzy stanu jest mnożona przez stałą macierz o wymiarach 4 x 4. Operacja ta wprowadza dalszą nielinearność do procesu szyfrowania i zapewnia mieszanie bitów w kolumnach macierzy stanu.
4. Dodawanie klucza rundowego (Add Round Key - AR). Do każdego bajtu bloku danych dodawany jest (operacja XOR) odpowiedni bajt podklucza rundowego. Operacja ta łączy klucz rundowy z blokiem danych, zwiększając złożoność szyfrowania.



Rysunek 4. Operacja Shift Rows
Źródło: https://en.m.wikipedia.org/wiki/Advanced_Encryption_Standard (dostęp 11.06.2024).

Algorytmy AES i DES, podobnie jak inne algorytmy szyfrujące, obejmują wiele operacji matematycznych, przez co ich implementacja może być podatna na błędy programistyczne. Nawet pojedyncze nieprawidłowe przekształcenie może spowodować, że otrzymany szyfrogram będzie różnił się od oczekiwanego, uniemożliwiając prawidłowe odszyfrowanie wiadomości.

Ostatnią omawianą metodą szyfrowania jest algorytm RSA. Został on opracowany w 1977 roku przez trzech kryptografów: Rona Rivesta, Adiego Shamira i Leonarda Adlemana. Nazwa metody wywodzi się od pierwszych liter nazwisk jego twórców.

Innowacyjność tego rozwiązania polegała na rozwiązaniu problemu bezpiecznej dystrybucji klucza tajnego, który stał się coraz bardziej istotny w latach 70. XX wieku. Technika RSA,

będąca algorytmem asymetrycznym, skutecznie eliminowała ten problem. Obecnie powszechnie stosowane długości kluczy to 1024 bity, 2048 bitów i 4096 bitów, przy czym 4096-bitowy klucz zapewnia najwyższy poziom bezpieczeństwa.

Algorytm generowania klucza publicznego i prywatnego jest następujący[3]:

1. W pierwszej kolejności wybierane są dwie różne liczby pierwsze, które oznaczymy przez p i q. Liczby te powinny składać się z podobnej liczby bitów.
2. Następnie obliczane jest n jako iloczyn wcześniej wspomnianych liczb. Długość n to długość klucza RSA.
3. Kolejno wyznaczana jest wartości funkcji Eulera F. Jest to iloczyn liczb pierwszych pomniejszonych o 1.
4. Spośród zakresu 1 do F dobierana jest losowo liczba e. Jednocześnie musi być ona względnie pierwsza względem F.
5. W ostatnim kroku algorytmu wyliczana jest liczba d. Jej iloczyn z e daje resztę z dzielenia przez F równą 1, co można zapisać: $d * e = 1\ (mod\ F)$. Liczba d jest obliczona z wykorzystaniem rozszerzonego algorytmu Euklidesa.

Klucz publiczny stanowi para (e,n), a prywatny to (d,n). Klucz publiczny może być bezpiecznie udostępniony, podczas gdy klucz prywatny musi być przechowywany w tajemnicy przez właściciela.

Dla tak przygotowanych kluczy, wiadomość M jest szyfrowana do szyfrogramu C z użyciem klucza publicznego. Operacja ta przyjmuje postać: $C = M^e (mod\ n)$. Aby zmniejszyć złożoność obliczeniową wykorzystywane jest twierdzenie Eulera, które pozwala na sprowadzenie potęgowania modulo do mnożenia reszt z dzielenia.

Odszyfrowanie do oryginalnej wiadomości M odbywa się analogicznie, ale z wykorzystaniem klucza prywatnego: $M = C^d (mod\ n)$.

Rozważając przykładowe liczby pierwsze p=11 i q=13:

- wartość n wynosi 143 (11 * 13),
- funkcja Eulera F przyjmuje wartość 120 [(11-1) * (13-1)],
- najmniejsza liczba e spełniającą warunek względnego pierwszeństwa z F to 7 (możliwy jest wybór większej liczby),
- zatem klucz publiczny to (e,n)=(7,143), a prywatny to (d,n)=(103,143),

[3] https://www.crypto-it.net/pl/asymetryczne/rsa.html (dostęp: 11.06.2024).

- przyjmując wiadomość M równa 9 (po konwersji znaków na reprezentację dziesiętną), szyfrogram C oblicza się jako: $M^e(mod\ n) = 9^7(mod\ 143) = 48$,
- następnie po odszyfrowaniu C otrzymana zostanie z powrotem wiadomość M równa: $C^d(mod\ n) = 48^{103}(mod\ 143) = 9$, co potwierdza poprawność procesu szyfrowania i deszyfrowania.

Należy zaznaczyć, że w rzeczywistych zastosowaniach algorytmu RSA wykorzystywane są znacznie większe liczby pierwsze. Przedstawiony przykład służy jedynie zilustrowaniu mechanizmu działania algorytmu.

Algorytm RSA odgrywa kluczową rolę w ochronie integralności i autentyczności danych w systemach cyfrowych. Służy do tworzenia kryptograficznych podpisów cyfrowych, które gwarantują, że wiadomość nie została zmieniona ani podszyta. Umożliwia również bezpieczną autentykację użytkowników w systemach, chroniąc przed nieautoryzowanym dostępem.

Ponadto RSA stanowi kluczowy element szyfrowania hybrydowego, łącząc wydajność szyfrów symetrycznych z wysokim poziomem bezpieczeństwa. Znajduje również zastosowanie w ochronie oprogramowania, zarządzaniu kluczami kryptograficznymi i wzmacnianiu bezpieczeństwa protokołów sieciowych, takich jak HTTPS i VPN.

5. Podsumowanie

Analiza historycznych przykładów obrazuje ewolucję algorytmów szyfrujących na przestrzeni lat. Początkowe, proste szyfry podstawieniowe ustąpiły miejsca bardziej złożonym metodom wieloalfabetowym. Nadejście ery komputerów bezlitośnie obnażyło słabość stosowanych technik szyfrowania. Na szczęście z pomocą przyszedł rozwój matematyki, oferując znacznie bezpieczniejsze rozwiązania. Pomimo to nieustanny wzrost mocy obliczeniowej wymusza dalsze poszukiwania udoskonaleń. Przykładem tego jest odejście od standardu DES na rzecz nowego, bezpieczniejszego AES. Choć zastosowania kryptografii nieustannie się rozszerzają, jej podstawowy cel – ochrona informacji – pozostaje niezmienny.

Należy jednak podkreślić, że nadchodząca era komputerów kwantowych może ponownie zrewolucjonizować tę dziedzinę. Wykorzystanie obliczeń kwantowych stwarza nowe wyzwania w postaci coraz łatwiejszego łamania kluczy kryptograficznych. Wymusza to opracowanie systemów kryptograficznych odpornych na tego rodzaju ataki.

Literatura

1. Karbowski M., *Podstawy kryptografii*, Wydawnictwa Helion, Gliwice 2014.

Źródła internetowe

1. https://crypto.interactive-maths.com/kasiski-analysis-breaking-the-code.html (dostęp: 11.06.2024).
2. https://www.crypto-it.net/pl/symetryczne/aes.html (dostęp: 11.06.2024).
3. https://www.crypto-it.net/pl/symetryczne/des.html (dostęp: 11.06.2024).
4. https://www.crypto-it.net/pl/asymetryczne/rsa.html (dostęp: 11.06.2024).
5. https://www.redhat.com/en/blog/brief-history-cryptography (dostęp: 11.06.2024).
6. https://wielkahistoria.pl/szyfry-starozytnych-grekow-tak-mieszkancy-antycznej-hellady-kodowali-poufne-wiadomosci/ (dostęp: 11.06.2024).

Nabożny Maciej, Łukasz Michnik, Karol Michoński, Mateusz Skali, Szymon Jabłoński
Studenckie Koło Naukowe Informatyków „KOD”

dr inż. Bartosz Trybus
Opiekun Koła Naukowego

Kompleksowe zabezpieczenie płatności online: implementacja z Typescript i nowoczesnymi technologiami backendowymi

Streszczenie

Artykuł opisuje znaczenie TypeScript w tworzeniu bezpiecznych aplikacji backendowych. Skupia się na kluczowych funkcjach, takich jak statyczne typowanie, które pozwala na wcześniejsze wykrywanie błędów i potencjalnych luk bezpieczeństwa, co redukuje ryzyko ataków, takich jak SQL injection. Omówiono również nowoczesne funkcje JavaScript wspierane przez TypeScript, jak async/await, które ułatwiają zarządzanie operacjami asynchronicznymi. Rozbudowany system interfejsów umożliwia precyzyjne definiowanie struktur danych, zwiększając bezpieczeństwo kodu. TypeScript jest kompatybilny z narzędziami do analizy statycznej kodu, jak ESLint, co pozwala na automatyczne sprawdzanie zgodności z najlepszymi praktykami. Mechanizmy modułowości ułatwiają organizowanie kodu i minimalizują ryzyko nieautoryzowanego dostępu do danych wrażliwych. Celem artykułu jest ukazanie, jak TypeScript wspiera deweloperów w tworzeniu bezpiecznych aplikacji, oraz poprzez opisane w artykule przykładów kodów, pokazanie jak mogą być one użyte w praktyce.

Słowa kluczowe: Typescript, Systemy płatności, Bezpieczeństwo, Autoryzacja i Uwierzytelnianie.

1. Wprowadzenie

Płatności online są kluczową częścią aplikacji e-commerce, idzie za tym również bezpieczeństwo tego typu aplikacji. Chęć posiadania przez firmy wysokiej jakości bezpieczeństwa w swoich produktach jest w głównej mierze podyktowane tym, że liczba różnego rodzaju płatnośći zabieranych przez internet z roku na rok rośnie. Pomimo takiego wysokiego zapotrzebowania na posiadanie takich technologii, wiele firm w nie niewystarczającym stopniu inwestuje fundusze w takie technologie, bo uważa je za, za bardzo zaawansowane, co doprowadza do różnych ataków cybernetycznych po których nie dość, że cierpi sama firma to i przede wszystkim konsumenci, którzy tracą zaufanie do marki. Standardy bezpieczeństwa zwiększają się z czasem, jest to spowodowane rozwojem technologii na której bazują opisywane aplikacje, jak i również zwiększająca się konkurencja na rynku aplikacji e-commerce, a również z nimi oczekiwania klientów. Klienci oczekują od firm, którym powierzają swoje dane finansowe i osobiste, że będą one przetwarzane i autoryzowane w

sposób bezpieczny. Takie działania są konieczne, jeżeli firma chce zachować wysoki standard i swoje dobre imię.

Systemy obsługujące płatności coraz częściej stają się celem ataków cybernetycznych co widać na przestrzeni ubiegłych lat, gdzie takie firmy jak Facebook (oraz jego użytkownicy) byli ofiarami ataków (takich jak phishing). Podczas tych ataków wykorzystywane są różne techniki kradzieży danych, są nimi między innymi phishing, czyli wymuszenie przez oszusta, zachowań użytkownika, który w sposób dla siebie nieświadomy podaje oszustowi swoje dane wrażliwe, bądź prowokuje sytuacje, które wymuszają na użytkowniku pewne działania co umożliwia oszustowi samodzielne wykradnięcie danych, bądź zasobów finansowych. Innym przykładem takich ataków są malware i ransomware, polegają na wgraniu na użądzenie użytkownika złośliwego oprogramowania które w specyficzny dla siebie sposób wykrada dane, bądź zmusza użytkownika do zapłacenia konkretnej sumy pieniędzy, aby móc odblokować dostęp do zasobów na urządzeniu (ma to miejsce przy ransomware). Bazy danych serwisów i aplikacji e-commerce są również ofiarami ataków cybernetycznym przez ataki DDoS(Distributed Denial of Service) oraz SQL injection. Pomagają one uzyskać nieautoryzowany dostę do danych finansowych i prywantych użytkowników atakowanego systemu.

Programiści, muszą w produktach firm stosować najbardziej skuteczne rozwiązania technologiczne, aby zabezpieczyć dane użytkowników, jest to konieczne jeżeli firma chce zabezpieczyć przechowane przez siebie dane. Takie zabezpieczenia można wprowadzać dzięki technologii TypeScript, która jest nadzbiorem JavaScript. Poprzez wykorzystanie zewnętrznych bibliotek deweloperzy są w stanie zabezpieczyć dane w sposób poprawny. Zakres tych technologi potrafi być różny, jest on w głównej mierze zależny od tego co ma robić dana aplikacja, ale najczęściej obejmują one technologie związane z szyfrowaniem danych oraz różnego rodzaju zaawansowane mechanizmy autoryzacji i uwierzytelniania użytkowników systemu. Poza rozwiązaniami technologicznymi, ważna jest również edukacja użytkowników z zakresu podstaw bezpiecznego poruszania się po aplikacjach, jak i samym internecie, taka wiedza pomogałaby w uniknięciu niektórych ataków cybernetycznych na które, są narażeni.

W tym artykule omówione zostaną przykłady, przy użyciu języka TypeScript oraz nowoczesnych technologii backendowych którymi można skutecznie zabezpieczyć procesy płatności. TypeScript, będący nadzbiorem JavaScript, wprowadza typowanie statyczne, które pozwala na wcześniejsze wykrywanie błędów i potencjalnych luk bezpieczeństwa. Dzięki temu deweloperzy mogą tworzyć bardziej niezawodne i bezpieczne aplikacje.

Artykuł skupia się na szerokopojętym bezpieczeństwie programowania backendowego, które można wyróżnić na różne wyspecjalizowane podsekcje, poruszone zostały tematy szyfrowania danych poprzez opisanie takich algorytmów jak AES (Advanced Encryption Standard), który zapewnia bezpieczeństwo w przechowywaniu informacji oraz została poruszona sama implementacja systemów płatniczych w kod źródłowy TypeScript. Protokoły służące do ochrony danych wrażliwych oraz służące do zarządzania sesjami użytkowników, takie jak OAuth2 oraz JWT (Json Web Tokens), również zostały opisane w treści artykułu. W atykule zostały przedstawione przykłady kodów źródłowych, które ukazują przykładowe przypadki użycia opisanych wyżej technologii.

Celem tego artykułu jest dostarczenie kompleksowego przewodnika, który pokazuje jak zbudować bezpieczne systemy płatności przy użyciu TypeScript oraz nowoczesnych technologii backendowych. Wiedza ta jest nie tylko teoretyczna, ale również praktyczna, z naciskiem na realne zastosowania. Odpowiednie zabezpieczenia mogą znacząco zwiększyć zaufanie klientów do marki, a tym samym przyczynić się do jej sukcesu na rynku.

2. Wprowadzenie do bezpiecznego programowania backendowego

2.1. Rola TypeScript w bezpieczeństwie backendowym

Stworzony przez Microsoft, język programowania TypeScript, będący nadzbiorem języka JavaScript, daje wiele korzyści deweloperom, w kontekście tworzenia bezpiecznego backendu dla aplikacji e-commerce. Dzięki takim rozwiązaniom jak wprowadzenie statycznego typowania (które jest główną zaletą tego języka programowania), TypeScript pozwala na ułatwione wykrywanie błędów w kodzie oraz luk bezpieczeństwa. JavaScript w przeciwieństwie do swojego nadzbioru nie ma takich możliwośći i typy danych są w nim dynamiczne. Powoduje to sytuacje, gdzie typy mogą być zmieniane w czasie trwania programu co może być luką w bezpieczeństwie dla potencjalnych ataków cybernetycznych. TypeScript za to wymaga w pierwszej kolejności, aby zadeklarować zmienną, funkcję czy obiekt na jeden konkretny typ danych. Dzięki takiemu poziomowi zaawansowania TypeScript, deweloperzy nie muszą się martwić tym czy zadeklarowane przez nich zmienne będą podczas działania programu zmieniać dynamicznie swój typ ponieważ jest to niemożliwe. Daje to twórcom aplikacji większą kontrolę nad tym jak działa program oraz zmniejsza to występowanie luk w bezpieczeństwie aplikacji. W przeciwnym razie kiedy nie ma takich praktyk przy tworzeniu aplikacji e-commerce (jak ma to miejsce w przypadku pisaniu kodu w TypeScript), niepoprawne zarządzanie typami może doprowadzić do wycieków z baz danych bądź błędów logicznych. Do ataków na bazy danych można zaliczyć wstrzyknięcia SQL, są one jednymi z naczęstszych rodzajów ataków cybernetycznych, polega na wstrzyknięciu złośliwego kodu SQL do zapytania bazy danych, w skutku prowadzi do nieautoryzowanego dostępu do danych wrażliwych, które można wykraść, bądź zmdoyfikować. Statyczne typowanie w TypeScript, umożliwia wcześniejsze wykrycie takiego ataku pod postacią błędu podczas kompilacji kodu,

co sprawia, że zapytanie koniec końców nie dojdzie do skutku i atak zostanie przerwany. TypeScript, jest również technologią, która spełnia najnowsze standardy języka programowania z którego się wywodzi, czyli JavaScript'a. Są nimi async/await, które pomagają deweloperom w rozdzieleniu operacji na te synchroniczne i asynchronicze, tak aby wykonywany program nie był zatrzymywany przez nie odpowiednio zoptymalizowany kod. Operacje określane jako asynchroniczne zajmują się obsługą zapytań do bazy danych, wysyłaniem wiadomości e-mail jak i również wysyłaniem zapytań do zewnętrznego API, są one niezastąpionym elementem aplikacji webowych, jeżeli mowa o kodzie źródłowym od strony backend'u. Zastosowanie operacji asynchroniczych pomaga w optymalizacji kodu, ponieważ nie blokują one innych części kodu, tylko działają w „tle". Wpływa to pozytywnie na wydajność aplikacji oraz powoduje to, że kod jest bardziej czytelny i łatwiejszy w debugowaniu, co pozwala na szybsze wykrywanie błędów oraz problemów z bezpieczeństwem. TypeScript oferuje również rozbudowany system interfejsów, który pozwala na dokładne definiowanie struktur danych. Interfejsy mogą być używane do określania, jakie właściwości i metody powinny mieć obiekty, co dodatkowo zwiększa bezpieczeństwo kodu. Dzięki interfejsom, deweloperzy mogą tworzyć bardziej przejrzysty i zrozumiały kod, co ułatwia jego utrzymanie i rozwój. Przykładowo, definiowanie interfejsu dla obiektu reprezentującego użytkownika może wyglądać jak na listingu 2.1:

```
interface User {
  id: number;
  name: string;
  email: string;
  passwordHash: string;
}
```

Listing 2.1. Kod deklarujący interface User

Źródło: Opracowanie własne

Poprzez zastosowanie interfejsu User jak na listingu 2.1, wszelkie operacje wykonywane na instancji obiektu User, będą musiały być zgodne ze zdefiniowanym wyżej interfejsem. Pomaga to w zapobieganiu przypadkowego dodawania jak i również usuwania właściwowości zadeklarowanych w kodzie źródłowym, które mogą prowadzić do różnego rodzaju błędów podczas kompilacji oraz luk w bezpieczeństwie.

Kolejną z opisywanych zalet nadzbioru JavaScript jest jego dopasowanie z narzędziami służącymi do analizy kodu pod kątem błędów, takimi jak ESLint czy TSLint. Wcześniej wspomniane narzędzia pomagają w automatycznej analizie kodu przez środowisko programistyczne. Są one używane do wykrywania błędów, oraz naruszeń w bezpieczeństwie kodu. Integracja z systemami ciągłej integracji (CI) umożliwia automatyczne sprawdzanie kodu pod kątem zgodności z wybranymi regułami, co dodatkowo zwiększa jakość i bezpieczeństwo kodu. Przykład konfiguracji ESLint dla projektu TypeScript może wyglądać jak w przykładzie pokazanym w Listingu 2.2:

```
{
  "extends": [
    "eslint:recommended",
    "plugin:@typescript-eslint/recommended"
  ],
  "parser": "@typescript-eslint/parser",
```

```
  "plugins": ["@typescript-eslint"],
  "rules": {
    "no-unused-vars": "error",
    "no-console": "warn",
    "@typescript-eslint/no-explicit-any": "error"
  }
}
```

Listing 2.2. Kod konfiugurjący Eslint

Źródło: Opracowanie własne

TypeScript wspiera również mechanizmy modułowości, które pozwalają na lepsze organizowanie i zarządzanie kodem. Moduły umożliwiają dzielenie aplikacji na mniejsze, niezależne części, co ułatwia testowanie i utrzymanie kodu. W kontekście bezpieczeństwa, modułowość pozwala na wyraźne oddzielenie warstw logiki biznesowej od warstwy dostępu do danych, co minimalizuje ryzyko nieautoryzowanego dostępu do danych wrażliwych. Ponadto, TypeScript jest w pełni kompatybilny z ekosystemem JavaScript, co oznacza, że deweloperzy mogą korzystać z istniejących bibliotek i narzędzi JavaScript, jednocześnie czerpiąc korzyści z typowania statycznego. Wiele popularnych bibliotek i frameworków, takich jak Express, NestJS czy Sequelize, oferuje wsparcie dla TypeScript, co umożliwia tworzenie bezpiecznych i wydajnych aplikacji backendowych. Przykład integracji Express z TypeScript może wyglądać jak w listingu 2.3:

```
import express, { Request, Response } from 'express';

const app = express();
const port = 3000;

app.get('/', (req: Request, res: Response) => {
  res.send('Hello, TypeScript!');
});

app.listen(port, () => {
  console.log(`Server is running at http://localhost:${port}`);
});
```

Listing 2.3. Kod konfiugurjący serwer HTTP

Źródło: Opracowanie własne

TypeScript umożliwia tworzenie zaawansowanych testów jednostkowych i integracyjnych, które mogą być używane do weryfikacji bezpieczeństwa aplikacji. Frameworki takie jak Jest czy Mocha wspierają TypeScript, co pozwala na pisanie testów w tym samym języku, co kod aplikacji. Testy te mogą być używane do automatycznego sprawdzania, czy wszystkie komponenty aplikacji działają zgodnie z oczekiwaniami, co minimalizuje ryzyko wprowadzenia błędów podczas zmian w kodzie.

Podsumowując, TypeScript oferuje szereg zaawansowanych funkcji i narzędzi, które znacząco zwiększają bezpieczeństwo aplikacji backendowych. Dzięki statycznemu typowaniu, wsparciu dla nowoczesnych standardów JavaScript, rozbudowanemu systemowi interfejsów, kompatybilności z narzędziami do analizy statycznej kodu, mechanizmom modułowości oraz wsparciu dla testowania, TypeScript jest idealnym wyborem dla deweloperów, którzy chcą

tworzyć bezpieczne i niezawodne aplikacje. Jego integracja z ekosystemem JavaScript sprawia, że jest to narzędzie nie tylko nowoczesne, ale także praktyczne i łatwe w użyciu.

2.2. Szyfrowanie danych

Szyfrowanie danych to jeden z fundamentów zabezpieczania płatności. Jest to proces przekształcania informacji w taki sposób, że tylko uprawnione osoby mogą je odczytać. Dzięki szyfrowaniu, dane są chronione przed nieautoryzowanym dostępem nawet w przypadku ich przechwycenia przez osoby trzecie. W kontekście backendu, wszystkie dane przesyłane między serwerem a klientem muszą być odpowiednio zaszyfrowane, aby zapobiec ich przechwyceniu przez nieuprawnione osoby. W dobie rosnącej liczby ataków cybernetycznych, szyfrowanie staje się nieodzownym elementem każdego bezpiecznego systemu płatności.

Najbardziej powszechną metodą zabezpieczania transmisji danych jest użycie protokołu HTTPS, który zapewnia bezpieczne połączenie poprzez wykorzystanie SSL/TLS. Protokół ten szyfruje dane przesyłane między serwerem a przeglądarką użytkownika, co uniemożliwia ich odczytanie przez osoby trzecie. HTTPS działa na warstwie transportowej (TLS/SSL), gdzie zapewnia integralność danych oraz ich poufność. Jest to standard stosowany we wszystkich nowoczesnych aplikacjach webowych, gdzie bezpieczeństwo danych jest priorytetem. Implementacja HTTPS w aplikacji jest stosunkowo prosta i polega na skonfigurowaniu serwera do używania certyfikatów SSL/TLS. Przykład konfiguracji serwera Node.js z wykorzystaniem biblioteki https może wyglądać ja w listingu 2.4:

```
const https = require('https');
const fs = require('fs');

const options = {
  key: fs.readFileSync('path/to/private-key.pem'),
  cert: fs.readFileSync('path/to/certificate.pem'),
};

https.createServer(options, (req, res) => {
  res.writeHead(200);
  res.end('Hello, secure world!');
}).listen(443);
```

Listing 2.4. Konfiguracja certyfikatu SSL

Źródło: Opracowanie własne

Dodatkowo, dane wrażliwe, takie jak numery kart kredytowych, powinny być przechowywane w bazie danych w postaci zaszyfrowanej. Algorytm AES (Advanced Encryption Standard) jest jednym z najczęściej stosowanych algorytmów do tego celu. AES oferuje wysoki poziom bezpieczeństwa i jest stosunkowo szybki w implementacji. AES jest symetrycznym algorytmem szyfrowania, co oznacza, że ten sam klucz jest używany do szyfrowania i odszyfrowania danych. Klucz ten musi być przechowywany w bezpieczny sposób, aby zapobiec jego kompromitacji. Przykład użycia AES do szyfrowania danych w Node.js z wykorzystaniem biblioteki crypto może wyglądaćjak w listingu 2.5:

```
const crypto = require('crypto');
const algorithm = 'aes-256-cbc';
const key = crypto.randomBytes(32);
const iv = crypto.randomBytes(16);

function encrypt(text) {
  const cipher = crypto.createCipheriv(algorithm,
Buffer.from(key), iv);
  let encrypted = cipher.update(text);
  encrypted = Buffer.concat([encrypted, cipher.final()]);
  return { iv: iv.toString('hex'), encryptedData: encrypt-
ed.toString('hex') };
}

function decrypt(text) {
  const iv = Buffer.from(text.iv, 'hex');
  const encryptedText = Buffer.from(text.encryptedData, 'hex');
  const decipher = crypto.createDecipheriv(algorithm,
Buffer.from(key), iv);
  let decrypted = decipher.update(encryptedText);
  decrypted = Buffer.concat([decrypted, decipher.final()]);
  return decrypted.toString();
}

const data = "Sensitive data";
const encrypted = encrypt(data);
console.log(encrypted);
const decrypted = decrypt(encrypted);
console.log(decrypted);
```

Listing 2.5. Bezpieczne szyfrowanie i deszyfrowanie danych

Źródło: Opracowanie własne

Oprócz szyfrowania danych przechowywanych w bazie, należy również zabezpieczyć dane przesyłane między różnymi usługami w architekturze mikroserwisów. Mikroserwisy często komunikują się ze sobą poprzez różne protokoły sieciowe, co stwarza potencjalne zagrożenia związane z przechwytywaniem danych. W celu zapewnienia pełnego bezpieczeństwa w całym ekosystemie aplikacji, dane te również powinny być szyfrowane. W przypadku komunikacji między mikroserwisami, można wykorzystać szyfrowanie na poziomie transportowym (TLS) lub zastosować techniki szyfrowania na poziomie aplikacyjnym. Dobrą praktyką jest również stosowanie bezpiecznych mechanizmów zarządzania kluczami. Klucze szyfrujące powinny być przechowywane w bezpieczny sposób, na przykład w dedykowanych usługach zarządzania kluczami, takich jak AWS KMS (Key Management Service) czy HashiCorp Vault. Umożliwia to centralne zarządzanie i rotację kluczy, co zwiększa bezpieczeństwo i ułatwia zgodność z regulacjami prawnymi.

Kolejnym istotnym aspektem jest stosowanie technik haszowania dla danych, które nie muszą być odszyfrowywane, takich jak hasła użytkowników. Haszowanie polega na przekształceniu danych w unikalny ciąg znaków, który jest trudny do odwrócenia. Algorytmy takie jak bcrypt, scrypt czy Argon2 są powszechnie stosowane do haszowania haseł, zapewniając wysoki poziom bezpieczeństwa nawet w przypadku kompromitacji bazy danych.

Podsumowując, szyfrowanie danych jest kluczowym elementem zabezpieczania płatności online. Zastosowanie protokołu HTTPS, szyfrowanie danych wrażliwych za pomocą algorytmu AES oraz zabezpieczenie komunikacji między mikroserwisami to niezbędne kroki, które zapewniają integralność i poufność danych. Dodatkowo, bezpieczne zarządzanie kluczami i stosowanie technik haszowania dla haseł użytkowników zwiększają ogólny poziom bezpieczeństwa systemu. Dzięki tym praktykom, można skutecznie chronić dane użytkowników i budować zaufanie klientów do marki.

2.3. Autoryzacja i uwierzytelnianie

Autoryzacja i uwierzytelnianie to kluczowe aspekty zabezpieczania płatności. Uwierzytelnianie to proces weryfikacji tożsamości użytkownika, natomiast autoryzacja polega na nadaniu uprawnień do wykonywania określonych operacji. W kontekście aplikacji e-commerce, skuteczne zarządzanie tymi procesami jest niezbędne do ochrony danych użytkowników i zapobiegania nieautoryzowanym operacjom finansowym. Współczesne systemy płatności często korzystają z protokołów OAuth2 lub JWT (JSON Web Tokens) do zarządzania sesjami użytkowników. OAuth2 umożliwia użytkownikom udzielanie aplikacjom dostępu do swoich zasobów bez ujawniania swoich danych uwierzytelniających, co znacząco zwiększa bezpieczeństwo. OAuth2 działa poprzez wydawanie tokenów dostępu, które aplikacje mogą wykorzystać do uzyskania dostępu do chronionych zasobów użytkownika na serwerze. Proces ten składa się z kilku kroków, w tym uzyskania zgody użytkownika, wymiany kodu autoryzacyjnego na token dostępu oraz użycia tokena dostępu do autoryzowania żądań. JWT, z kolei, to kompaktowy, bezpieczny sposób na przekazywanie informacji między stronami jako obiekt JSON. JWT może być używany do autoryzacji, ponieważ tokeny mogą być weryfikowane i zaufane. Tokeny JWT są podpisywane cyfrowo za pomocą klucza tajnego lub klucza publiczno-prywatnego, co zapewnia integralność i autentyczność danych zawartych w tokenie. JWT składa się z trzech części: nagłówka, ładunku i podpisu. Nagłówek zawiera informacje o algorytmie szyfrowania i typie tokena, ładunek zawiera dane użytkownika i inne informacje, a podpis zapewnia, że token nie został zmodyfikowany.

Implementacja tych protokołów w TypeScript jest stosunkowo prosta dzięki dostępności licznych bibliotek, takich jak Passport dla OAuth2 czy jsonwebtoken dla JWT. Passport jest wszechstronnym middlewarem do uwierzytelniania w Node.js, który obsługuje różne strategie uwierzytelniania, w tym OAuth2. Jsonwebtoken to popularna biblioteka do pracy z JWT w Node.js, która umożliwia łatwe generowanie i weryfikowanie tokenów JWT. W listingu 2.6 przedstawiony został przykład implementacji uwierzytelniania za pomocą JWT w TypeScript :

```
import jwt from 'jsonwebtoken';

// Interfejs użytkownika, którego dane będą przechowywane w
tokenie
interface User {
  id: number;
  email: string;
}

// Funkcja generująca token JWT dla danego użytkownika
const generateToken = (user: User): string => {
  return jwt.sign({ id: user.id, email: user.email }, pro-
cess.env.JWT_SECRET, {
```

```
    expiresIn: '1h',
  });
};

// Funkcja weryfikująca poprawność tokena JWT
const verifyToken = (token: string): User | null => {
  try {
    return jwt.verify(token, process.env.JWT_SECRET) as User;
  } catch (error) {
    return null;
  }
};
```

Listing 2.6. Generowanie i weryfikacja tokenów JWT w TypeScript

Źródło: Opracowanie własne

Kod przedstawiony w listingu 2.6 pokazuje, jak w prosty sposób można generować i weryfikować tokeny JWT w aplikacji Node.js. Funkcja generateToken przyjmuje obiekt użytkownika i zwraca wygenerowany token JWT, który zawiera dane użytkownika (takie jak jego ID i adres e-mail). Funkcja verifyToken weryfikuje poprawność tokena i zwraca dane użytkownika, jeśli token jest poprawny, lub null, jeśli token jest niepoprawny lub wygasł. Kolejnym krokiem jest integracja generowania i weryfikacji tokenów z logiką aplikacji. Przykładowo, podczas logowania użytkownika, aplikacja może wygenerować token JWT i zwrócić go do klienta, który będzie go przechowywał (np. w ciasteczkach lub lokalnej pamięci przeglądarki) i dołączał do każdego żądania wymagającego autoryzacji.

```
import express, { Request, Response } from 'express';
import bcrypt from 'bcrypt';
import jwt from 'jsonwebtoken';
const app = express();
app.use(express.json());
interface User {
  id: number;
  email: string;
  passwordHash: string;
}
// Przykładowa baza danych użytkowników
const users: User[] = [
  { id: 1, email: 'user@example.com', passwordHash:
bcrypt.hashSync('password', 10) }
];
// Endpoint logowania
app.post('/login', (req: Request, res: Response) => {
  const { email, password } = req.body;
  const user = users.find(u => u.email === email);

  if (!user || !bcrypt.compareSync(password, us-er.passwordHash)) {
    return res.status(401).send('Invalid email or pass-word');
  }
  const token = generateToken(user);
  res.json({ token });
});
// Middleware weryfikujący token JWT
```

```
const authenticateJWT = (req: Request, res: Response, next: Function) =>
{
  const authHeader = req.headers.authorization;
  if (authHeader) {
    const token = authHeader.split(' ')[1];
    const user = verifyToken(token);
    if (user) {
      req.user = user;
      next();
    } else {
      res.status(403).send('Forbidden');
    }
  } else {
    res.status(401).send('Unauthorized');
  }
};
// Przykładowy chroniony endpoint
app.get('/protected', authenticateJWT, (req: Request, res: Response) => {
  res.send('This is a protected route');
});
const PORT = process.env.PORT || 3000;
app.listen(PORT, () => {
  console.log(`Server is running on port ${PORT}`);
});
```

Listing 2.7. Uwierzytelnianie JWT z hashowaniem haseł w Express

Źródło: Opracowanie własne

W listingu 2.7, endpoint /login obsługuje logowanie użytkownika, weryfikując jego dane logowania i zwracając token JWT, jeśli dane są poprawne. Middleware authenticateJWT weryfikuje token JWT dołączony do żądania i, jeśli token jest poprawny, pozwala na dostęp do chronionych zasobów.

Podsumowując, autoryzacja i uwierzytelnianie to nieodzowne elementy zabezpieczania płatności online. Wykorzystanie protokołów OAuth2 i JWT pozwala na skuteczne zarządzanie sesjami użytkowników i ochronę danych przed nieautoryzowanym dostępem. Dzięki dostępności bibliotek w TypeScript, implementacja tych mechanizmów jest prosta i efektywna, co pozwala na tworzenie bezpiecznych aplikacji płatniczych.

2.4. Detekcja i zapobieganie oszustwom

Detekcja i zapobieganie oszustwom to kolejne istotne elementy zabezpieczania płatności. Systemy płatności powinny być wyposażone w mechanizmy monitorujące podejrzane aktywności, takie jak nagłe wzrosty transakcji z jednego konta czy płatności z różnych lokalizacji geograficznych. W tym celu można wykorzystać narzędzia analizy danych oraz machine learning, które będą w stanie wykrywać anomalie i zapobiegać oszustwom w czasie rzeczywistym. Implementacja takich mechanizmów może obejmować monitorowanie wzorców transakcji oraz użycie algorytmów predykcyjnych do identyfikacji potencjalnych zagrożeń. Przykładem może być system wykrywający oszustwa w czasie rzeczywistym, który analizuje dane transakcyjne pod kątem nietypowych wzorców. Gdy system wykryje podejrzane działania, może automatycznie zablokować transakcję i powiadomić odpowiedni zespół bezpieczeństwa w celu dalszej analizy. Takie podejście pozwala na szybkie reagowanie na zagrożenia i minimalizowanie potencjalnych strat.

3. Podsumowanie

Bezpieczeństwo płatności online to skomplikowany i wieloaspektowy proces, który wymaga zastosowania wielu technologii i najlepszych praktyk. TypeScript, dzięki swojej typizacji i rozbudowanej ekosystemie, stanowi doskonałe narzędzie do budowy bezpiecznych systemów płatności. Szyfrowanie danych, autoryzacja, uwierzytelnianie oraz detekcja oszustw to tylko niektóre z kluczowych elementów, które należy uwzględnić w procesie projektowania i implementacji backendu dla aplikacji finansowych. Stosowanie tych technologii i praktyk może znacząco zwiększyć bezpieczeństwo systemów płatności, a tym samym zbudować zaufanie klientów do marki. W dynamicznie rozwijającym się świecie technologii, ciągłe doskonalenie i aktualizowanie zabezpieczeń jest niezbędne, aby skutecznie chronić dane użytkowników i utrzymać przewagę konkurencyjną na rynku.

Źródła internetowe

1. https://www.honeybadger.io/blog/encryption-and-decryption-in-typescript/(dostęp: 13.06.2024).
2. https://clouddevs.com/typescript/secure-authentication/ (dostęp: 13.06.2024).
3. https://dev.to/vapourisation/east-encryption-in-typescript-3948 (dostęp: 13.06.2024).

Kamil Uchwat
SKNI KOD

Opiekun naukowy / dr inż. Bartosz Trybus

Klasyfikatory tekstu z użyciem głębokich sieci neuronowych

Streszczenie

Artykuł przedstawia przegląd nowoczesnych metod klasyfikacji tekstu z użyciem głębokich sieci neuronowych, koncentrując się na porównaniu tych technik z klasycznymi metodami oraz omówieniu ich zastosowań, wyzwań i przyszłości. Klasyfikacja tekstu, będąca kluczowym zadaniem przetwarzania języka naturalnego (NLP), znajduje szerokie zastosowania, takie jak filtrowanie spamu, analiza sentymentu i klasyfikacja dokumentów.

Tradycyjne metody, takie jak Naive Bayes i SVM, mimo że skuteczne, mają swoje ograniczenia, zwłaszcza w obliczu złożonych danych tekstowych. Głębokie sieci neuronowe, w szczególności rekurencyjne sieci neuronowe (RNN), sieci konwolucyjne (CNN) oraz transformery, oferują zaawansowane rozwiązania pozwalające na lepsze uchwycenie kontekstu i złożonych zależności w danych tekstowych.

Artykuł omawia proces przetwarzania danych, w tym tokenizację i tworzenie reprezentacji wektorowych (embeddings), a także techniki przygotowania danych wejściowych. Szczegółowo przedstawione są architektury modeli głębokiego uczenia, metody optymalizacji oraz techniki unikania nadmiernego dopasowania. Praktyczne przykłady zastosowań, takie jak analiza sentymentu i filtrowanie spamu, ilustrują możliwości tych technologii.

Słowa kluczowe: uczenie głębokie, sieci neuronowe, klasyfikacja tekstu.

1. Wprowadzenie

Klasyfikacja tekstu to jedno z kluczowych zadań w przetwarzaniu języka naturalnego (NLP), polegające na automatycznym przypisywaniu etykiet do fragmentów tekstu na podstawie ich zawartości. Ta technologia ma szerokie zastosowania w wielu dziedzinach, takich jak filtrowanie spamu, analiza sentymentu, klasyfikacja dokumentów, tłumaczenie maszynowe i systemy rekomendacji. Tradycyjne metody klasyfikacji tekstu, takie jak Naive Bayes, SVM (Support Vector Machines) i drzewa decyzyjne, były przez wiele lat podstawą tej dziedziny, jednak mają swoje ograniczenia, zwłaszcza w kontekście dużych i złożonych zbiorów danych tekstowych.

W ostatnich latach głębokie sieci neuronowe zrewolucjonizowały przetwarzanie języka naturalnego. Głębokie uczenie, będące poddziedziną uczenia maszynowego, wykorzystuje sieci neuronowe z wieloma warstwami do modelowania skomplikowanych wzorców w danych. Rekurencyjne sieci neuronowe (RNN), sieci konwolucyjne (CNN) oraz transformery, takie jak BERT i GPT, zdobyły popularność dzięki zdolności do przetwarzania i analizy sekwencji danych tekstowych w zaawansowany i efektywny sposób.

Celem tego artykułu jest przedstawienie osiągnięć w dziedzinie klasyfikacji tekstu z wykorzystaniem głębokich sieci neuronowych, porównanie tych metod z klasycznymi technikami oraz omówienie ich zastosowań.

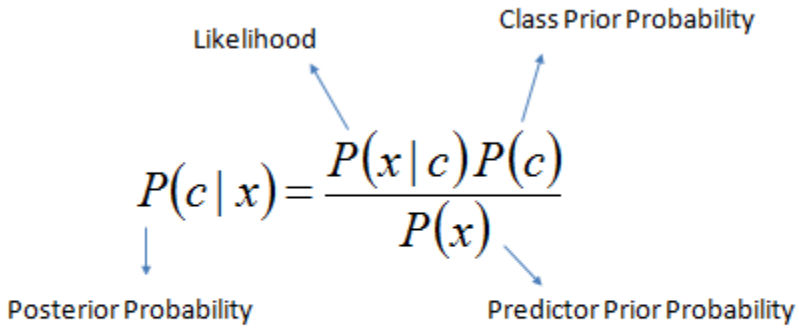
2. Klasyczne metody klasyfikacji tekstu

2.1.Naiwny klasyfikator Bayesa

Pierwszą klasyczną metodą klasyfikacji jaką omówię, będzie **naiwny klasyfikator Bayesa** , który jest prostym algorytmem probabilistycznym.

Naiwny Bayes należy do rodziny algorytmów uczenia się generatywnego, co oznacza, że stara się modelować rozkład danych wejściowych danej klasy lub kategorii. W przeciwieństwie do klasyfikatorów dyskryminacyjnych, takich jak regresja logistyczna, nie uczy się, które cechy są najważniejsze w celu rozróżnienia klas.

Ta metoda działa inaczej, ponieważ działa w oparciu o kilka kluczowych założeń, dzięki czemu zyskuje miano „naiwnej". Zakłada, że predyktory w modelu Naiwnego Bayesa są warunkowo niezależne lub niepowiązane z żadną inną cechą modelu. Zakłada również, że wszystkie cechy w równym stopniu przyczyniają się do wyniku. Chociaż założenia te są często łamane w rzeczywistych scenariuszach (np. kolejne słowo w wiadomości e-mail zależy od słowa, które je poprzedza), upraszcza to problem klasyfikacji, czyniąc go bardziej wykonalnym obliczeniowo. Oznacza to, że dla każdej zmiennej będzie teraz wymagane tylko jedno prawdopodobieństwo, co z kolei ułatwi obliczenia modelu. Pomimo tego nierealistycznego założenia o niezależności algorytm klasyfikacji działa dobrze, szczególnie w przypadku małych próbek.



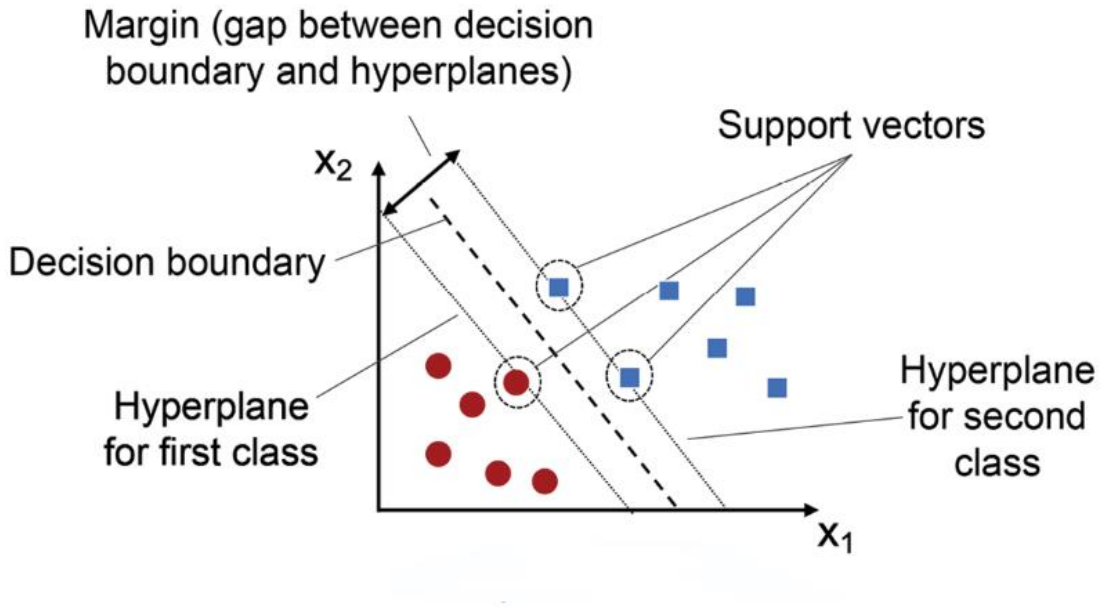
$$P(c \mid \mathrm{X}) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

Rys. 1 Źródło : https://mateuszgrzyb.pl/wybor-odpowiedniego-algorytmu-czesc-2-algorytmy-klasyfikacyjne/

2.2. Metoda wektorów nośnych

Metoda Wektorów Nośnych (Support Vector Machines, SVM) jest jednym z najbardziej efektywnych i popularnych algorytmów uczenia maszynowego używanych do klasyfikacji i regresji. Jest szczególnie ceniona za swoją zdolność do radzenia sobie z danymi wielowymiarowymi i nieciągłymi, co czyni ją wszechstronnym narzędziem w rozwiązywaniu problemów klasyfikacyjnych.

SVM działa na zasadzie znalezienia optymalnej granicy decyzyjnej (zwanej hiperplanem) oddzielającej klasy w przestrzeni cech. W klasyfikacji binarnej celem SVM jest znalezienie hiperplanu, który maksymalizuje margines, czyli odległość między najbliższymi punktami danych obu klas (zwanymi wektorami nośnymi) a tym hiperplanem. Im większy margines, tym lepsze będzie ogólne działanie klasyfikatora na nowych, niewidzianych danych.



Rys. 2 Źródło: https://vitalflux.com/classification-model-svm-classifier-python-example/

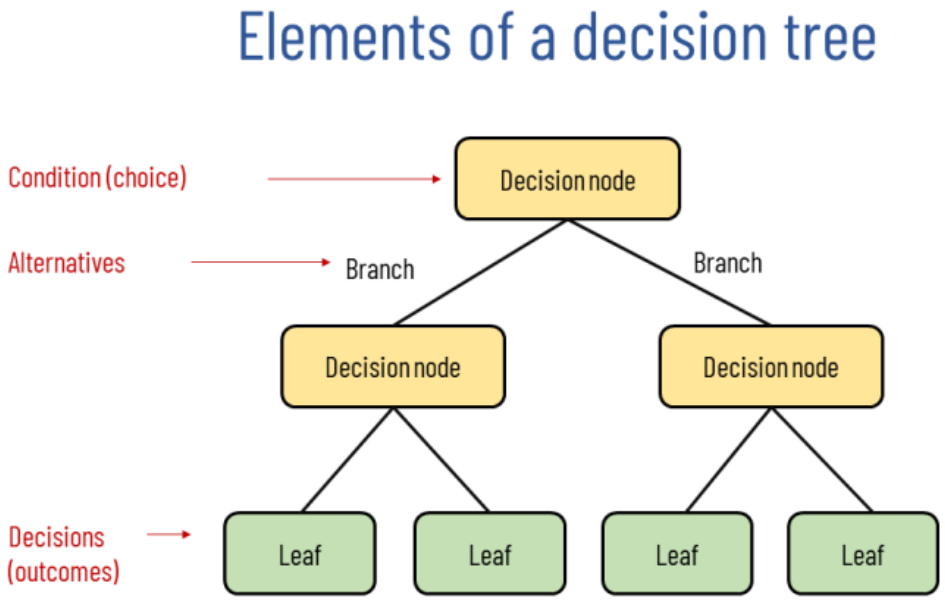
2.3. Drzewa decyzyjne

Drzewa decyzyjne są jedną z najbardziej intuicyjnych i przystępnych metod uczenia maszynowego, stosowaną zarówno w klasyfikacji, jak i regresji. Ta metoda opiera się na hierarchicznym podziale danych na podstawie cech, co pozwala na podejmowanie decyzji krok po kroku, przypominając proces myślowy człowieka.

Drzewo decyzyjne składa się z węzłów (nodes) i gałęzi (branches). Struktura ta zaczyna się od pojedynczego węzła początkowego, zwanego korzeniem (root), który następnie dzieli się na dwa lub więcej węzłów podrzędnych. Każdy węzeł wewnętrzny reprezentuje test na jednej z cech (atrybutów), a każda gałąź reprezentuje wynik tego testu. Liście (leaves) są końcowymi węzłami, które przedstawiają decyzję lub przewidywaną wartość.

Proces budowy drzewa:

1. **Wybór cechy podziału:**
    a. Na początku wybierana jest cecha, która najlepiej dzieli dane. Popularnymi miarami oceny jakości podziału są:
        i. Gini impurity: Stosowana w klasyfikacji, mierzy czystość węzła.
        ii. Entropy: Miara niepewności w zbiorze danych.
        iii. Variance reduction: Używana w regresji, mierzy zmniejszenie wariancji.
2. **Podział danych:**
    a. Na podstawie wybranej cechy, dane są podzielone na podzbiory. Proces ten jest rekurencyjnie powtarzany dla każdego z nowych podzbiorów.
3. **Kryterium zakończenia:**
    a. Budowa drzewa kończy się, gdy spełnione są określone kryteria, takie jak osiągnięcie maksymalnej głębokości drzewa, minimalna liczba próbek w węźle lub brak dalszego podziału, który poprawiłby jakość decyzji.



Rys. 3 Źródło: https://why-change.com/2021/11/13/how-to-create-decision-trees-for-business-rules-analysis/

3. Przygotowanie danych

3.1.Technika TF-IDF

TF-IDF, czyli **Term Frequency-Inverse Document Frequency**, jest popularną metodą używaną do przetwarzania tekstu i analizy tekstu, szczególnie w zadaniach związanych z wyszukiwaniem informacji i klasyfikacją tekstu. Metoda ta pozwala na ocenę znaczenia słów w dokumencie w kontekście całego zbioru dokumentów. TF-IDF łączy dwa wskaźniki: częstotliwość terminu (TF) i odwrotną częstość dokumentów (IDF).

Częstotliwość terminu (TF) mierzy, jak często dany termin (słowo) pojawia się w dokumencie. Jest to stosunkowo prosty wskaźnik, który można obliczyć na kilka sposobów, ale najczęściej jest to stosunek liczby wystąpień słowa w dokumencie do całkowitej liczby słów w dokumencie. Można to zapisać wzorem:

$$TF(t,d) = \frac{Liczba\ wystąpień\ słowa\ `t`\ w\ dokumencie\ 'd'}{Całkowita\ liczba\ słów\ w\ dokumencie\ 'd'}$$

Odwrotna częstość dokumentów (IDF) mierzy, jak unikalne jest dane słowo w całym zbiorze dokumentów. Częstość dokumentów (DF) oznacza, w ilu dokumentach dane słowo się pojawia. Jeśli słowo występuje w wielu dokumentach, jego IDF będzie niskie, ponieważ nie jest ono specyficzne ani unikalne dla żadnego konkretnego dokumentu. Wzór na IDF wygląda następująco:

$$IDF(t) = \log(\frac{N}{DF(t)})$$

Gdzie:

- N – całkowita liczba dokumentów w zbiorze.
- DF(t) – liczba dokumentów zawierających słowo t.

TF-IDF łączy oba wskaźniki, aby ocenić znaczenie słowa w danym dokumencie w kontekście całego zbioru dokumentów. Wzór na TF-IDF to:

$$TF - IDF(t, d) = TF(t, d) * IDF(t)$$

3.2.Metoda bag of words

Metoda Bag of Words (BoW) to prosty i popularny sposób reprezentacji tekstu, używany w przetwarzaniu języka naturalnego (NLP) i uczeniu maszynowym. Główną ideą tej metody jest przekształcenie dokumentu tekstowego w zbiór słów (tzw. "torba słów"), nie uwzględniając ich kolejności w zdaniu, ale zachowując ich ilość. Dzięki temu tekst można łatwo analizować i przetwarzać za pomocą algorytmów.

Jak działa metoda Bag of Words?

1. **Zbiór dokumentów**:
    - Na początku zbieramy zbiór dokumentów, które chcemy analizować. Może to być kilka artykułów, wiadomości e-mail, recenzji produktów itp.
2. **Tokenizacja**:
    - Każdy dokument jest dzielony na pojedyncze słowa lub tokeny. Tokenizacja to proces rozdzielania tekstu na poszczególne słowa.
    - Przykład: Dokument "Kot lubi mleko" zostanie rozdzielony na słowa ["Kot", "lubi", "mleko"].
3. **Tworzenie słownika**:
    - Następnie tworzymy słownik wszystkich unikalnych słów, które występują w zbiorze dokumentów. Słownik to po prostu lista wszystkich słów, które występują w naszych dokumentach.
    - Przykład: Jeśli mamy trzy dokumenty:
        1. "Kot lubi mleko"
        2. "Pies lubi kości"
        3. "Kot i pies lubią zabawę" Słownik będzie zawierał słowa: ["Kot", "lubi", "mleko", "Pies", "kości", "i", "lubią", "zabawę"].
4. **Tworzenie wektorów**:
    - Każdy dokument jest reprezentowany jako wektor o długości równej liczbie unikalnych słów w słowniku. Wartości w wektorze odpowiadają liczbie wystąpień poszczególnych słów w dokumencie.
    - Przykład: Dokument "Kot lubi mleko" będzie reprezentowany jako wektor [1, 1, 1, 0, 0, 0, 0, 0], gdzie:

- "Kot" występuje 1 raz,
- "lubi" występuje 1 raz,
- "mleko" występuje 1 raz,
- a reszta słów ze słownika nie występuje w tym dokumencie.

4. Wprowadzenie do głębokiego uczenia

Uczenie głębokie (deep learning) to poddziedzina uczenia maszynowego, która koncentruje się na modelach zwanych sieciami neuronowymi. Uczenie głębokie jest inspirowane strukturą i funkcjonowaniem ludzkiego mózgu, a jego celem jest nauczanie komputerów wykonywania złożonych zadań, takich jak rozpoznawanie obrazów, przetwarzanie języka naturalnego czy granie w gry, bez konieczności programowania każdej reguły z osobna.

4.1.Podstawowe pojęcia

- **Sztuczna Sieć Neuronowa (Artificial Neural Network, ANN)**:
  - To system składający się z połączonych ze sobą jednostek, zwanych neuronami. Każdy neuron otrzymuje wejście, przetwarza je i przesyła wynik do innych neuronów. Sieć neuronowa składa się zazwyczaj z trzech typów warstw: warstwy wejściowej, warstw ukrytych i warstwy wyjściowej.
- **Neuron**:
  - Podstawowy element sieci neuronowej, który naśladuje działanie biologicznego neuronu. Otrzymuje sygnały wejściowe, przetwarza je za pomocą funkcji aktywacji i wysyła wynik do następnych neuronów.
- **Warstwa**:
  - Zbiór neuronów na tym samym poziomie. Sieci neuronowe składają się z warstw:
    - **Warstwa wejściowa**: Otrzymuje surowe dane wejściowe.
    - **Warstwy ukryte**: Przetwarzają dane na różnych poziomach abstrakcji.
    - **Warstwa wyjściowa**: Produkuje ostateczny wynik modelu.
- **Funkcja aktywacji**:
  - Funkcja matematyczna stosowana do wyjścia neuronu. Pomaga modelowi uczyć się nieliniowych wzorców. Przykłady to ReLU (Rectified Linear Unit), sigmoid i tanh.
- **Propagacja wsteczna (Backpropagation)**:
  - Algorytm służący do trenowania sieci neuronowych. Polega na propagacji błędu wstecz przez sieć, aby dostosować wagi neuronów. Jest to kluczowy element procesu uczenia się w sieciach neuronowych.

- **Wagi (Weights)**:
    - Parametry neuronów, które są modyfikowane podczas procesu uczenia się. Wagi kontrolują, jak mocno sygnały wejściowe wpływają na wynik neuronu.
- **Epoka (Epoch)**:
    - Jedno pełne przejście przez cały zbiór treningowy. W uczeniu głębokim, sieci neuronowe są trenowane przez wiele epok, aby zoptymalizować swoje wagi i zminimalizować błąd.
- **Optymalizator**:
    - Algorytm, który pomaga w procesie uczenia się, dostosowując wagi neuronów w celu minimalizacji funkcji kosztu. Przykłady to gradientowy spadek stochastyczny (SGD), Adam i RMSprop.
- **Zbiór treningowy, walidacyjny i testowy**:
    - Dane są zazwyczaj dzielone na trzy części:
        - **Zbiór treningowy**: Używany do trenowania modelu.
        - **Zbiór walidacyjny**: Używany do tuningu hyperparametrów i zapobiegania przeuczeniu.
        - **Zbiór testowy**: Używany do oceny ostatecznej wydajności modelu.

4.2.Dlaczego uczenie głębokie jest tak potężne?

Uczenie głębokie zyskało ogromną popularność dzięki swojej zdolności do automatycznego wykrywania skomplikowanych wzorców w danych. Jest wyjątkowo skuteczne w analizie obrazów, dźwięku i tekstu, co umożliwiło postęp w dziedzinach takich jak rozpoznawanie mowy, tłumaczenie maszynowe, diagnozowanie medyczne i autonomiczne pojazdy.

Uczenie głębokie korzysta z dużych zbiorów danych oraz potężnych jednostek obliczeniowych, takich jak procesory graficzne (GPU), co pozwala na skuteczne trenowanie bardzo głębokich sieci neuronowych z wieloma warstwami ukrytymi.

5. Głębokie sieci neuronowe w klasyfikacji tekstu

5.1.Transformery

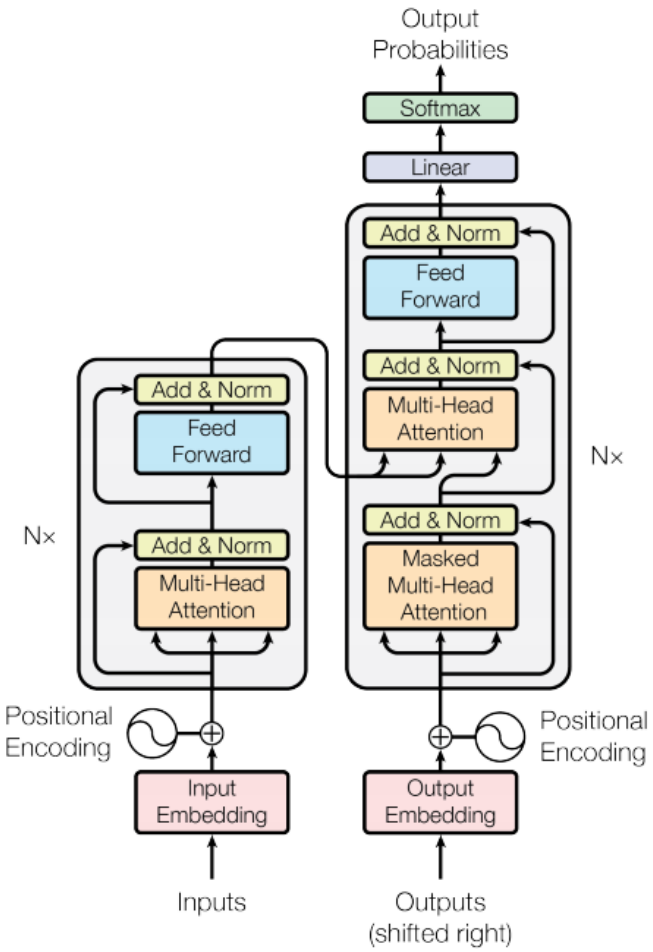
Transformery to nowoczesna architektura sieci neuronowych, która zrewolucjonizowała przetwarzanie języka naturalnego (NLP). Wprowadzone w artykule "Attention is All You Need" w 2017 roku przez zespół badaczy Google, transformery charakteryzują się zdolnością do skutecznego przetwarzania sekwencji danych bez potrzeby korzystania z rekurencyjnych struktur.

Jak działają transformery?

1. **Warstwy samoatencji (Self-Attention)**:

- Główna innowacja w transformerach polega na mechanizmie samoatencji, który pozwala modelowi zwracać uwagę na różne części sekwencji wejściowej jednocześnie. Dzięki temu, każde słowo w zdaniu może wpływać na tłumaczenie lub analizę innych słów, niezależnie od ich odległości.
- Mechanizm samoatencji oblicza ważność każdego słowa względem innych słów w sekwencji, co pozwala na lepsze zrozumienie kontekstu.

2. **Koder i dekoder**:
   - Transformery składają się z dwóch głównych części: kodera i dekodera.
   - **Koder** przetwarza dane wejściowe i generuje reprezentacje wewnętrzne.
   - **Dekoder** przekształca te reprezentacje w dane wyjściowe, takie jak tłumaczenia lub odpowiedzi na pytania.
3. **Warstwy liniowe i nieliniowe**:
   - Transformery wykorzystują kombinację warstw liniowych (dense layers) i nieliniowych (activation functions) do przetwarzania danych na różnych poziomach abstrakcji.



Rys. 4 Źródło: https://miroslawmamczur.pl/czym-jest-i-jak-dziala-transformer-siec-neuronowa/

Transformery są szeroko stosowane w zadaniach NLP, takich jak tłumaczenie maszynowe, generowanie tekstu, analiza sentymentu, oraz systemy pytanie-odpowiedź. Modele takie jak BERT, GPT-3 i T5 opierają się na architekturze transformerów i osiągają znakomite wyniki w wielu zadaniach językowych.

5.2.Rekurencyjne sieci neuronowe

Rekurencyjne sieci neuronowe (RNN) są specjalnym typem sieci neuronowych zaprojektowanym do przetwarzania sekwencji danych. RNN są szczególnie przydatne w zadaniach, gdzie kolejność danych ma znaczenie, takich jak analiza tekstu, rozpoznawanie mowy czy prognozowanie czasowe.

Jak działają RNN?

1. **Pamięć stanu (Hidden State)**:
    - RNN mają zdolność przechowywania informacji o poprzednich elementach sekwencji dzięki ukrytemu stanowi (hidden state), który jest aktualizowany na każdym kroku sekwencji.
    - Na wejściu RNN przetwarza jeden element sekwencji i aktualizuje swój ukryty stan, który jest następnie używany do przetwarzania kolejnego elementu.
2. **Kolejność danych**:
    - RNN są zaprojektowane do uwzględniania kolejności danych. Na każdym kroku, ukryty stan przechowuje informacje z poprzednich kroków, co pozwala modelowi na rozumienie kontekstu w sekwencji.
3. **Problemy z uczeniem**:
    - Klasyczne RNN mają trudności z zapamiętywaniem długoterminowych zależności w sekwencjach, co prowadzi do problemów z zanikiem gradientu (vanishing gradient problem).

Ulepszenia RNN: LSTM i GRU

Aby rozwiązać problemy klasycznych RNN, wprowadzono ulepszone architektury:

- **LSTM (Long Short-Term Memory)**: Zawiera specjalne mechanizmy, takie jak komórki pamięci i bramki, które pozwalają na lepsze przechowywanie i manipulowanie długoterminowymi zależnościami.
- **GRU (Gated Recurrent Unit)**: Uproszczona wersja LSTM, która także skutecznie radzi sobie z problemem zanikania gradientu, ale jest mniej złożona.

5.3.Konwolucyjne sieci neuronowe

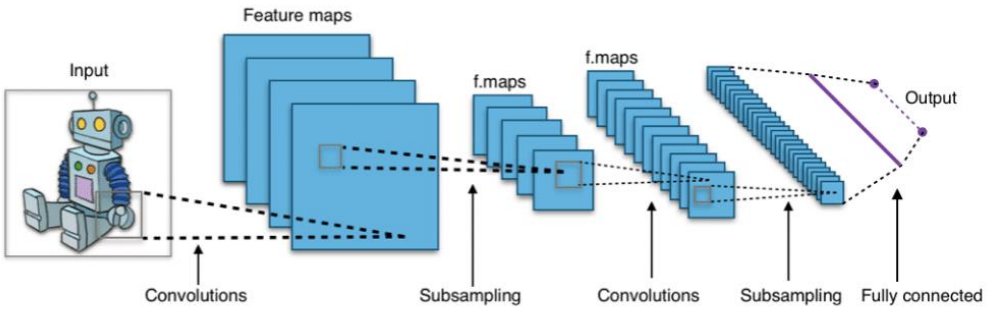
Sieci konwolucyjne (Convolutional Neural Networks, CNN) są specjalnym typem sieci neuronowych, które są szczególnie skuteczne w przetwarzaniu danych o strukturze siatki, takich jak obrazy. CNN są powszechnie stosowane w zadaniach związanych z wizją komputerową, takich jak rozpoznawanie obrazów, klasyfikacja obiektów i segmentacja obrazu.

Jak działają CNN?

1. **Warstwy konwolucyjne**:

- Podstawowym elementem CNN jest warstwa konwolucyjna, która przetwarza dane wejściowe za pomocą filtrów (kernels). Filtry przesuwają się po obrazie (lub innej siatce danych) i wykrywają różne cechy, takie jak krawędzie, tekstury czy wzory.
- Każdy filtr generuje mapę cech (feature map), która reprezentuje wykryte cechy w obrazie.

2. **Warstwy spłaszczające i łączeniowe (Pooling Layers)**:
   - Warstwy łączeniowe, takie jak max-pooling, zmniejszają wymiarowość map cech, agregując informacje i redukując ilość parametrów. To pomaga w zredukowaniu kosztów obliczeniowych i zapobiega przeuczeniu (overfitting).
3. **Warstwy gęste (Fully Connected Layers)**:
   - Po warstwach konwolucyjnych i łączeniowych, dane są przekształcane przez warstwy gęste, które łączą wszystkie neurony. Warstwy te służą do ostatecznej klasyfikacji lub regresji na podstawie wyuczonych cech.



Rys. 5Źródło: https://www.google.com/url?sa=i&url=https%3A%2F%2Fen.wikipedia.org%2Fwiki%2FConvolutional_neural_network&psig=AOvVaw2mlF_PJ0spceOaNAy88ZZM&ust=1726049161588000&source=images&cd=vfe&opi=89978449&ved=0CBcQjhxqFwoTCJDqoqyQuIgDFQAAAAAdAAAAABAE

Podsumowując, transformery, RNN i CNN to trzy różne typy sieci neuronowych, z których każdy jest zaprojektowany do specyficznych typów zadań i danych. Transformery są wyjątkowo skuteczne w przetwarzaniu sekwencji tekstowych, RNN są stosowane tam, gdzie istotna jest kolejność danych, a CNN doskonale radzą sobie z analizą obrazów i innych danych o strukturze siatki.

6. Trenowanie i ewolucja modelu

6.1.Zbiory danych

W klasyfikacji tekstu używa się wielu różnych zbiorów danych, które są specjalnie zaprojektowane do trenowania i testowania modeli uczenia maszynowego. Oto kilka popularnych i często wykorzystywanych zbiorów danych:

1. IMDB

- **Opis**: Zbiór danych z recenzjami filmów z bazy IMDb (Internet Movie Database).
- **Zawartość**: Zawiera 50 000 recenzji filmów, z których połowa to recenzje pozytywne, a połowa negatywne.
- **Zastosowanie**: Klasyfikacja sentymentu, czyli określenie, czy recenzja jest pozytywna czy negatywna.
- **Przykład**: Jeśli model zostanie przeszkolony na tym zbiorze danych, będzie mógł przewidzieć, czy nowa recenzja filmu jest pozytywna czy negatywna.

2. 20 Newsgroups

- **Opis**: Zbiór danych z artykułami z 20 różnych grup dyskusyjnych Usenet.
- **Zawartość**: Zawiera około 20 000 dokumentów, które są przypisane do jednej z 20 kategorii, takich jak sport, polityka, nauka itp.
- **Zastosowanie**: Klasyfikacja tekstu według tematu, czyli określenie, do której kategorii należy dany artykuł.
- **Przykład**: Jeśli model zostanie przeszkolony na tym zbiorze danych, będzie mógł przewidzieć, czy nowy artykuł dotyczy np. sportu czy polityki.

3. Reuters-21578

- **Opis**: Zbiór danych z artykułami z agencji informacyjnej Reuters.
- **Zawartość**: Zawiera 21 578 artykułów, podzielonych na różne kategorie tematyczne, takie jak finanse, handel, rynki, itp.
- **Zastosowanie**: Klasyfikacja tekstu według tematu.
- **Przykład**: Model przeszkolony na tym zbiorze danych może być używany do automatycznego kategoryzowania nowych wiadomości.

4. AG News

- **Opis**: Zbiór danych z artykułami wiadomości z czterech różnych kategorii: świat, sport, biznes i technologia.
- **Zawartość**: Zawiera 120 000 artykułów podzielonych równomiernie między te cztery kategorie.
- **Zastosowanie**: Klasyfikacja wiadomości według kategorii.

- **Przykład**: Model przeszkolony na tym zbiorze danych będzie w stanie przewidzieć, czy nowy artykuł dotyczy np. technologii czy sportu.

5. Yelp Reviews

- **Opis**: Zbiór danych z recenzjami różnych biznesów (restauracji, sklepów itp.) z platformy Yelp.
- **Zawartość**: Zawiera miliony recenzji ocenionych w skali od 1 do 5 gwiazdek.
- **Zastosowanie**: Klasyfikacja sentymentu oraz analiza opinii klientów.
- **Przykład**: Model przeszkolony na tym zbiorze danych może określić, czy recenzja biznesu jest pozytywna czy negatywna oraz przewidzieć ocenę w gwiazdkach.

6. SpamAssassin

- **Opis**: Zbiór danych z wiadomościami e-mail sklasyfikowanymi jako spam lub nie-spam.
- **Zawartość**: Zawiera tysiące wiadomości e-mail, z których każda jest oznaczona jako spam lub nie-spam.
- **Zastosowanie**: Filtracja spamu, czyli automatyczne wykrywanie i segregowanie wiadomości spamowych.
- **Przykład**: Model przeszkolony na tym zbiorze danych może być używany do automatycznego filtrowania spamu w skrzynce e-mail.

Każdy z tych zbiorów danych ma swoje specyficzne zastosowania i jest wykorzystywany do trenowania modeli, które mogą wykonywać różnorodne zadania związane z klasyfikacją tekstu. W zależności od rodzaju zadania – czy to klasyfikacja sentymentu, klasyfikacja tematyczna czy filtrowanie spamu – różne zbiory danych mogą być bardziej odpowiednie. Dzięki tym zbiorom danych, badacze i inżynierowie mogą tworzyć i testować swoje modele, aby były one jak najbardziej skuteczne w realnych zastosowaniach.

6.2.Metryki oceny

Metryki oceny są kluczowe do mierzenia wydajności modeli klasyfikacyjnych. Oto cztery podstawowe metryki:

1. Dokładność (Accuracy)

- **Opis**: Dokładność to stosunek liczby poprawnych przewidywań do ogólnej liczby przewidywań.
- **Formuła**: $Dokładność = \frac{Poprawne\ trafienia}{Liczba\ danych\ w\ zbiorze}$
- **Przykład**: Jeśli model poprawnie przewidział 90 z 100 przykładów, dokładność wynosi 90%.

2. Precyzja (Precision)

- **Opis**: Precyzja to stosunek liczby prawdziwych pozytywnych przewidywań do liczby wszystkich pozytywnych przewidywań.
- **Formuła**: $Precyzja = \frac{Prawdziwe\ pozytywne}{Prawdziwe\ pozytywne + prawdziwe\ negatywne}$

- **Przykład**: Jeśli model przewidział 50 pozytywnych przykładów, z czego 45 było poprawnych, precyzja wynosi 90%

3. Czułość (Recall, inaczej Sensitivity lub TPR)

- **Opis**: Czułość to stosunek liczby prawdziwych pozytywnych przewidywań do liczby wszystkich rzeczywistych pozytywnych przypadków.
- **Formuła**: $Czułość = \frac{Prawdziwe\ pozytywne}{Prawdziwe\ pozytywne + fałszywe\ negatywne}$
- **Przykład**: Jeśli model wykrył 45 z 50 rzeczywistych pozytywnych przykładów, czułość wynosi 90%

4. F1-score

- **Opis**: F1-score to średnia harmoniczna precyzji i czułości. Jest używana, gdy potrzebna jest równowaga między precyzją a czułością.
- **Formuła**: $F1 - score = 2 * \frac{precyzja * czułość}{Precyzja + czułość}$
- **Przykład**: Jeśli precyzja i czułość wynoszą 0.9, F1-score wynosi 2×0.9×0.9/0.9+0.9=0.9

6.3.Techniki unikania nadmiernego dopasowania tzw. Overfittingu

Nadmierne dopasowanie (overfitting) występuje, gdy model jest zbyt dobrze dopasowany do danych treningowych i nie radzi sobie dobrze z nowymi danymi. Oto trzy popularne techniki unikania nadmiernego dopasowania:

1. Regularizacja

- **Opis**: Regularizacja polega na dodaniu dodatkowego ograniczenia do funkcji kosztu, które penalizuje duże wartości wag.
- **Rodzaje regularizacji**:
    - **L1 (Lasso)**: Dodaje sumę wartości bezwzględnych wag do funkcji kosztu.
    - **L2 (Ridge)**: Dodaje sumę kwadratów wag do funkcji kosztu.
- **Efekt**: Regularizacja zmniejsza złożoność modelu, co pomaga uniknąć nadmiernego dopasowania.

2. Dropout

- **Opis**: Dropout to technika, w której losowo "wyłączane" są niektóre neurony podczas treningu.
- **Jak działa**: Na każdej iteracji treningu, losowo wybierany procent neuronów jest ignorowany (ustawiany na zero).
- **Efekt**: Dropout zmniejsza ryzyko, że model będzie polegał na zbyt specyficznych wzorcach danych treningowych, co poprawia jego zdolność generalizacji.

3. Early Stopping

- **Opis**: Early stopping to technika, w której trening jest zatrzymywany, gdy model przestaje poprawiać się na zbiorze walidacyjnym.
- **Jak działa**: Podczas treningu, co kilka epok sprawdza się wydajność modelu na zbiorze walidacyjnym. Jeśli wydajność przestaje się poprawiać przez określoną liczbę epok (patience), trening zostaje przerwany.
- **Efekt**: Early stopping zapobiega nadmiernemu dopasowaniu modelu do danych treningowych, umożliwiając lepszą generalizację do nowych danych.

Te metryki oceny i techniki unikania nadmiernego dopasowania są kluczowe dla skutecznego trenowania i oceny modeli klasyfikacyjnych. Dokładność, precyzja, czułość i F1-score pozwalają mierzyć, jak dobrze model radzi sobie z różnymi aspektami przewidywań. Regularizacja, dropout i early stopping pomagają zapewnić, że model jest dobrze dostosowany do danych treningowych, ale jednocześnie ma zdolność do radzenia sobie z nowymi, niewidzianymi wcześniej danymi.

7. Podsumowanie

Artykuł omawia nowoczesne metody klasyfikacji tekstu z wykorzystaniem głębokich sieci neuronowych. Skupia się na porównaniu tych technik z tradycyjnymi metodami, takimi jak Naive Bayes i SVM, oraz na omówieniu ich. W artykule przedstawiono proces przetwarzania danych, w tym tokenizację i tworzenie reprezentacji wektorowych (embeddings), oraz techniki przygotowania danych wejściowych. Omówiono również architektury modeli głębokiego uczenia, metody optymalizacji oraz techniki unikania nadmiernego dopasowania. Praktyczne przykłady zastosowań, takie jak analiza sentymentu i filtrowanie spamu, ilustrują możliwości tych technologii.

Źródła internetowe

https://www.unite.ai/pl/jak-dzia%C5%82a-klasyfikacja-tekstu/

https://www.ibm.com/topics/naive-bayes

https://www.statsoft.pl/textbook/stathome_stat.html?https%3A%2F%2Fwww.statsoft.pl%2Ftextbook%2Fstsvm.html

https://home.agh.edu.pl/~pmarynow/pliki/iwmet/drzewa

https://pl.wikipedia.org/wiki/Naiwny_klasyfikator_bayesowski

https://mateuszgrzyb.pl/wybor-odpowiedniego-algorytmu-czesc-2-algorytmy-klasyfikacyjne/

https://vitalflux.com/classification-model-svm-classifier-python-example/

https://www.google.com/url?sa=i&url=https%3A%2F%2Fwhy-change.com%2F2021%2F11%2F13%2Fhow-to-create-decision-trees-for-business-rules-analysis%2F&psig=AOvVaw3yS1DiPZq-2YFqP-JYLWZQ&ust=1719823528202000&source=images&cd=vfe&opi=89978449&ved=0CBQQjhxqFwoTCKjEiIf4gocDFQAAAAAdAAAAABAJ (rys.3)

https://www.google.com/url?sa=i&url=https%3A%2F%2Fmiroslawmamczur.pl%2Fczym-jest-i-jak-dziala-transformer-siec-neuronowa%2F&psig=AOvVaw00rLVZJZA5Evy_2bulTzk_&ust=1719826735377000&source=images&cd=vfe&opi=89978449&ved=0CBQQjhxqFwoTCNCe2P6Dg4cDFQAAAAAdAAAAABAE

https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.unite.ai%2Fpl%2Fwhat-are-convolutional-neural-networks%2F&psig=AOvVaw3ksHLUzEKQ0iGRQzOMRxji&ust=1719826879688000&source=images&cd=vfe&opi=89978449&ved=0CBQQjhxqFwoTCKij2vKEg4cDFQAAAAAdAAAAABAE