AKADEMIA BIALSKA
IM. JANA PAWŁA II
Jubileusz 25-lecia

SZTUCZNA INTELIGENCJA W KSZTAŁTOWANIU BEZPIECZEŃSTWA INTERNETU

Redakcja naukowa
Tomasz Grudniewski
Marta Chodyka
Zbigniew Omiotek

Biała Podlaska 2025

AKADEMIA BIALSKA
IM. JANA PAWŁA II
Jubileusz 25-lecia

SZTUCZNA INTELIGENCJA W KSZTAŁTOWANIU BEZPIECZEŃSTWA INTERNETU

Redakcja naukowa
Tomasz Grudniewski
Marta Chodyka
Zbigniew Omiotek

Biała Podlaska 2025

Wydawca
Akademia Bialska im. Jana Pawła II

Recenzja
prof. dr hab. Zbigniew Ciekanowski
dr Sławomir Żurakowski

© Copyright by Akademia Bialska im. Jana Pawła II

Książka jest udostępniana na licencji Creative Commons Uznanie autorstwa – Użycie niekomercyjne – Bez utworów zależnych 4.0 Międzynarodowa

ISBN 978-83-68103-30-4
https://doi.org/10.29316/ab.9788368103304

Wydawnictwo AB JPII
ul. Sidorska 95/97
21-500 Biała Podlaska
wydawnictwo.akademiabialska.pl

Projekt okładki, skład

druk-24h.com.pl
DRUKARNIA CYFROWA
Hurtowa 21, 15-399 Białystok
tel. 535 500 898
e-mail: studio@druk-24h.com.pl

SPIS TREŚCI

Wprowadzenie 5

Aleksander Anchimiuk
Insecure Output Handling in Large Language Models. 7

Sławomir Krzysztof Czubaj
Rola sztucznej inteligencji w kształtowaniu bezpieczeństwa informacyjnego Internetu jako medium współczesnej cywilizacji. 25

Marek Kania
Rola sztucznej inteligencji w wykrywaniu i neutralizacji zagrożeń cybernetycznych 41

Magda Konieczna
Systemy komputerowego wspomagania diagnostyki oparte na głębokim uczeniu i technikach wyjaśniania decyzji: aktualne trendy i kierunki rozwoju 52

Michał Świątczak, Tomasz Grudniewski
Automatyczna klasyfikacja obrazów mikroskopowych okrzemek z wykorzystaniem bibliotek Pythona i transfer learningu: badanie porównawcze EfficientNetB0, ResNet50 i VGG16 69

Krzysztof Łuczka
Zmiany i innowacje w ocenie zdolności kredytowej. 90

Ewelina Melaniuk
Wykorzystanie technologii AI w kreacji wirtualnych influencerów 108

Volodymyr Malovichko, Roman Rybalka, Kostiantyn Honcharov
Proposal for Enhancing Sustainability in the Digitalisation of Education in the USUST 123

Barbara Drapikowska, Robert Zakrzewski
Zaufanie i lęk wobec sztucznej inteligencji – psychologiczne mechanizmy a odpowiedzialne wdrażanie technologii 145

Zofia Lubańska, Gabriel Tarasiuk
Automatyczne rozpoznawanie mowy z ruchu ust przy użyciu sztucznej inteligencji – przegląd metod, zastosowań i wyzwań 164

Piotr Lichograj, Marta Chodyka, Michał Szałkowski, Vladimir Golovko
Analiza metod YOLO w detekcji obiektów z obrazów rejestrowanych przez UAV oraz ich zastosowanie w pomiarze prędkości pieszych 189

WPROWADZENIE

Dynamiczny rozwój sztucznej inteligencji (AI) sprawia, że technologie oparte na algorytmach uczenia maszynowego i głębokiego uczenia stają się integralnym elementem współczesnej rzeczywistości cyfrowej. Zastosowania AI obejmują zarówno obszary o kluczowym znaczeniu dla bezpieczeństwa informacyjnego, jak i sektory wymagające wysokiej precyzji analitycznej, kreatywności technologicznej czy odporności na zmieniające się warunki środowiska cyfrowego. Współczesne systemy oparte na AI nie tylko wspierają procesy decyzyjne i upraszczają analizę danych, lecz coraz częściej przejmują odpowiedzialność za zadania wymagające szybkości reakcji, skalowalności oraz interpretacji złożonych wzorców.

W przestrzeni bezpieczeństwa cyfrowego sztuczna inteligencja pełni funkcję kluczowego narzędzia ochrony, umożliwiając wykrywanie anomalii, neutralizację zaawansowanych zagrożeń oraz zwiększanie odporności systemów informacyjnych na ataki. W obszarze medycyny i diagnostyki AI wspiera analizę obrazów, wzmacnia interpretowalność modeli oraz ułatwia automatyzację procesów, które dotychczas wymagały specjalistycznej wiedzy i dużego nakładu pracy. Równie istotny jest sektor finansowy, gdzie algorytmy predykcyjne umożliwiają nowatorskie podejście do oceny ryzyka i tworzenia dynamicznych modeli zdolności kredytowej. Z kolei w sferze komunikacji cyfrowej sztuczna inteligencja otwiera zupełnie nowe możliwości, pozwalając na kreowanie wirtualnych postaci, automatyzację interakcji oraz tworzenie immersyjnych doświadczeń, które redefiniują relacje między użytkownikiem a systemem cyfrowym. Wspólnym mianownikiem przedstawionych w monografii zagadnień jest szeroko pojęta transformacja cyfrowa, której fundamentem stały się technologie AI. Rozdziały zebrane w tomie ukazują różnorodność zastosowań sztucznej inteligencji w praktyce – od bezpieczeństwa i diagnostyki, przez edukację i infrastrukturę cyfrową, aż po sektor kreatywny i finansowy. Prace te podkreślają zarówno możliwości, jak i wyzwania związane z automatyzacją, interpretowalnością modeli, odpowiedzialnością

algorytmiczną oraz koniecznością tworzenia systemów odpornych, transparentnych i dostosowanych do realnych potrzeb użytkowników.

Myśl przewodnia tomu koncentruje się na roli sztucznej inteligencji jako technologii jednocześnie wzmacniającej możliwości analityczne i automatyzacyjne, a zarazem generującej nowe ryzyka dla bezpieczeństwa, wiarygodności informacji oraz zaufania do systemów cyfrowych. Publikacja jest adresowana do czytelnika technicznego i interdyscyplinarnego: badaczy i praktyków AI, cyberbezpieczeństwa oraz analityki danych, a także osób wdrażających rozwiązania algorytmiczne w organizacjach (w tym w edukacji, medycynie i sektorze finansowym), zainteresowanych zarówno stroną inżynierską, jak i konsekwencjami społecznymi. Kompozycja tomu ma charakter dwublokowy. Pierwszy blok obejmuje zagadnienia „AI w bezpieczeństwie i zaufaniu" – od ryzyk technicznych i podatności rozwiązań generatywnych, przez bezpieczeństwo informacyjne i cyberzagrożenia, po mechanizmy kształtujące zaufanie i postawy użytkowników wobec AI. Drugi blok prezentuje „AI w zastosowaniach sektorowych", pokazując metody oraz studia przypadków w obszarach takich jak diagnostyka i analiza obrazów, finanse, kreacja cyfrowa, edukacja czy systemy wizyjne. Dobór rozdziałów oparto na trzech kryteriach: (1) jednoznacznym osadzeniu tematu w problematyce bezpieczeństwa Internetu i/lub wiarygodnego wykorzystania AI, (2) dominacji perspektywy aplikacyjnej (metody, narzędzia, wyniki), oraz (3) komplementarności – tak, aby tom obejmował zarówno warstwę ryzyka i kontroli, jak i przekrojowe przykłady wdrożeń w różnych domenach. Układ treści koresponduje z zakresem rozdziałów ujętych w spisie treści.

Monografia stanowi przegląd aktualnych kierunków badań i wdrożeń, ukazując, że sztuczna inteligencja nie jest już wyłącznie narzędziem wspomagającym działania człowieka, lecz integralnym elementem infrastruktury współczesnego społeczeństwa. Prezentowane opracowania mają na celu przybliżenie czytelnikowi zarówno teoretycznych fundamentów, jak i praktycznych aspektów stosowania AI w kluczowych obszarach funkcjonowania świata cyfrowego. Dzięki temu monografia oferuje wieloaspektowe spojrzenie na rolę inteligentnych systemów i ich wpływ na rozwój bezpieczeństwa, innowacji oraz efektywności procesów w erze szybkiej i wszechobecnej cyfryzacji.

Miłej lektury – Redaktorzy

Aleksander Anchimiuk[1]

Politechnika Białostocka

https://doi.org/10.29316/9788368103304_1

INSECURE OUTPUT HANDLING IN LARGE LANGUAGE MODELS

NIEBEZPIECZNE PRZETWARZANIE DANYCH WYJŚCIOWYCH W DUŻYCH MODELACH JĘZYKOWYCH

Abstract

Large Language Models (LLMs) are increasingly deployed across domains such as automation, healthcare, and software development. While existing research predominantly focuses on input-related risks, such as prompt injection and adversarial attacks, the critical challenge of Insecure Output Handling (IOH), as identified by the OWASP Top 10 for LLMs, remains understudied in the academic literature. This paper addresses this gap by providing a formal analysis of IOH as a distinct security threat, in which model-generated outputs introduce vulnerabilities into downstream systems. Through controlled experiments focusing on XSS vulnerabilities, we demonstrate significant variation in model security postures, with baseline exploit rates ranging from 4.2% to 27.6% across different architectures. Crucially, our findings validate OWASP's concerns by showing that standard sanitisation techniques, recommended to mitigate IOH, provide 100% protection against successful XSS exploits in our experimental setup while maintaining negligible performance overhead. Furthermore, we examine how the European Union's AI

[1] student, e-mail: aleksander.anchimiuk.118968@student.pb.edu.pl

Act establishes concrete obligations for addressing such risks, positioning IOH mitigation not just as a technical best practice but as a regulatory imperative. The empirical evidence presented demonstrates that implementing output validation is both technically feasible and increasingly necessary for the deployment of compliant AI systems.

Keywords: AI security, large language models (LLMs), insecure output handling, application security, cross-site scripting (XSS), output sanitisation, EU AI Act, regulatory compliance

Streszczenie

Modele językowe dużej skali (LLM) są coraz częściej wykorzystywane w różnych dziedzinach, takich jak automatyzacja, opieka zdrowotna czy rozwój oprogramowania. Podczas gdy dotychczasowe badania koncentrują się głównie na zagrożeniach związanych z danymi wejściowymi, takich jak wstrzykiwanie promptów czy ataki adwersarialne, kluczowe wyzwanie, jakim jest niebezpieczne przetwarzanie danych wyjściowych (Insecure Output Handling, IOH) – zidentyfikowane przez listę OWASP Top 10 dla LLM – pozostaje słabo zbadane w literaturze naukowej. Niniejszy rozdział wypełnia tę lukę, dostarczając formalną analizę IOH jako odrębnego zagrożenia dla bezpieczeństwa, w ramach której generowane przez modele dane wyjściowe wprowadzają podatności do systemów je wykorzystujących. Na podstawie kontrolowanych eksperymentów dotyczących podatności typu XSS wykazano istotne różnice w poziomie bezpieczeństwa różnych architektur modeli – odsetek udanego wykorzystania podatności w badanej próbie wynosił od 4,2% do 27,6%. Co istotne, wyniki weryfikują obawy OWASP, pokazując, że w przeprowadzonym eksperymencie XSS standardowe techniki sanityzacji – rekomendowane do mitygacji IOH – zapewniły 100% ochrony przed skutecznymi atakami w badanej próbie, przy jednoczesnym minimalnym wpływie na wydajność. Ponadto rozdział analizuje, w jaki sposób Rozporządzenie UE o Sztucznej Inteligencji (AI Act) nakłada konkretne obowiązki w zakresie ograniczania tego typu ryzyk, czyniąc łagodzenie skutków IOH nie tylko najlepszą praktyką techniczną, lecz także wymogiem regulacyjnym. Przedstawione dane empiryczne dowodzą, że wdrożenie walidacji danych wyjściowych jest zarówno technicznie wykonalne, jak i coraz bardziej niezbędne dla zgodnego z przepisami wdrażania systemów AI.

Słowa kluczowe: bezpieczeństwo AI, modele językowe dużej skali (LLM), niebezpieczne przetwarzanie danych wyjściowych (Insecure Output Handling, IOH), bezpieczeństwo aplikacji, cross-site scripting (XSS), sanityzacja danych wyjściowych, Rozporządzenie UE o Sztucznej Inteligencji (AI Act), zgodność regulacyjna

Introduction

Large Language Models (LLMs) have become essential components of modern digital ecosystems, supporting chatbots, virtual assistants, code generation, automated testing, and database query automation (Sajadi et al., 2025). Security research on LLMs has predominantly focused on input-oriented threats, including prompt injection, adversarial manipulation, and training data poisoning (He et al., 2023; Gupta et al., 2017). However, outputs generated by LLMs can also serve as vectors for malicious activity. This phenomenon, which this paper defines as Insecure Output Handling (IOH), occurs when downstream systems consume model outputs without proper safeguards, such as sanitisation, schema validation, or parameterisation (Siddiq et al., 2023; Alwan et al., 2017; Hong et al., 2024; Luo et al., 2024).

The OWASP GenAI Top 10 2025 list identifies Improper Output Handling as a critical risk, highlighting that unsafe consumption of LLM outputs can lead to exploits, including XSS, SQL Injection, and remote code execution (OWASP GenAI Security Project, 2025). Building on this, the paper contributes by (1) formalising IOH as a distinct LLM security risk, (2) providing an empirical case study for XSS with mitigation analysis, and (3) situating IOH within the AI Act's regulatory framework, highlighting its technical and compliance relevance.

Empirical studies further demonstrate that LLM-generated code and HTML may propagate vulnerabilities if handled insecurely (He et al., 2023; Alwan et al., 2017; Hong et al., 2024). Despite these findings, systematic analyses of IOH remain limited. From a regulatory perspective, the European Union's Artificial Intelligence Act (AI Act, 2024) mandates risk management, robustness, technical documentation, and cybersecurity for high-risk AI systems (European Union, 2024 (AI Act)). While IOH is not named explicitly, its mitigation directly supports compliance with these provisions.

By synthesising perspectives from classical application security, emerging AI risk classifications, and regulatory policy, this work advances the discourse on AI security engineering. It provides actionable insights for researchers, practitioners, and policymakers.

Threat Definition: Insecure Output Handling (IOH)

Concept and Definition

Insecure Output Handling (IOH) refers to security risks arising when outputs generated by a Large Language Model (LLM) are integrated into downstream systems without sufficient safeguards, such as sanitisation, schema enforcement, or context-aware filtering. Unlike adversarial prompt injection, where an attacker manipulates the model's input, IOH exploits the fact that model-generated text or code may contain unsafe patterns itself. These outputs, when executed, displayed, or queried by other software components, can unintentionally enable attacks such as script injection, query manipulation, or even remote execution of commands. Prior research on code generation and AI-assisted programming confirms that LLMs often replicate insecure practices present in their training data, making them prone to propagate vulnerabilities by default (Siddiq et al., 2023; Alwan et al., 2017; Hong et al., 2024; Luo et al., 2024).

Adversarial Assumptions and Scope

To analyse IOH systematically, we consider the following assumptions:

- adversary model: the attacker does not require direct access to model internals (weights, training set). Instead, the attack surface is the downstream system that consumes raw LLM output,
- capabilities: by crafting natural prompts or leveraging ordinary user inputs, an attacker can steer the model toward producing code, queries, or markup that contain exploitable patterns,
- impact scope: any environment where LLM outputs are automatically processed – such as web rendering engines, database interfaces, automation pipelines, or monitoring tools – is within the risk boundary.

The threat of Insecure Output Handling (IOH) materialises when vulnerable downstream components process unsanitized LLM output. The following scenarios exemplify how this core vulnerability can lead to classic software exploits:

- Cross-Site Scripting (XSS): LLMs tasked with generating HTML, Markdown, or other content for web interfaces can inadvertently produce

payloads containing executable JavaScript. For instance, an LLM might be prompted to create a user-friendly message and output a string like <script>alert(“XSS”) </script>. If this output is directly embedded into a web page without proper encoding or sanitisation, it will execute in the user’s browser, leading to a client-side attack (Gupta et al., 2017; Siddiq et al., 2023). This scenario highlights IOH as the critical failure point where the application trusts and renders the model’s raw output.

- SQL Injection (SQLi): The widespread use of natural-language-to-SQL systems presents a significant IOH risk. An LLM might generate a SQL query constructed via string concatenation based on a user’s request. Without mandatory output parameterisation, a maliciously crafted prompt can induce the model to create a query containing a tautology (e.g., ‘OR ‘1’=’1’) or a UNION-based payload. If this generated string is executed directly against the database, it can lead to unauthorised data access, manipulation, or exfiltration (Alwan et al., 2017; Hong et al., 2024; Luo et al., 2024).
- Extended Attack Vectors:
  - Remote Code Execution (RCE): LLMs generating shell commands, CI/CD pipeline configurations, or code snippets can produce output that, when executed automatically by a downstream interpreter or shell, results in unauthorised command execution on the host system.
  - Server-Side Request Forgery (SSRF): An LLM suggesting URLs or API call structures might be manipulated into outputting an internal endpoint (e.g., http://169.254.169.254/). If a downstream component executes this request without validation, it could access sensitive metadata or internal services.
  - Configuration and Log Injection: LLM outputs embedded into configuration files, log entries, or monitoring dashboards can poison these systems. Subsequent parsing of these poisoned entries by administrative tools can lead to data corruption, false alerts, or even command execution, depending on the parsing logic.

Characteristics of the IOH Risk

IOH is distinctive because it fuses classical software security concerns with probabilistic AI behaviour:

- Unpredictability of outputs: LLM responses are stochastic, making the presence of unsafe patterns difficult to predict or detect in advance (Siddiq et al., 2023).
- Dependency on integration context: The risk severity is shaped less by the LLM itself than by how its output is consumed (e.g., web rendering, SQL execution, shell interpretation).
- Effectiveness of lightweight mitigations: Empirical studies show that simple measures – such as HTML sanitisation, strict schema validation, and query parameterisation – dramatically reduce exploitability without degrading system performance (Alwan, Younis, 2017; Hong et al., 2024; Luo et al., 2024).

This framing positions IOH as a distinct class of AI-enabled vulnerability. It demonstrates the necessity of handling model outputs with the same rigour as that traditionally applied to untrusted user inputs, thereby bridging the gap between AI-specific risk categories and established secure coding practices.

Methodology

To empirically assess the risks of Insecure Output Handling (IOH) for Cross-Site Scripting (XSS) in LLM-powered applications, we designed and implemented a controlled, automated experimentation pipeline. The study focused on the security implications of model-generated HTML outputs, simulating a realistic scenario where an LLM assists in generating user interface components.

Experimental methodology employs a black-box empirical approach to evaluate Insecure Output Handling (IOH) risks in operational contexts. The framework treats LLMs as opaque systems, focusing on observable input-output behaviours rather than internal mechanisms. We adopt a universal validation paradigm where all model outputs undergo security screening, reflecting zero-trust principles in production AI systems. The design incorporates comparative

analysis across different model architectures to identify consistent security patterns versus model-specific vulnerabilities.

Experimental Design and Model Selection

To quantitatively assess IOH risks, we employed a comparative framework analysing two architecturally distinct open-weight models:

- DeepSeek-R1 1.5B (DeepSeek, 2025): Representing the emerging class of efficient, reasoning-optimised models.
- Llama 3 8B (Meta, 2024 (Llama 3): A state-of-the-art general-purpose model.

This selection enables the investigation of how model scale and specialisation affect security properties and output safety.

Both selected models were deployed and interacted with locally using Ollama (version 0.12.3), a widely adopted framework for running LLMs. This ensured a controlled and consistent inference environment across all experiments. All prompts were submitted to, and responses collected from, the models via the Ollama API, simulating a typical developer workflow for integrating open-weight LLMs into applications.

Threat Modelling and Test Corpus

We constructed a stratified prompt corpus (N=500) designed to probe different aspects of IOH vulnerability:

- Tier 1: Innocuous Tasks (25%) – Baseline HTML generation tasks
- Tier 2: Security-Ambiguous Tasks (25%) – Requests implying interactivity
- Tier 3: Explicitly Risky Tasks (50%) – Direct requests for executable code

This distribution reflects real-world usage patterns while enabling granular analysis of risk factors.

Multi-Layer Defence Evaluation

Model output underwent systematic testing across a defence-in-depth matrix to evaluate protection effectiveness. The tested defence layers included:

- L1: Server-side Sanitization
  - Implementation: Bleach (allow-list based HTML sanitiser)
  - Scope: Universal
  - Purpose: Primary server-side filtering of malicious HTML patterns
- L2: Client-side Sanitization
  - Implementation: DOMPurify + Content Security Policy (CSP)
  - Scope: Universal
  - Purpose: Secondary client-side sanitisation and policy enforcement

This layered approach allowed for the assessment of individual and combined defence effectiveness against LLM-generated XSS vectors.

Analytical Metrics Framework

Primary Risk Assessment

- Baseline Exploit Rate (BER):

$$\text{BER} = \frac{\text{Number of Successfully Executed Outputs}}{\text{Total Outputs}} \times 100$$

Measures the inherent risk of a model generating exploitable outputs without any safeguards.

- Pattern Prevalence (PP):

$$\text{PP} = \frac{\text{Outputs with Detected XSS Patterns}}{\text{Total Outputs}} \times 100\%$$

Quantifies how often a model produces code that appears dangerous based on pattern matching.

- Exploitation Efficiency (EE):

$$\text{EE} = \frac{\text{Successfully Executed Outputs}}{\text{Outputs with Detected Patterns}} \times 100\%$$

Measures the quality and functional effectiveness of the detected attack vectors.

Defence Efficacy Metrics

- Exploit Neutralisation Rate (ENR):

$$\text{ENR} = \frac{\text{Exploits Blocked}}{\text{Total Exploits}} \times 100\%$$

The percentage of successfully executed attacks prevented by a defence mechanism.

- Pattern Reduction Rate (PRR):

$$\text{PRR} = \frac{\text{Patterns Removed}}{\text{Total Patterns}} \times 100\%$$

Measures the effectiveness of sanitisation in removing potentially dangerous code patterns.

- Cross-Model Defence Consistency: Evaluates whether a defence mechanism performs equally well across different model architectures.

Experimental Protocol

The sequential testing protocol ensured controlled comparisons:

- Baseline Characterisation: Establish inherent risk profiles for each model,
- Defence Isolation Testing: Measure individual layer effectiveness,
- Combined Defence Analysis: Assess synergistic effects of multi-layer protection,
- Robustness Validation: Test edge cases and adversarial variations.

This comprehensive framework enables not only the quantification of vulnerability but also a deeper investigation into the relationship between model architecture, training methodology, and security properties in generative AI systems.

Experimental Artefacts and Reproducibility

To ensure full transparency and reproducibility of our findings, the complete experimental pipeline, including prompt templates, sanitisation implementations, evaluation scripts, and raw results, is publicly available under an open-source license. The repository can be accessed at: GitHub Repository: https://github.com/AAnchimiuk/IOH-xss-experiment. This includes:

- Python source code of experimental pipeline,
- Raw results for pipeline execution for both models,
- Raw files of computed metrics for both models.

Results

Comparative Model Vulnerability Analysis

The experimental results reveal significant differences in security posture between the two evaluated models. DeepSeek-R1 exhibited a Baseline Exploit Rate (BER) of 4.2% (21/500), whereas Llama 3 8B demonstrated a substantially higher BER of 27.6% (138/500). This 6.6-fold difference in inherent vulnerability highlights the profound impact of model architecture and training methodology on security properties.

Pattern Detection vs. Actual Exploitation

A critical finding emerges from the disparity between pattern detection and successful exploitation, quantified by the Exploitation Efficiency (EE) metric:

- DeepSeek-R1 showed extensive Pattern Prevalence (PP) of 87.8% but a very low EE of 4.8%. This indicates the model generates a high volume of low-quality, non-functional, or easily mitigated attack vectors.
- Llama 3 demonstrated a lower PP (74.2%) but a significantly higher EE of 37.2%, indicating it produces fewer but more potent and readily exploitable vulnerabilities.

The scale of this “false positive” problem in static analysis is stark: 95.2% of DeepSeek’s and 62.8% of Llama 3’s flagged patterns were non-exploitable,

revealing a critical limitation of relying solely on pattern detection for security assessment.

Vector Distribution and Attack Patterns

Analysis of attack vectors reveals distinct model-specific tendencies:
DeepSeek-R1 Vector Profile:

- Primary Mechanism: High-volume, low sophistication,
- Script Tags: 924 instances (56.9% of total vectors),
- Event Handlers: 560 on_attr patterns,
- Advanced Vectors: Minimal (7 iframe srcdoc, two JavaScript protocols).

Llama 3 Vector Profile:

- Primary Mechanism: Balanced, high sophistication,
- Script Tags: 287 instances (41.6%),
- Advanced Vectors: Significant numbers of iframe srcdoc (28), meta refresh (27), and javascript: protocols (20),
- Attack Quality: Demonstrated higher capability for multi-vector attacks.

Defence Mechanism Efficacy

The implemented sanitisation defences demonstrated exceptional performance:

- Exploit Neutralisation Rate (ENR): both server-side (Bleach) and client-side (DOMPurify + CSP) sanitisation achieved 100% ENR, eliminating all 159 successful exploits,
- Pattern Reduction Rate (PRR): sanitisation reduced XSS patterns by 76.1% for DeepSeek-R1 and 88.4% for Llama 3,
- Functional Integrity: benign content maintained full functionality post-sanitisation,
- Cross-Model Defence Consistency: both defence layers provided identical 100% protection regardless of model architecture or vulnerability profile.

Risk Assessment Summary

Table 1. Summary of results from experimental metrics

Metric	DeepSeek-R1:1.5B	Llama 3:8B
Baseline Exploit Rate (BER)	4.2%	27.6%
Pattern Prevalence (PP)	87.8%	74.2%
Exploitation Efficiency (EE)	4.8%	37.2%
Residual Risk (Post-Defence)	0%	0%
Primary Risk Characteristic	Volume of low-quality vectors	Quality of exploitable vectors

Source: own study.

The results demonstrate that while Llama 3 generates more immediately dangerous outputs, both models produce vulnerabilities that necessitate comprehensive output sanitisation. The 100% effectiveness of standard web security measures against LLM-generated XSS provides compelling evidence for integrating these established practices into AI-augmented systems (Table 1).

Discussion

Technical Perspective

This study demonstrates that Insecure Output Handling (IOH) bridges two domains that have traditionally been treated in isolation: adversarial risks in Large Language Models (LLMs) and well-known web and software vulnerabilities. Our empirical investigation focused specifically on Cross-Site Scripting (XSS), providing systematic, quantitative evidence that even benign prompts, without adversarial manipulation, can yield exploitable HTML/JavaScript outputs when passed directly to downstream applications.

The XSS scenario reflects long-standing security concerns: generated HTML and Markdown outputs often contain inline JavaScript or unsafe attributes, reproducing vectors that have been extensively documented in XSS research (Gupta et al., 2017). Our experiments confirm that "off-the-shelf" LLMs do not incorporate security considerations into their generative logic by default (Sajadi et al., 2025; He et al., 2023). A significant contribution of this work is the quantification of the Exploit Success Rate (ESR) for XSS under both baseline

and secured configurations. This empirical evidence shows that lightweight safeguards – such as HTML sanitisation and restrictive Content Security Policies – can significantly reduce or even eliminate exploitability.

While our experimental validation centred on XSS, the IOH paradigm extends to other critical vulnerabilities. The SQL injection (SQLi) scenario, drawn from prior literature (Alwan, Younis, 2017; Hong et al., 2024), illustrates this clearly. LLMs frequently generate queries via unsafe string concatenation, a practice identified as insecure in both classical surveys (Alwan et al., 2017) and more recent evaluations of NL-to-SQL systems (Hong et al., 2024). Similarly, threats like Remote Code Execution (RCE) or Server-Side Request Forgery (SSRF) represent high-risk IOH vectors where generated code or suggested URLs could compromise systems if executed without validation.

However, the technical perspective also highlights significant limitations and nuances revealed by our data.

First, mitigations are highly context dependent. Our experiments confirmed the total effectiveness of HTML sanitisation against XSS, but this defence is specific to that vector. A Content Security Policy (CSP) may block inline JavaScript but not malicious data URLs; parameterisation eliminates SQLi but does nothing against insecure shell script generation. Therefore, the choice of output sanitisation must be precisely tailored to the integration context and the potential attack vectors it enables.

Second, our results underscore that vulnerability detection and successful exploitation are distinct problems. We observed a high volume of detected XSS patterns (e.g. 87.8% for DeepSeek-R1) but a much lower rate of successful execution (4.2% for the same model). This "pattern-to-exploit gap" means that static analysis alone is an insufficient metric for risk assessment, as it generates numerous false positives and fails to account for the functional quality and context of the attack vector.

Finally, the probabilistic nature of LLMs means that a model's security posture cannot be guaranteed by testing a finite set of prompts. A seemingly safe model, or one that generates mostly low-quality, non-functional vectors, could still produce a high-quality, exploitable output in a previously untested scenario. This inherent unpredictability, coupled with the high stakes of a single successful

exploit, necessitates continuous runtime monitoring and defence-in-depth rather than relying on one-off audits or the assumption of general model ‹safety›.

Regulatory Perspective

From a regulatory standpoint, IOH raises direct concerns under the European Union's Artificial Intelligence Act (European Union, 2024 (AI Act)). Although IOH is not named explicitly, the Act requires high-risk AI systems to comply with several obligations that IOH directly undermines. Our empirically demonstrated XSS risks provide a concrete example of how IOH can violate these provisions:

- risk management (Art. 9): our ESR-based evaluation for XSS offers a structured way to quantify IOH risk and integrate it into risk management pipelines. Similar to penetration testing in classical software security, IOH testing can become a verifiable compliance artefact,
- robustness and accuracy (Art. 15): systems that propagate XSS through generated outputs cannot be considered robust. By contrast, applying sanitisation fulfils the robustness expectation, as demonstrated by our results,
- transparency and documentation (Art. 13): maintaining records of how outputs are sanitised, validated, and monitored supports compliance with documentation requirements. Such logs can also serve as evidence during regulatory audits,
- human oversight (Art. 14): oversight cannot be limited to inputs; it must also extend to outputs. Unsafe outputs must be either blocked automatically or flagged for human review, aligning with the principle of "meaningful human control".

These obligations resonate with guidance from the OWASP GenAI Security Project (2025) (OWASP GenAI Security Project, 2025), which explicitly identifies Improper Output Handling as a top-tier risk for generative AI. IOH thus becomes not just a best-practice issue but a compliance necessity: failure to mitigate it could result in enforcement action, liability for damages, or loss of certification to deploy high-risk AI systems in the EU.

Broader Implications

The implications of IOH extend beyond immediate technical controls and regulatory compliance, encompassing questions of trust, liability, and systemic risk.

First, the sectoral impact is profound. In healthcare, insecure outputs embedded in electronic health records or clinical decision support tools could compromise patient safety and violate GDPR. In finance, LLM-generated code or queries vulnerable to injection attacks could distort fraud detection pipelines or risk assessments, raising compliance concerns under MiFID II or Basel III. In critical infrastructure, IOH could surface in CI/CD pipelines, where generated configuration scripts introduce vulnerabilities into industrial control systems.

Second, IOH introduces significant liability and governance challenges. Under the AI Act, non-compliance may lead to substantial fines (European Union, 2024 (AI Act)). But even outside formal regulation, organisations risk civil liability if insecure outputs lead to breaches. For instance, if an LLM-generated SQL injection (as identified in prior works (Alwan et al., 2017; Hong et al., 2024)) compromises a hospital's patient database, legal disputes could emerge over whether responsibility lies with the AI provider, the integrator, or the deploying institution.

Finally, IOH points toward an evolving security paradigm. Historically, input validation became a non-negotiable baseline in secure system design. Our results suggest that output validation may soon be regarded as equally essential for AI-enabled systems. Extending this into practice could mean that future development pipelines treat every LLM response as "untrusted input," subject to the same rigour of sanitisation and validation as external user data.

Conclusion

This study introduces and empirically grounds the concept of Insecure Output Handling (IOH) as a critical class of risks in LLM-based systems. Through a controlled experimental framework focused on Cross-Site Scripting (XSS), this work provides quantitative evidence that:

- LLMs inherently generate insecure outputs, with the nature and severity of the risk being highly model-dependent. Our results showed a dramatic 6.6-fold difference in the baseline exploit rate between the two tested models (DeepSeek-R1 at 4.2% vs Llama 3 at 27.6%), underscoring that architectural and training choices directly impact security posture.
- IOH poses immediate and practical security threats when outputs are consumed without validation. The successful execution of generated XSS vectors in a simulated production environment confirms that classical web vulnerabilities are readily reintroduced through LLM outputs.
- Conventional, lightweight defences are highly effective and essential. Our central finding is that standard security measures – specifically HTML sanitisation and Content Security Policies (CSPs) – achieved 100% effectiveness in neutralising all successful XSS exploits across both models. This provides a strong, evidence-based argument for integrating these established practices into AI-integrated systems.
- The mitigation of IOH is not merely a technical best practice but a cornerstone of regulatory compliance. Our empirical framework for quantifying IOH risk directly supports obligations under the EU AI Act concerning risk management, robustness, and documentation, transforming a technical vulnerability into a manageable compliance requirement.

Overall, this study positions IOH as a critical convergence point between AI safety research and established application security. By demonstrating that a probabilistic, generative component can be secured using deterministic, classical defences, we bridge a crucial gap in the engineering of trustworthy AI systems.

Future work must expand upon this experimental foundation. The immediate priority is to extend rigorous empirical testing to other critical IOH vectors, such as SQL Injection, remote code execution, and server-side request forgery, to build a comprehensive, data-driven understanding of the IOH threat landscape. Furthermore, proactive approaches – such as fine-tuning models on secure coding corpora, developing standardised output validation libraries, and exploring runtime monitoring solutions – present vital pathways for building more secure and resilient generative AI applications.

References

1. Alwan, Z. S., Younis, M. F. (2017). Detection and prevention of SQL injection attack: a survey. *International Journal of Computer Science and Mobile Computing*, *6*(8), 5-17.
2. Gupta, S., Gupta, B. B. (2017). Cross-site scripting (XSS) attacks and defence mechanisms: classification and state of the art. *International Journal of System Assurance Engineering and Management*, *8*, 512-530. https://doi.org/10.1007/s13198-015-0376-0
3. He, J., Vechev, M. (2023). Large Language Models for Code: Security Hardening and Adversarial Testing. Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security (CCS '23), 1865-1879. https://doi.org/10.1145/3576915.3623175
4. Hong, Z., Yuan, Z., Chen, H., Zhang, Q., Huang, F., Huang, X. (2024). Knowledge-to-SQL: Enhancing SQL Generation with Data Expert LLM. Findings of the Association for Computational Linguistics: ACL 2024, 10997-11008. https://doi.org/10.18653/v1/2024.findings-acl.653
5. Luo, Y., Lin, Z., Zhang, Y., Sun, J., Lin, C., Xu, C., Su, X., Shen, Y., Guo, J., Gong, Y. (2024). Ensuring Safe and High-Quality Outputs: A Guideline Library Approach for Language Models. *Proceedings of NAACL-HLT 2024,* 1152-1197. https://doi.org/10.18653/v1/2024.naacl-long.65
6. Sajadi, A., Le, B., Nguyen, A., Damevski, K., Chatterjee, P. (2025). Do LLMs consider security? An empirical study on responses to programming questions. *Empirical Software Engineering*. https://doi.org/10.1007/s10664-025-10658-6
7. Siddiq, M. L., Santos, J. C. (2023). Generate and Pray: Using SALLMs to Evaluate the Security of LLM Generated Code. *IEEE/ACM International Conference on Automated Software Engineering Workshops (ASEW)*. https://doi.org/10.1145/3691621.3694934

Netography

1. DeepSeek. (2025). *DeepSeek-R1: A Series of Efficient Reasoning Models*. Retrieved from: https://github.com/deepseek-ai/DeepSeek-R1 (access: 06.11.2025).
2. European Union (2024). *Artificial Intelligence Act*. *Official Journal of the European Union*.

3. Meta AI. (2024). *The Llama 3 Herd of Models*. Retrieved from: https://github.com/meta-llama/llama3 (access: 06.11.2025).
4. OWASP GenAI Security Project. (2025). *LLM052025 – Improper Output Handling*. Retrieved from: https://genai.owasp.org/llmrisk/llm052025-improper-output-handling/ (access: 06.11.2025).

Sławomir Krzysztof Czubaj[1]

Akademia Bialska im. Jana Pawła II, Biała Podlaska

https://doi.org/10.29316/9788368103304_2

ROLA SZTUCZNEJ INTELIGENCJI W KSZTAŁTOWANIU BEZPIECZEŃSTWA INFORMACYJNEGO INTERNETU JAKO MEDIUM WSPÓŁCZESNEJ CYWILIZACJI

THE ROLE OF ARTIFICIAL INTELLIGENCE IN SHAPING INFORMATION SECURITY ON THE INTERNET AS A MEDIUM OF MODERN CIVILISATION

Streszczenie

Celem pracy jest analiza wpływu sztucznej inteligencji na bezpieczeństwo informacyjne Internetu jako kluczowego medium współczesnej cywilizacji. Rozdział bada transformację Internetu z tradycyjnej infrastruktury komunikacyjnej w złożone środowisko algorytmiczne, w którym systemy generatywne, rekomendacyjne i agentowe aktywnie współuczestniczą w procesach poznawczych, komunikacyjnych i decyzyjnych użytkowników. Przedmiotem analizy są korzyści, zagrożenia oraz wyzwania regulacyjne wynikające z dynamicznego rozwoju technologii AI. W pracy zastosowano metodę przeglądu literatury naukowej i raportów instytucjonalnych z lat 2023-2025, w tym publikacji OECD, UNESCO, Stanford Institute for Human-Centered Artificial Intelligence, Europolu,

[1] https://orcid.org/0000-0002-7561-2277, e-mail: s.czubaj@dyd.akademiabialska.pl

DataReportal, Nature Machine Intelligence oraz ACM. Dokonano krytycznej analizy najważniejszych modeli generatywnych i systemów algorytmicznych wpływających na środowisko informacyjne, obejmującej zarówno aspekty technologiczne, jak i społeczne. Uwzględniono również aktualne ramy regulacyjne, takie jak Artificial Intelligence Act UE czy amerykańskie Executive Orders. Analiza ma charakter jakościowy i opiera się na metodzie studiów przypadków (case-based review) oraz analizie porównawczej aktualnych podejść regulacyjnych. Przeprowadzona analiza wykazała, że sztuczna inteligencja stała się jednym z kluczowych mechanizmów współczesnego Internetu, wpływając na sposób organizacji i kontroli informacji. Modele generatywne i systemy rekomendacyjne zwiększają efektywność dostępu do wiedzy, lecz jednocześnie wzmacniają zagrożenia informacyjne, takie jak dezinformacja generatywna, manipulacja poznawcza, nieprzejrzystość algorytmów oraz erozja autonomii poznawczej użytkowników. Stwierdzono również rosnące znaczenie systemów agentowych i „Internetu agentów", które radykalnie zmieniają strukturę uczestników ekosystemu informacyjnego. Ponadto odnotowano znaczące różnice w tempie i zakresie wdrażania regulacji AI między regionami świata. Sztuczna inteligencja redefiniuje Internet jako środowisko informacyjne, przekształcając go w układ kognitywno-algorytmiczny współtworzony przez ludzi i systemy autonomiczne. Odpowiedzialne wykorzystanie AI wymaga wzmocnienia infrastruktury regulacyjnej, wprowadzenia mechanizmów audytu algorytmicznego, budowania odporności cyfrowej społeczeństwa oraz rozwijania kompetencji krytycznych użytkowników. Internet przyszłości będzie środowiskiem współpracy człowieka i inteligentnych systemów, jednak warunkiem zachowania zaufania społecznego i integralności informacyjnej jest transparentne, etyczne i kontrolowane wdrażanie technologii AI.

Słowa kluczowe: sztuczna inteligencja, bezpieczeństwo informacyjne, Internet, dezinformacja, generatywna AI, algorytmizacja, Internet agentów, odporność cyfrowa

Abstract

The aim of this study is to examine the influence of artificial intelligence on the information security of the Internet as a key medium of contemporary civilisation. The article investigates the transformation of the Internet from a traditional communication infrastructure into a complex algorithmic environment in which generative, recommender, and agent-based systems actively participate in users' cognitive, communicative, and decision-making processes. The study focuses on the benefits, risks, and regulatory challenges arising from the rapid development of AI technologies. The study applies a qualitative literature review of scientific publications and institutional reports from 2023-2025, including sources from OECD, UNESCO, Stanford Institute for Human-Centered Artificial Intelligence, Europolu, DataReportal, Europol, DataReportal, Nature Machine Intelligence, and ACM.

A critical analysis of the main generative models and algorithmic systems shaping the contemporary information environment was conducted, encompassing technological, social, and epistemic perspectives. Current regulatory frameworks – such as the EU Artificial Intelligence Act and U.S. Executive Orders – were examined using comparative policy analysis. Additionally, selected case studies were used to identify emerging risks and societal impacts of algorithmic systems. The findings indicate that artificial intelligence has become a central mechanism of the modern Internet, reshaping how information is organised, interpreted, and distributed. Generative models and recommendation systems enhance access to knowledge but simultaneously amplify information security threats, including generative disinformation, cognitive manipulation, algorithmic opacity, and the erosion of users' cognitive autonomy. The analysis also reveals the growing importance of agent-based systems and the emerging "agentic Internet," which fundamentally alters the structure of participants in the digital ecosystem. Significant regional disparities were identified in the development and enforcement of AI regulations. Artificial intelligence is redefining the Internet as an information environment, transforming it into a cognitive-algorithmic system co-constructed by humans and autonomous technologies. Ensuring responsible AI adoption requires strengthening regulatory infrastructures, implementing algorithmic audit mechanisms, fostering digital resilience, and developing users' critical competencies. The future Internet will increasingly rely on collaboration between humans and intelligent systems; however, maintaining public trust and safeguarding information integrity will depend on transparent, ethical, and accountable AI governance.

Keywords: artificial intelligence, information security, Internet, disinformation, generative AI, algorithmization, agentic Internet, digital resilience

Wstęp

Internet stanowi fundament współczesnego świata, będąc zarówno podstawowym kanałem dystrybucji informacji, jak i narzędziem pracy, edukacji, kultury oraz komunikacji społecznej. Według najnowszych danych Digital Poland 2025 ponad 88% mieszkańców Polski korzysta z Internetu, spędzając średnio ponad sześć godzin dziennie w przestrzeni cyfrowej, w tym ponad dwie godziny w mediach społecznościowych (DataReportal, 2025b). Skala uczestnictwa cyfrowego potwierdzona jest także w danych globalnych, które wskazują na systematyczny wzrost intensywności korzystania z Internetu w ujęciu międzynarodowym (DataReportal, 2025a). Internet staje się zatem nie tylko narzędziem, ale kluczowym

medium funkcjonowania człowieka XXI wieku oraz obszarem o znaczeniu strategicznym dla bezpieczeństwa informacyjnego.

W ostatnich latach obserwujemy jednak fundamentalną transformację tego środowiska. Internet przestaje być wyłącznie przestrzenią treści tworzonych i zarządzanych przez ludzi – coraz większą rolę pełni w nim sztuczna inteligencja, która staje się aktywnym uczestnikiem procesów obiegu informacji. Zmianę tę wzmacniają także trendy technologiczne wskazywane w raportach globalnych, takich jak Ericsson Mobility Report (Ericsson, 2024b), które podkreślają rosnącą integrację systemów komunikacji, danych i sztucznej inteligencji. Współczesne systemy AI:

- generują i przetwarzają treści,
- filtrują i moderują informacje,
- rekomendują zasoby i kształtują ekspozycję informacyjną użytkowników,
- analizują i przewidują zachowania odbiorców,
- automatyzują komunikację i inicjują interakcje,
- współuczestniczą w procesach poznawczych, decyzyjnych i komunikacyjnych.

Modele generatywne, takie jak GPT, Claude, Gemini czy Llama, a także multimodalne systemy obrazu i wideo (DALL-E, Stable Diffusion, Sora) wpływają nie tylko na strukturę Internetu jako zasobu, lecz przede wszystkim na integralność i wiarygodność informacji. Wraz z pojawieniem się autonomicznych architektur agentowych rośnie także znaczenie koncepcji Internetu agentów i Internetu zmysłów (6G + AI), w których systemy sztucznej inteligencji przejmują część funkcji tradycyjnych procesów komunikacji i przetwarzania danych (Sehad i in., 2024).

Równocześnie intensyfikują się wyzwania związane z bezpieczeństwem informacyjnym: manipulacja treściami, dezinformacja generatywna, automatyzacja propagandy, naruszenia prywatności, filtracja algorytmiczna czy ryzyka wynikające z nierównowagi informacyjnej. Rozwija się również zjawisko uzależnienia od systemów algorytmicznych i ograniczania autonomii poznawczej użytkowników.

Rozdział podejmuje analizę tych procesów oraz ich konsekwencji dla funkcjonowania Internetu jako złożonego, inteligentnego medium współczesnej cywilizacji, ze szczególnym uwzględnieniem roli sztucznej inteligencji w kształtowaniu bezpieczeństwa informacyjnego i integralności przestrzeni cyfrowej.

Czym jest Internet we współczesnej cywilizacji

Internet należy do kluczowych osiągnięć cywilizacji informacyjnej i stanowi fundamentalną infrastrukturę społeczeństwa wiedzy. Od momentu powstania pełni funkcję globalnego medium komunikacyjnego, umożliwiającego transfer informacji, wymianę danych, rozwój nauki, gospodarki i kultury oraz kształtowanie relacji społecznych. Jego istota wykracza jednak poza rolę kanału komunikacyjnego – Internet stał się środowiskiem życia społecznego, ekonomicznego i edukacyjnego, a obecnie, za sprawą sztucznej inteligencji, również obszarem o kluczowym znaczeniu dla bezpieczeństwa informacyjnego.

Zgodnie z najnowszymi danymi dotyczącymi użytkowania Internetu, w Polsce 88,4% mieszkańców korzysta z sieci, a średni dzienny czas aktywności online wynosi ponad 6 godzin (DataReportal, 2025b). Globalne raporty potwierdzają systematyczny wzrost intensywności uczestnictwa cyfrowego oraz rosnącą zależność społeczeństw i gospodarek od infrastruktury internetowej (DataReportal, 2025a). Jednocześnie analizy *Ericsson Mobility Report* wskazują, że zwiększa się zapotrzebowanie na transmisję danych oraz integrację systemów sieciowych z technologiami sztucznej inteligencji, co wpływa na odporność i stabilność cyfrowego ekosystemu informacyjnego (Ericsson, 2024b; Ericsson, 2024a).

Wraz z rozwojem technologii informacyjnych Internet przekształcił się z pasywnego repozytorium danych w środowisko interaktywne, a następnie – dzięki dynamicznie rozwijającym się modelom AI – w przestrzeń współtworzoną przez systemy inteligentne. Zjawisko to jest szeroko opisywane we współczesnych pracach poświęconych cyfryzacji i społeczeństwu sieci, które wskazują na przejście od Internetu komunikacyjnego do Internetu algorytmicznego, w którym przetwarzanie danych staje się integralnym elementem struktury społecznej (van Dijck i in., 2025; Perriam, Carter, 2022). W literaturze opisuje się współczesny Internet jako układ hybrydowy, w którym współdziałają:

- użytkownicy,
- instytucje i organizacje,
- infrastruktura sieciowa i obliczeniowa,
- algorytmy decyzyjne,
- systemy generatywne i agentowe.

Takie ujęcie jest zgodne z koncepcjami nowoczesnych społeczeństw cyfrowych, które wskazują na coraz bardziej współzależny charakter relacji między człowiekiem a infrastrukturą technologiczną (Camarinha-Matos i in., 2022). Jednocześnie rozwój narzędzi sztucznej inteligencji wywołuje redefinicję klasycznego modelu komunikacji – algorytmy zaczynają pełnić funkcję współuczestników procesów informacyjnych, a nie tylko narzędzi przesyłu treści.

Trend ten znajduje potwierdzenie w globalnych raportach dotyczących SI, które opisują gwałtowne zwiększenie roli modeli generatywnych, systemów rekomendacyjnych i architektur agentowych w kształtowaniu przepływu informacji oraz organizacji wiedzy w Internecie (Stanford HAI, 2025). Wraz z tym procesem Internet przechodzi od modelu transmisji danych do modelu poznawczo-algorytmicznego, w którym technologie AI interpretują, filtrują i generują treści, przyczyniając się zarówno do wzrostu efektywności komunikacyjnej, jak i nowych wyzwań dotyczących integralności informacyjnej. Zjawisko to prowadzi do szeregu implikacji:

- redefinicji pojęcia autorytetu informacyjnego,
- transformacji modelu wyszukiwania wiedzy – od indeksacji do generatywnych odpowiedzi AI,
- algorytmicznego kształtowania opinii publicznej,
- automatyzacji interakcji społecznych i komunikacji,
- przekształcenia modeli gospodarczych i edukacyjnych,
- wzrostu podatności na manipulację informacyjną i dezinformację generowaną przez modele AI.

Wraz z ekspansją sztucznej inteligencji Internet przestaje być neutralną infrastrukturą odbicia rzeczywistości społecznej – staje się środowiskiem współtworzącym tę rzeczywistość, wpływającym na percepcję, decyzje i zachowania użytkowników. W tym sensie Internet nie jest już jedynie technologią pomocniczą, lecz strukturą poznawczą współczesnej cywilizacji, w której bezpieczeństwo informacyjne odgrywa rolę fundamentalną.

Sztuczna inteligencja jako mechanizm Internetu

Rozwój sztucznej inteligencji w ostatnich latach doprowadził do fundamentalnej transformacji Internetu. Współcześnie nie jest on jedynie przestrzenią

gromadzenia i dystrybucji informacji, lecz staje się środowiskiem poznawczym, w którym algorytmy pełnią funkcję filtrów, moderatorów, a coraz częściej także generatorów treści oraz autonomicznych uczestników ekosystemu informacyjnego. Najnowsze analizy globalne, m.in. *Artificial Intelligence Index Report* (Stanford HAI, 2025) wskazują, że systemy oparte na uczeniu maszynowym i modelach generatywnych stają się kluczowym elementem przetwarzania danych w sieci.

Postęp technologiczny w obszarze uczenia głębokiego, transformatorów oraz multimodalnych sieci neuronowych doprowadził do powstania systemów zdolnych do rozumienia kontekstu, analizy semantycznej, przewidywania zachowań użytkowników oraz generowania treści tekstowych, graficznych i dźwiękowych zgodnych z ludzkimi wzorcami kulturowymi. Rozwiązania te są integrowane praktycznie we wszystkich sektorach funkcjonowania Internetu – od wyszukiwarek i mediów społecznościowych, przez edukację i finanse, aż po sektor zdrowia i administrację publiczną. Tendencja ta wpisuje się w globalną koncepcję Internetu zmysłów (Internet of Senses), w której generatywna AI umożliwia coraz bardziej imersyjne formy komunikacji (Sehad i in., 2024).

W początkowej fazie Internet opierał się na algorytmach statycznych, które indeksowały zasoby i porządkowały treści według słów kluczowych. Wprowadzenie uczenia maszynowego doprowadziło do powstania modelu dynamicznego, w którym dostęp do informacji jest kształtowany przez personalizację i predykcyjne modelowanie zachowań użytkowników. Współczesne mechanizmy wyszukiwawcze interpretują zapytania semantycznie, filtrują treści, rekomendują zasoby, moderują komunikację, automatyzują przepływy pracy i generują odpowiedzi zamiast jedynie wskazywać źródła. Zmiana ta prowadzi do stopniowego przejścia od klasycznej wyszukiwarki (search engine) do tzw. silnika odpowiedzi (answer engine), co jest obserwowane m.in. w badaniach nad ruchem internetowym w kontekście generatywnego Google AI Overviews (Scire, 2025).

Modele generatywne zmieniają epistemologię Internetu – wiedza nie musi być już samodzielnie wyszukiwana, ponieważ jest wytwarzana i syntetyzowana przez systemy AI w odpowiedzi na potrzeby użytkownika. Systemy te potrafią generować tekst, kod, obrazy, wideo, a także personalizować narracje w czasie rzeczywistym. Najnowsze przeglądy literatury wskazują, że technologie generatywne zaczynają pełnić funkcję współtwórców wiedzy i informacji cyfrowej, a ich wpływ na społeczeństwo będzie narastał (Musiol, 2024).

Kolejnym istotnym etapem rozwoju Internetu jest popularyzacja systemów agentowych. Agenci AI są zdolni do autonomicznego działania w środowisku cyfrowym – planują i wykonują operacje wieloetapowe, komunikują się z innymi systemami oraz API usług internetowych, a także automatyzują procesy społeczne i gospodarcze. Najnowsze opracowania opisują rodzącą się koncepcję Agentic Web – ekosystemu, w którym głównymi uczestnikami komunikacji są nie tylko ludzie, lecz także autonomiczne systemy AI współdziałające ze sobą (Caetano i in., 2025). Internet staje się w ten sposób przestrzenią koegzystencji ludzi i cyfrowych bytów zdolnych do podejmowania decyzji i zarządzania informacją. Współczesny Internet można zatem opisać jako system złożony z trzech współzależnych warstw:

- warstwy ludzkiej (użytkownicy, instytucje, twórcy treści),
- warstwy infrastrukturalnej (sieci, centra danych, chmura, urządzenia mobilne),
- warstwy algorytmiczno-kognitywnej (systemy predykcyjne, modele generatywne, filtry informacyjne, agenci AI).

Integracja tych warstw prowadzi do powstania symbiotycznego ekosystemu poznawczego, w którym informacja nie jest wyłącznie gromadzona i przesyłana, lecz jest aktywnie interpretowana, modulowana i generowana. To właśnie w tej warstwie algorytmicznej rozstrzygają się współcześnie kwestie bezpieczeństwa informacyjnego – takie jak integralność treści, autentyczność przekazów, odporność na manipulację czy kontrola nad przepływem informacji.

Współczesne społeczeństwo informacyjne w epoce sztucznej inteligencji

Społeczeństwo informacyjne, którego fundamentalnym zasobem jest informacja, ulega obecnie kolejnemu etapowi transformacji – przejściu do społeczeństwa algorytmicznego. W strukturze społecznej coraz większą rolę odgrywają dane, modele predykcyjne oraz inteligentne systemy informacyjne, które w coraz większym stopniu wpływają na procesy poznawcze, decyzyjne i komunikacyjne. Raporty OECD (2023a) oraz UNESCO (2023) podkreślają, że w erze sztucznej inteligencji informacja przestaje być kategorią bierną – staje się dynamicznie modelowaną strukturą, której interpretacja i dystrybucja zależą od działania algorytmów.

Internet stanowi podstawową przestrzeń pracy, edukacji, komunikacji i rozrywki współczesnego człowieka, a jego rola w procesie socjalizacji młodego pokolenia jest szczególnie znacząca. Badania pokazują, że dzieci i młodzież należą do grup najbardziej intensywnie korzystających z technologii cyfrowych, co sprawia, że środowisko online pełni funkcję zarówno narzędzia edukacyjnego, jak i platformy formowania tożsamości cyfrowej oraz kompetencji społecznych (OECD, 2023b; UNESCO, 2023). Współczesna młodzież funkcjonuje w warunkach stałej ekspozycji na systemy algorytmiczne i traktuje Internet jako naturalne środowisko nauki, rozwoju i komunikacji, jednocześnie będąc odbiorcą treści modelowanych przez mechanizmy personalizacji. Najnowsze przeglądy badań potwierdzają, że personalizacja może prowadzić do ograniczenia różnorodności informacyjnej, powstawania baniek filtrujących oraz kształtowania opinii na podstawie zawężonego zestawu treści (Ahmmad i in., 2025; Ludwig i in., 2025). Jednocześnie rozwój modeli generatywnych prowadzi do pojawienia się zjawiska automatyzowanej komunikacji społecznej, w której użytkownik wchodzi w interakcję nie tylko z innymi ludźmi, lecz także z systemami AI, często nieświadomie, ponieważ treści generowane przez algorytmy nie zawsze są odróżnialne od treści tworzonych przez człowieka (Musiol, 2024).

Rozszerzenie funkcji Internetu oraz intensywna personalizacja doprowadziły także do nasilenia zjawiska uzależnienia cyfrowego. Mechanizmy natychmiastowej gratyfikacji, algorytmiczne wzmacnianie zaangażowania oraz projektowanie bodźców opartych na zmiennej nagrodzie sprzyjają kompulsywnemu korzystaniu z technologii, co znajduje potwierdzenie w najnowszych metaanalizach (Lu i in., 2025). Wraz z rozwojem sztucznej inteligencji pojawia się również nowa forma zależności poznawczej – uzależnienie od systemów wspomagania myślenia i decyzji. Narzędzia generatywne, takie jak modele językowe, ułatwiają tworzenie treści, wyszukiwanie informacji i rozwiązywanie problemów, lecz jednocześnie mogą ograniczać potrzebę samodzielnej refleksji, redukować kompetencje krytyczne oraz prowadzić do delegowania procesów poznawczych na algorytmy (UNESCO, 2023). Zjawisko to określa się mianem algorytmicznego ubezwłasnowolnienia poznawczego, gdy jednostka zaczyna polegać na systemach AI w zakresie uczenia się i podejmowania decyzji, jednocześnie tracąc część autonomii intelektualnej.

W obszarze edukacji sztuczna inteligencja wymusza głęboką przebudowę klasycznego modelu nauczania. Raporty UNESCO (2023) oraz OECD (2023c) wskazują, że edukacja XXI wieku powinna koncentrować się na rozwijaniu kompetencji krytycznych, metapoznawczych oraz umiejętności oceny wiarygodności źródeł. Równocześnie konieczne jest wprowadzanie edukacji dotyczącej etyki technologicznej, bezpieczeństwa cyfrowego oraz zasad odpowiedzialnego korzystania z AI. W literaturze podkreśla się, że narzędzia oparte na sztucznej inteligencji powinny pełnić funkcję wspierającą, a nie zastępującą proces dydaktyczny, a kluczową rolę odgrywa rozwój kompetencji algorytmicznych – czyli podstawowej świadomości dotyczącej sposobu działania i ograniczeń systemów AI (OECD, 2023c).

W rezultacie społeczeństwo informacyjne wchodzi w fazę, w której relacje między człowiekiem a technologią ulegają złożonym przekształceniom. Internet nie jest już przestrzenią neutralną ani bierną, lecz środowiskiem aktywnie współtworzonym przez algorytmy. Sztuczna inteligencja staje się integralnym elementem procesów społecznych, poznawczych i edukacyjnych, prowadząc do powstania nowego modelu organizacji życia społecznego – społeczeństwa algorytmicznego, w którym bezpieczeństwo informacyjne i autonomia poznawcza stają się kluczowymi wyzwaniami współczesnej cywilizacji.

Zagrożenia, regulacje i wyzwania etyczne w erze sztucznej inteligencji

Rozwój sztucznej inteligencji w środowisku internetowym generuje nie tylko nowe możliwości, lecz także zagrożenia o charakterze społecznym, gospodarczym, politycznym i poznawczym. Wraz ze wzrostem złożoności modeli algorytmicznych zwiększa się ich wpływ na życie publiczne, co rodzi konieczność analizy kwestii bezpieczeństwa, przejrzystości, odpowiedzialności i ochrony praw jednostki. Najnowsze raporty OECD (2024) oraz Europol Innovation Lab (2024) wskazują, że technologie generatywne radykalnie przekształcają krajobraz komunikacji cyfrowej i wymagają systemowego podejścia regulacyjnego.

Jednym z najbardziej dynamicznie rozwijających się zagrożeń jest dezinformacja generatywna i manipulacja informacyjna. Modele tworzą realistyczne teksty, obrazy, materiały audio i wideo, często niemożliwe do odróżnienia od treści rzeczywistych. Ułatwia to powstawanie syntetycznych wiadomości, deepfake'ów,

automatycznych profili, kampanii wpływu oraz narzędzi propagandowych opartych na modelach językowych. Badania ACM Computing Surveys podkreślają, że deepfake'i stały się kluczowym wyzwaniem dla bezpieczeństwa informacyjnego i zaufania społecznego (Wang i in., 2024). Zjawisko to ma wymiar polityczny i kulturowy – generatywna AI staje się narzędziem formowania opinii publicznej, wpływając na procesy demokratyczne oraz integralność informacyjną (OECD, 2024; Negreiro, 2025).

Równolegle rośnie znaczenie zagrożeń związanych z prywatnością, danymi osobowymi i bezpieczeństwem cyfrowym. Sztuczna inteligencja opiera się na analizie ogromnych zbiorów danych, co rodzi napięcie pomiędzy postępem technologicznym a prawem jednostki do prywatności. Kluczowe problemy obejmują masowe profilowanie użytkowników, predykcyjne modelowanie ich zachowań, wykorzystywanie danych biometrycznych oraz ryzyko cyberataków na infrastrukturę zasilającą modele AI. Report UE *Facing Reality?* (Europol Innovation Lab, 2024) podkreśla dodatkowo nieprzejrzystość algorytmów, która utrudnia audyt, ocenę zagrożeń oraz przypisanie odpowiedzialności za decyzje podejmowane przez modele.

Istotne wyzwania pojawiają się również w obszarze prawa autorskiego i własności treści generowanych przez sztuczną inteligencję. Rozwój modeli generatywnych prowadzi do debat dotyczących praw własności do treści wytwarzanej przez SI, legalności wykorzystywania danych treningowych, ryzyka naruszania praw twórców oraz odpowiedzialności za nieprawdziwe lub szkodliwe treści. Dyskusje te potwierdzają, że dotychczasowe koncepcje autorstwa, kreatywności i własności intelektualnej stają się niewystarczające wobec twórczości algorytmicznej.

W odpowiedzi na te wyzwania rządy oraz organizacje międzynarodowe opracowują systemy regulacji i nadzoru nad rozwojem sztucznej inteligencji. Unia Europejska przyjęła w 2024 r. przełomowy *Artificial Intelligence Act* – pierwszy kompleksowy system klasyfikacji ryzyka i nadzoru nad AI, obejmujący wymogi transparentności, obowiązki audytowe oraz regulacje dotyczące modeli wysokiego ryzyka (European Union, 2024). W Stanach Zjednoczonych administracja federalna wydała *Executive Order 14110* dotyczący zgodności, bezpieczeństwa i ochrony danych (Federal Register, 2023), choć część jego przepisów została w 2025 r. uchylona przez nową administrację (Reuters, 2025; The White House, 2025). Debaty regulacyjne dotyczą również wpływu silników odpowiedzi

(answer engines) – według analiz (Deck, 2025; Kint, 2025) i Reutersa (2025), generatywne systemy wyszukiwania wpływają na rynek informacji, ruch wydawców oraz konstrukcję ekosystemu medialnego.

Nadrzędnym wyzwaniem pozostają etyka algorytmiczna i odpowiedzialność technologiczna. Nowoczesne systemy Ai muszą być rozwijane w sposób zgodny z wartościami demokratycznymi, zasadami sprawiedliwości oraz ochroną praw człowieka. Kluczowe zagadnienia obejmują przejrzystość modeli, zapobieganie dyskryminacji algorytmicznej, określanie odpowiedzialności za skutki działania systemów autonomicznych oraz wprowadzanie mechanizmów nadzoru i kontroli. W centrum dyskusji znajduje się koncepcja odpowiedzialności technologicznej, zakładająca, że rozwój sztucznej inteligencji musi być nierozerwalnie związany z oceną jej konsekwencji społecznych i ochroną jednostki przed nadużyciami.

Podsumowanie

Rozwój sztucznej inteligencji stał się jednym z najważniejszych czynników redefiniujących współczesny Internet oraz architekturę globalnej cywilizacji informacyjnej. Internet, który pierwotnie pełnił funkcję infrastruktury komunikacyjnej i informacyjnej, przekształcił się w złożone środowisko kognitywno-algorytmiczne, w którym systemy AI nie tylko wspierają procesy poznawcze, lecz także współuczestniczą w ich przebiegu. Generatywne modele językowe, systemy rekomendacyjne i agentowe modele działania zmieniły sposób pozyskiwania, interpretowania oraz dystrybucji informacji, czyniąc z AI jeden z kluczowych mechanizmów współczesnego Internetu.

Transformacja ta niesie zarówno wymierne korzyści, jak i istotne zagrożenia. Z jednej strony sztuczna inteligencja zwiększa dostępność wiedzy, usprawnia analizę danych, automatyzuje procesy komunikacji i tworzenia treści oraz wspiera rozwój gospodarczy i naukowy. Z drugiej strony prowadzi do nowych wyzwań dotyczących bezpieczeństwa informacyjnego, odporności społecznej, integralności treści cyfrowych oraz autonomii poznawczej użytkowników. Wraz z ekspansją systemów generatywnych rośnie ryzyko dezinformacji, manipulacji, uprzedzeń algorytmicznych, naruszeń prywatności oraz koncentracji władzy informacyjnej w rękach podmiotów kontrolujących infrastrukturę AI. Jednocześnie niejasność

autorstwa, przejrzystości modeli i odpowiedzialności za działania systemów autonomicznych staje się jednym z kluczowych dylematów współczesnej polityki technologicznej.

Wobec tych wyzwań konieczne jest całościowe podejście do rozwoju sztucznej inteligencji, obejmujące regulacje prawne, mechanizmy audytu algorytmicznego oraz rozwój kompetencji społecznych. Dokumenty prawne, takie jak europejski Artificial Intelligence Act czy amerykańskie regulacje federalne, potwierdzają, że AI wymaga nadzoru instytucjonalnego i normatywnego. Jednak równie istotnym elementem jest edukacja – kształtowanie kompetencji cyfrowych, świadomości algorytmicznej, umiejętności krytycznej analizy treści oraz zdolności do świadomego korzystania z zaawansowanych systemów informacyjnych.

Najbliższe lata przyniosą konieczność wypracowania równowagi między innowacyjnością technologii a ochroną wartości demokratycznych i humanistycznych. Internet przyszłości będzie środowiskiem współpracy człowieka i inteligentnych systemów, w którym rola człowieka nie powinna zostać zredukowana, lecz wzmocniona poprzez odpowiednią refleksję, kompetencje i narzędzia. Odpowiedzialne, transparentne i etyczne wykorzystanie sztucznej inteligencji stanowi fundament zrównoważonego rozwoju społeczeństwa cyfrowego oraz warunek bezpiecznej ewolucji Internetu jako medium kluczowego dla funkcjonowania współczesnej cywilizacji.

Literatura

1. Ahmmad, M., Shahzad, K., Iqbal, A., Latif, M. (2025). Trap of social media algorithms: A systematic review of filter bubbles and echo chambers and their impact on youth. *Societies, 15*(11), 301. https://doi.org/10.3390/soc15110301
2. Caetano, A., Verma, K., Taheri, A., Kumaran, R., Chen, Z., Chen, J., Höllerer, T., Sra, M. (2025). *Agentic workflows for conversational human-AI interaction design* (preprint). arXiv. https://doi.org/10.48550/arXiv.2501.18002
3. Camarinha-Matos, L. M., Ortiz, A., Boucher, X., Osório, A. L. (Eds.). (2022). *Collaborative Networks in Digitalization and Society 5.0: 23rd IFIP WG 5.5 Working Conference on Virtual Enterprises, PRO-VE 2022, Lisbon, Portugal, September 19-21, 2022, Proceedings* (IFIP Advances in Information and Communication Technology, Vol. 662). Springer Cham. https://doi.org/10.1007/978-3-031-14844-6

4. Europol Innovation Lab. (2024, March 4). *Facing reality? Law enforcement and the challenge of deepfakes: An observatory report from the Europol Innovation Lab*. Publications Office of the European Union.
5. Federal Register. (2023, November 1). *Executive Order 14110: Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence* (88 Fed. Reg. 76,403).
6. Ludwig, K., Müller, P., Nikolajevic, N., Grote, A. (2025). Putting 'filter bubble' effects to the test: Evidence on the polarizing impact of ideology-based news recommendation from two experiments in Germany and the U.S. *Information, Communication & Society*, 1-20. Advance online publication. https://doi.org/10.1080/1369118X.2024.2435998
7. Lu, P., Qiu, J., Huang, S., Wang, X., Han, S., Zhu, S., Ning, Y., Zeng, F.-F., Yuan, Y. (2025). Interventions for digital addiction: Umbrella review of meta-analyses. *Journal of Medical Internet Research, 27*, e59656. https://doi.org/10.2196/59656
8. Musiol, M. (2024). *Generative AI: Navigating the course to the artificial general intelligence future* (ISBN 978-1-394-20591-2). Wiley.
9. OECD. (2023a). *AI and the future of skills: Volume 2*. OECD Publishing. https://doi.org/10.1787/ae96ca6c-en
10. OECD. (2023b). *Employment outlook 2023: Skill needs in the age of AI*. OECD Publishing. https://doi.org/10.1787/7c5219e7-en
11. OECD. (2023c). *Artificial intelligence and the future of skills: Project overview*. OECD Publishing. https://doi.org/10.1787/11f00efb-en
12. OECD. (2024). *Facts not fakes: Tackling disinformation, strengthening information integrity*. OECD Publishing. https://doi.org/10.1787/8055b8f6-en
13. Perriam, J., Carter, S. (Eds.). (2021). *Understanding Digital Societies. SAGE Publications Ltd.*
14. Sehad, N., Bariah, L., Hamidouche, W., Hellaoui, H., Jäntti, R., Debbah, M. (2025). Generative AI for immersive communication: The next frontier in Internet-of-Senses through 6G. *IEEE Communications Magazine, 63*(2), 31-43. https://doi.org/10.1109/MCOM.001.2400199
15. Sehad, N., Bariah, L., Hamidouche, W., Hellaoui, H., Jäntti, R., Debbah, M. (2024). *Generative AI for immersive communication: The next frontier in Internet-of-Senses through 6G* (preprint). arXiv. https://doi.org/10.48550/arXiv.2404.01713
16. van Dijck, J., van Es, K., Helmond, A., van der Vlist, F. (Eds.). (2025). *Governing the digital society: Platforms, artificial intelligence, and public values*. Amsterdam University Press. https://doi.org/10.5117/9789048562718

17. Wang, T., Liao, X., Chow, K. P., Lin, X., Wang, Y. (2024). *Deepfake detection: A comprehensive survey from the reliability perspective*. *ACM Computing Surveys*, *57*(3), 1-35. https://doi.org/10.1145/3699710

Netografia

1. DataReportal. (2025a). *Digital 2025: Global Overview Report*. DataReportal – Global Digital Insights. Pobrane z: https://datareportal.com/reports/digital-2025-global-overview-report (dostęp: 06.11.2025).
2. DataReportal. (2025b). *Digital in Poland 2025*. DataReportal – Global Digital Insights. Pobrane z: https://datareportal.com/reports/digital-2025-poland (dostęp: 06.11.2025).
3. Deck, A. (2025, May 5). *Google is using content from publishers who "opt out" of other AI training to power AI Overviews*. Nieman Journalism Lab. Pobrane z: https://www.niemanlab.org/2025/05/google-is-using-content-from-publishers-who-opt-out-of-other-ai-training-to-power-ai-overviews/ (dostęp: 06.11.2025).
4. Ericsson. (2024a, November). *Ericsson mobility report: November 2024*. Ericsson. Pobrane z: https://www.ericsson.com/en/reports-and-papers/mobility-report (dostęp: 06.11.2025).
5. Ericsson. (2024b, June). *Ericsson mobility report: June 2024* (PDF). Ericsson. Pobrane z: https://www.ericsson.com/en/reports-and-papers/mobility-report (dostęp: 06.11.2025).
6. European Union. (2024, July 12). *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence*. Official Journal of the European Union L 273, 1-175. Pobrane z: https://data.europa.eu/eli/reg/2024/1689/oj (dostęp: 06.11.2025).
7. Kint, J. (2025, August 14). Facts: Google's push to AI hurts publisher traffic. Digital Content Next. Pobrane z: https://digitalcontentnext.org/blog/2025/08/14/facts-googles-push-to-ai-hurts-publisher-traffic/ (dostęp: 06.11.2025).
8. Negreiro Achiaga, M. D. M. (2025, July 3). *Children and deepfakes* (Briefing). European Parliamentary Research Service, European Parliament. Pobrane z: https://www.europarl.europa.eu/RegData/etudes/BRIE/2025/775855/EPRS_BRI(2025)775855_EN.pdf (dostęp: 06.11.2025).
9. Reuters. (2025, January 21). *Trump revokes Biden executive order addressing AI risks*. Reuters. Pobrane z: https://www.reuters.com/world/us/trump-revokes-biden-executive-order-addressing-ai-risks-2025-01-21/ (dostęp: 06.11.2025).

10. Reuters. (2025, July 4). *EU antitrust complaint vs. Google AI Overviews*. Reuters. Pobrane z: https://www.reuters.com/world/europe/eu-antitrust-complaint-vs-google-ai-overviews-2025-07-04/ (dostęp: 06.11.2025).
11. Scire, S. (2025). *Google referrals are steady, but social and direct are down, according to new Chartbeat data*. Nieman Journalism Lab. Pobrane z: https://www.niemanlab.org/reading/google-referrals-are-steady-but-social-and-direct-are-down-according-to-new-chartbeat-data/ (dostęp: 06.11.2025).
12. Stanford Institute for Human-Centered Artificial Intelligence. (2025). *Artificial Intelligence Index Report 2025*. Stanford University. Pobrane z: https://aiindex.stanford.edu/report/ (dostęp: 06.11.2025).
13. Ten Teije, S. (2025, August 11). *Google's AI Overview is not yet harming traffic, but publishers should remain alert*. INMA. Pobrane z: https://www.inma.org/blogs/big-data-for-news-publishers/post.cfm/google-s-ai-overview-is-not-yet-harming-traffic-but-publishers-should-remain-alert (dostęp: 06.11.2025).
14. The White House. (2025, January 23). *Removing barriers to American leadership in AI*. Pobrane z: https://www.whitehouse.gov/briefing-room/presidential-actions/2025/01/23/removing-barriers-to-american-leadership-in-ai/ (dostęp: 06.11.2025).
15. UNESCO. (2023). *Guidance for generative AI in education and research*. UNESCO Publishing. Pobrane z: https://unesdoc.unesco.org/ark:/48223/pf0000386081 (dostęp: 06.11.2025).

Marek Kania[1]

Akademia Policji w Szczytnie

https://doi.org/10.29316/9788368103304_3

ROLA SZTUCZNEJ INTELIGENCJI W WYKRYWANIU I NEUTRALIZACJI ZAGROŻEŃ CYBERNETYCZNYCH

THE ROLE OF ARTIFICIAL INTELLIGENCE IN DETECTING AND NEUTRALISING CYBER THREATS

Streszczenie

Rozwój technologii sztucznej inteligencji (SI) w ostatnich latach znacząco zmienił sposób identyfikowania, monitorowania i neutralizowania zagrożeń w cyberprzestrzeni. Dzięki algorytmom uczenia maszynowego i głębokiego uczenia, systemy bezpieczeństwa są w stanie analizować ogromne wolumeny danych w czasie rzeczywistym, wykrywając anomalie i wzorce charakterystyczne dla działań cyberprzestępczych. Rozdział przedstawia przegląd aktualnych zastosowań AI w cyberbezpieczeństwie, ze szczególnym uwzględnieniem metod detekcji ataków typu phishing, ransomware oraz APT (Advanced Persistent Threats). Omówione zostaną także wyzwania związane z implementacją SI, w tym ryzyko błędnych alarmów, podatność algorytmów na manipulacje oraz kwestie etyczne. Analiza obejmuje również studia przypadków zastosowań AI w infrastrukturze

[1] e-mail: marekkania@interia.pl

krytycznej oraz w sektorze prywatnym. Wnioski wskazują, że skuteczne wykorzystanie Ai wymaga synergii technologii z czynnikiem ludzkim, ciągłego doskonalenia algorytmów oraz integracji z polityką bezpieczeństwa organizacji.

Słowa kluczowe: sztuczna inteligencja, cyberbezpieczeństwo, wykrywanie zagrożeń, neutralizacja cyberataków, uczenie maszynowe, APT

Abstract

The development of artificial intelligence (AI) technology in recent years has significantly altered the way cyber threats are identified, monitored, and neutralised. Thanks to machine learning and deep learning algorithms, security systems can analyse vast volumes of data in real time, detecting anomalies and patterns characteristic of cybercriminal activity. This chapter provides an overview of current AI applications in cybersecurity, with a particular focus on methods for detecting phishing, ransomware, and APT (Advanced Persistent Threats). The challenges associated with AI implementation, including the risk of false alarms, the vulnerability of algorithms to manipulation, and ethical issues, will also be discussed. The analysis also includes case studies of AI applications in critical infrastructure and the private sector. The conclusions suggest that the effective use of AI necessitates synergy between technology and the human factor, continuous algorithm improvement, and integration with the organisation's security policy.

Keywords: artificial intelligence, cybersecurity, threat detection, cyberattack neutralisation, machine learning, APT

Wstęp

Sztuczna inteligencja (SI) stała się jednym z kluczowych narzędzi w walce z rosnącą liczbą i złożonością zagrożeń cybernetycznych. W dobie globalnej cyfryzacji, gdy codziennie generowane są ogromne ilości danych, tradycyjne metody monitorowania i ochrony systemów informatycznych stają się niewystarczające (Nguyen i in., 2019). Algorytmy uczenia maszynowego oraz głębokiego uczenia pozwalają na analizę ruchu sieciowego w czasie rzeczywistym, identyfikację nietypowych wzorców i przewidywanie potencjalnych ataków, zanim spowodują one znaczące szkody (Sommer, Paxson, 2010).

Zastosowanie AI w cyberbezpieczeństwie nie ogranicza się jedynie do pasywnej detekcji zagrożeń. Coraz częściej jest ona integrowana z systemami automatycznego reagowania, które mogą izolować zainfekowane segmenty sieci, blokować podejrzane transakcje czy modyfikować reguły zapór sieciowych w odpowiedzi na dynamicznie zmieniającą się sytuację (Buczak, Guven, 2016). W efekcie Ai staje się nie tylko narzędziem wspierającym zespoły SOC (Security Operations Center), ale również autonomicznym elementem infrastruktury bezpieczeństwa.

Równocześnie rozwój tych technologii generuje nowe wyzwania. Pojawia się problem fałszywych alarmów, które mogą obciążać systemy i zespół reagowania. Istnieje również ryzyko, że zaawansowani cyberprzestępcy będą w stanie manipulować algorytmami, wykorzystując tzw. adversarial AI – specjalnie spreparowane dane wejściowe mające na celu zmylenie systemów detekcji (Biggio, Roli, 2018). Dodatkowo, kwestie etyczne, takie jak prywatność użytkowników, przejrzystość działania algorytmów czy odpowiedzialność za błędne decyzje SI, pozostają otwarte i wymagają dalszych badań.

Celem niniejszego rozdziału jest przedstawienie roli sztucznej inteligencji w wykrywaniu i neutralizacji zagrożeń cybernetycznych, analiza jej potencjału oraz ograniczeń, a także omówienie studiów przypadków wdrożeń w sektorze publicznym i prywatnym.

Podsumowując, wprowadzenie ukazuje rosnące znaczenie sztucznej inteligencji w kontekście globalnego krajobrazu zagrożeń cybernetycznych. Wyraźnie zaznaczono, że rozwój technologii informatycznych niesie zarówno potencjał zwiększenia bezpieczeństwa, jak i ryzyko eskalacji ataków o wysokim stopniu zaawansowania. Celem rozdziału jest kompleksowa analiza roli AI w procesach detekcji i neutralizacji takich zagrożeń.

Podstawy teoretyczne

Definicje i typy zagrożeń cybernetycznych

Zagrożenia cybernetyczne obejmują wszelkie działania skierowane na naruszenie poufności, integralności lub dostępności zasobów informacyjnych, które mogą mieć charakter celowy (atak) lub przypadkowy (incydent bezpieczeństwa) (ENISA, 2023). Do najczęściej spotykanych typów zagrożeń należą:

- Phishing – próby wyłudzenia danych uwierzytelniających lub informacji poufnych poprzez podszywanie się pod zaufane podmioty (Jagatic i in., 2007).
- Ransomware – oprogramowanie blokujące dostęp do systemu lub danych do czasu uiszczenia okupu (Kharraz i in., 2015).
- Advanced Persistent Threats (APT) – długotrwałe, wysoce zaawansowane kampanie cybernetyczne prowadzone w celu szpiegostwa lub sabotażu (Chen i in., 2014).
- DDoS (Distributed Denial of Service) – ataki mające na celu przeciążenie systemu lub usługi, uniemożliwiając jej normalne działanie.
- Ataki socjotechniczne – manipulowanie zachowaniem użytkowników w celu uzyskania dostępu do zasobów lub informacji.

Każdy z tych typów zagrożeń może mieć poważne konsekwencje finansowe, prawne i wizerunkowe dla organizacji. Z tego względu wymagają one coraz bardziej zaawansowanych metod detekcji i neutralizacji.

Techniki AI w analizie zagrożeń

Sztuczna inteligencja w cyberbezpieczeństwie opiera się na algorytmach zdolnych do automatycznego uczenia się na podstawie danych. Najważniejsze techniki obejmują:

- Uczenie nadzorowane – modele trenowane na oznaczonych zbiorach danych, pozwalające na klasyfikację ruchu sieciowego jako „normalny" lub „złośliwy" (Sahoo i in., 2019).
- Uczenie nienadzorowane – wykrywanie anomalii w danych bez wcześniejszego oznaczenia przykładów, co jest szczególnie przydatne przy identyfikacji nowych, nieznanych zagrożeń (Chandola i in., 2009).
- Uczenie głębokie (deep learning) – wykorzystanie wielowarstwowych sieci neuronowych do analizy złożonych wzorców w ruchu sieciowym i logach systemowych (Javaid i in., 2016).
- Przetwarzanie języka naturalnego (NLP) – analiza treści wiadomości e-mail, postów w mediach społecznościowych i dokumentów pod kątem treści phishingowych lub dezinformacyjnych (Zhang i in., 2017).

Połączenie tych technik umożliwia tworzenie systemów, które nie tylko rozpoznają znane zagrożenia, ale również adaptują się do nowych metod ataku.

Rozdział ten porządkuje kluczowe pojęcia oraz typologie cyberzagrożeń, co stanowi fundament dla dalszej analizy wykorzystania AI w cyberbezpieczeństwie. Przedstawiono różne kategorie ataków – od phishingu po ransomware i APT – wskazując na ich dynamiczną ewolucję. Ujęcie to podkreśla konieczność ciągłego dostosowywania narzędzi ochrony do nowych form zagrożeń.

Zastosowania AI w wykrywaniu zagrożeń

Analiza ruchu sieciowego i detekcja anomalii

Jednym z kluczowych zastosowań sztucznej inteligencji w cyberbezpieczeństwie jest analiza ruchu sieciowego w czasie rzeczywistym. Algorytmy uczenia maszynowego potrafią identyfikować wzorce nietypowego zachowania, takie jak nagły wzrost liczby pakietów, niestandardowe porty komunikacyjne czy nietypowe adresy IP, które mogą wskazywać na próbę ataku (Sommer, Paxson, 2010). Modele nienadzorowane, np. algorytmy klastrowania, wykrywają anomalie nawet w przypadku ataków typu zero-day, które nie zostały wcześniej zidentyfikowane.

Wykrywanie phishingu i ataków socjotechnicznych

SI odgrywa rosnącą rolę w wykrywaniu ataków phishingowych poprzez analizę treści wiadomości, struktur linków oraz metadanych nadawcy. Wykorzystanie przetwarzania języka naturalnego (NLP) pozwala na identyfikację podejrzanych zwrotów, nietypowych schematów gramatycznych oraz manipulacyjnego języka perswazji (Zhang i in., 2017). Ponadto, algorytmy mogą automatycznie blokować domeny, które podszywają się pod zaufane instytucje, minimalizując ryzyko interakcji użytkownika z fałszywymi stronami.

Ochrona przed ransomware i APT

Ransomware i APT należą do najbardziej kosztownych i destrukcyjnych rodzajów cyberzagrożeń. W ich przypadku AI jest stosowana do monitorowania integralności plików, wykrywania nietypowych operacji na systemach plików

oraz identyfikowania podejrzanych procesów (Kharraz i in., 2015). W przypadku APT algorytmy analizują sekwencje zdarzeń w długim okresie czasu, co pozwala na wykrycie ukrytych działań, takich jak stopniowe eksfiltracje danych czy ciche instalowanie backdoorów (Chen i in., 2014).

Analiza pokazuje, że systemy oparte na AI znacząco zwiększają skuteczność wykrywania ataków dzięki zdolności do przetwarzania ogromnych wolumenów danych w czasie rzeczywistym. Przedstawione przykłady zastosowań, takie jak sieci neuronowe czy algorytmy uczenia maszynowego, wskazują na rosnącą adaptacyjność i precyzję w identyfikacji anomalii. Wskazano również na potrzebę dalszego doskonalenia algorytmów, aby minimalizować fałszywe alarmy.

Neutralizacja zagrożeń z wykorzystaniem SI

Automatyczne reagowanie i izolacja incydentów

Jedną z najważniejszych przewag sztucznej inteligencji w obszarze cyberbezpieczeństwa jest możliwość automatycznego reagowania na wykryte incydenty. Systemy klasy SOAR (Security Orchestration, Automation and Response) integrują mechanizmy detekcji z narzędziami reagowania, umożliwiając natychmiastowe odcięcie zainfekowanego segmentu sieci, zablokowanie konta użytkownika lub wycofanie podejrzanego pliku z obiegu (Islam, Babar, Nepal, 2019). Dzięki temu czas reakcji liczony jest w sekundach, a skala szkód jest minimalizowana.

Wsparcie dla zespołów SOC (Security Operations Center)

SI nie zastępuje w pełni zespołów SOC, ale stanowi ich kluczowe wsparcie, filtrując i priorytetyzując zgłoszenia. Dzięki automatycznemu analizowaniu ogromnych wolumenów logów i alertów, analitycy mogą skupić się na najbardziej krytycznych incydentach. Ponadto, algorytmy uczenia maszynowego potrafią tworzyć predykcyjne modele, które wskazują prawdopodobieństwo wystąpienia określonego ataku w najbliższym czasie, co umożliwia działania prewencyjne (Buczak, Guven, 2016).

Podsumowując, sztuczna inteligencja znajduje zastosowanie nie tylko w detekcji, lecz również w aktywnej neutralizacji incydentów. Automatyczne reakcje systemów AI pozwalają ograniczyć czas reakcji i skalę szkód. Przedstawione przypadki dowodzą, że integracja AI z istniejącymi narzędziami bezpieczeństwa zwiększa odporność infrastruktury cyfrowej.

Wyzwania i ograniczenia

Fałszywe alarmy i nadwrażliwość systemów

Jednym z największych problemów systemów Ai w cyberbezpieczeństwie jest wysoki poziom false positives (fałszywych alarmów). Nadwrażliwość algorytmów, szczególnie w początkowej fazie wdrożenia, może prowadzić do przeciążenia zespołów SOC i ignorowania części alertów (Sommer, Paxson, 2010). Wymaga to odpowiedniego kalibrowania modeli oraz stosowania metod hybrydowych, łączących analizę automatyczną z weryfikacją ekspercką.

Manipulacja algorytmami i adversarial AI

Coraz większym wyzwaniem jest adversarial AI, czyli celowe wprowadzanie danych mających zmylić systemy SI. Przykładem może być modyfikacja pakietów sieciowych lub plików w taki sposób, aby algorytmy klasyfikowały je jako bezpieczne, mimo że zawierają złośliwy kod (Biggio, Roli, 2018). Obrona przed tego typu atakami wymaga opracowywania odpornych modeli oraz stosowania mechanizmów detekcji prób manipulacji.

Aspekty etyczne i prawne

Zastosowanie AI w cyberbezpieczeństwie wiąże się również z pytaniami o prywatność, przejrzystość działania algorytmów oraz odpowiedzialność prawną za ich decyzje. Przetwarzanie dużych zbiorów danych może prowadzić do naruszeń RODO lub innych regulacji o ochronie danych osobowych (European Union, 2016). Ponadto brak transparentności w działaniu sieci neuronowych („black box") utrudnia audyt i zrozumienie podstaw podjętych decyzji, co jest szczególnie istotne w kontekście incydentów wpływających na infrastrukturę krytyczną.

Rozdział ten podkreśla, że pomimo wielu zalet, wdrożenie AI w obszarze bezpieczeństwa wiąże się z wyzwaniami technicznymi, prawnymi i etycznymi. Kluczowe problemy to m.in. podatność algorytmów na ataki typu adversarial, ryzyko nadmiernej automatyzacji oraz konieczność ochrony danych wykorzystywanych w procesach uczenia. Wskazano, że równoważenie innowacji z bezpieczeństwem i zgodnością prawną pozostaje kluczowe.

Studia przypadków

Wdrożenie systemu AI w sektorze publicznym – ochrona infrastruktury krytycznej

W 2023 roku jeden z europejskich operatorów systemów energetycznych wdrożył platformę AI opartą na uczeniu głębokim, której celem była analiza ruchu w sieciach sterowania przemysłowego (ICS/SCADA). System monitorował sygnały w czasie rzeczywistym i porównywał je z modelem „normalnego" funkcjonowania instalacji. W przypadku wykrycia anomalii następowało natychmiastowe odcięcie podejrzanego segmentu oraz automatyczne powiadomienie zespołu SOC. W ciągu pierwszych sześciu miesięcy wykryto i zneutralizowano trzy incydenty związane z próbami wstrzyknięcia złośliwego kodu do systemu sterowania turbinami (ENISA, 2023).

Zastosowanie Ai w sektorze prywatnym – ochrona danych klientów banku

Duży bank komercyjny w Azji wdrożył narzędzie oparte na uczeniu maszynowym, które analizowało wszystkie transakcje w czasie rzeczywistym. Algorytmy były trenowane na historycznych danych dotyczących oszustw finansowych, co pozwoliło na szybkie wychwytywanie nietypowych wzorców, takich jak nagłe wypłaty dużych kwot w nietypowych lokalizacjach czy próby logowania z kilku krajów w krótkim czasie. W ciągu pierwszego roku od wdrożenia liczba udanych ataków typu account takeover spadła o 72% (Nguyen i in., 2019).

Analiza skuteczności rozwiązań

Porównanie obu przypadków pokazuje, że skuteczność AI zależy nie tylko od jakości algorytmów, ale również od stopnia integracji systemu z procesami operacyjnymi organizacji. W obu wdrożeniach kluczowe było połączenie automatycznych mechanizmów detekcji i reakcji z czynnikiem ludzkim – analitykami zdolnymi do weryfikacji decyzji systemu oraz dostosowywania jego parametrów.

Analizowane przykłady wdrożeń AI w systemach bezpieczeństwa dowodzą realnej skuteczności tych rozwiązań w różnych sektorach – od finansowego po administrację publiczną. Każdy z opisanych przypadków potwierdza, że właściwie zaprojektowane systemy AI mogą znacząco skrócić czas reakcji i zmniejszyć skalę incydentów. Praktyczne doświadczenia są jednocześnie źródłem wiedzy dla dalszych innowacji w tej dziedzinie.

Podsumowanie

Analiza roli sztucznej inteligencji w wykrywaniu i neutralizacji zagrożeń cybernetycznych pokazuje, że technologia ta ma ogromny potencjał w zwiększaniu skuteczności działań ochronnych. Algorytmy uczenia maszynowego i głębokiego uczenia umożliwiają wykrywanie anomalii w czasie rzeczywistym, automatyczną reakcję na incydenty oraz przewidywanie nowych form ataków. Studia przypadków potwierdzają, że właściwie wdrożone systemy AI mogą znacząco redukować liczbę skutecznych ataków zarówno w sektorze publicznym, jak i prywatnym.

Jednocześnie, technologia ta nie jest wolna od ograniczeń. Fałszywe alarmy, podatność na manipulacje typu adversarial AI oraz problemy z przejrzystością działania algorytmów wymagają stałego doskonalenia modeli i mechanizmów bezpieczeństwa. Kluczowym elementem pozostaje synergia między automatyzacją a czynnikiem ludzkim – wykwalifikowani analitycy nadal są niezbędni do weryfikacji decyzji systemów i reagowania na sytuacje wyjątkowe.

Rekomenduje się, aby organizacje:

1. Inwestowały w hybrydowe systemy bezpieczeństwa, łączące AI z nadzorem eksperckim.
2. Regularnie aktualizowały modele i dane treningowe, aby nadążać za ewoluującymi zagrożeniami.

3. Wdrażały mechanizmy obrony przed adversarial AI, w tym detekcję prób manipulacji algorytmami.
4. Zapewniały zgodność z regulacjami prawnymi oraz przejrzystość działania systemów w kontekście audytu i odpowiedzialności.

W perspektywie najbliższych lat można spodziewać się, że rozwój AI w cyberbezpieczeństwie będzie postępował w kierunku większej autonomii systemów, integracji z innymi technologiami (takimi jak blockchain czy kwantowa kryptografia) oraz rozszerzenia zastosowań na kolejne obszary, w tym ochronę prywatności i przeciwdziałanie dezinformacji.

Podsumowując całość rozważań, rozdział wskazuje na strategiczne znaczenie AI w walce z zagrożeniami cybernetycznymi. Wdrażanie inteligentnych systemów wymaga jednak nie tylko inwestycji w technologię, ale i w rozwój kompetencji specjalistów. Rekomenduje się dalsze badania nad bezpieczeństwem algorytmów oraz rozwój standardów prawnych i etycznych dla ich stosowania.

Literatura

1. Biggio, B., Roli, F. (2018). Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition*, 84, 317-331. https://doi.org/10.1016/j.patcog.2018.07.023
2. Buczak, A. L., Guven, E. (2016). A survey of data mining and machine learning methods for cyber security intrusion detection. *IEEE Communications Surveys & Tutorials*, *18*(2), 1153-1176. https://doi.org/10.1109/COMST.2015.2494502
3. Chandola, V., Banerjee, A., Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys*, 41(3), 1-58. https://doi.org/10.1145/1541880.1541882
4. Chen, P., Desmet, L., Huygens, C. (2014). A study on advanced persistent threats. *IFIP International Conference on Communications and Multimedia Security*, 63-72. https://doi.org/10.1007/978-3-662-44885-4_5
5. Islam, C., Babar, M.A., Nepal, S. (2019). A Multi-Vocal Review of Security Orchestration. ACM Computing Surveys, *52(2), Article 37*. https://doi.org/10.1145/3305268
6. Jagatic, T. N., Johnson, N. A., Jakobsson, M., Menczer, F. (2007). Social phishing. *Communications of the ACM*, 50(10), 94-100. https://doi.org/10.1145/1290958.1290968

7. Javaid, A., Niyaz, Q., Sun, W., Alam, M. (2016). A deep learning approach for network intrusion detection system. *Proceedings of the 9th EAI International Conference on Bio-inspired Information and Communications Technologies*, 21-26. https://doi.org/10.4108/eai.3-12-2015.2262516
8. Kharraz, A., Robertson, W., Balzarotti, D., Bilge, L., Kirda, E. (2015). Cutting the Gordian knot: A look under the hood of ransomware attacks. *International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment*, 3-24. https://doi.org/10.1007/978-3-319-20550-2_1
9. Nguyen, T.D., Marchal, S., Miettinen, M., Fereidooni, H., Asokan, N., Sadeghi, A.-R. (2019). DÏoT: A Federated Self-learning Anomaly Detection System for IoT. *2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS*), 756–767. https://doi.org/10.1109/ICDCS.2019.00080.
10. Sahoo, K.S., Panda, S.K., Sahoo, S. et al. Toward secure software-defined networks against distributed denial of service attack. J Supercomput 75, 4829–4874 (2019). https://doi.org/10.1007/s11227-019-02767-z
11. Sommer, R., Paxson, V. (2010). Outside the closed world: On using machine learning for network intrusion detection. *IEEE Symposium on Security and Privacy*, 305-316. https://doi.org/10.1109/SP.2010.25
12. Zhang, W., Jiang, Q., Chen, L., Li, C. (2017). Two-stage ELM for phishing Web pages detection using hybrid features. *World Wide Web, 20, 797–813.* https://doi.org/10.1007/s11280-016-0418-9

Netografia

1. ENISA. (2023). *Threat Landscape 2023*. European Union Agency for Cybersecurity. Pobrane z: https://www.enisa.europa.eu/topics/threat-risk-management/threats-and-trends (dostęp: 06.11.2025).
2. European Union. (2016). *Regulation (EU) 2016/679 of the European Parliament and of the Council* (General Data Protection Regulation). Official Journal of the European Union. Pobrane z: https://eur-lex.europa.eu/eli/reg/2016/679/oj (dostęp: 06.11.2025).

Magda Konieczna[1]

Akademia Bialska im. Jana Pawła II, Biała Podlaska

https://doi.org/10.29316/9788368103304_4

SYSTEMY KOMPUTEROWEGO WSPOMAGANIA DIAGNOSTYKI OPARTE NA GŁĘBOKIM UCZENIU I TECHNIKACH WYJAŚNIANIA DECYZJI: AKTUALNE TRENDY I KIERUNKI ROZWOJU

COMPUTER-AIDED DIAGNOSTIC SYSTEMS BASED ON DEEP LEARNING AND DECISION EXPLANATION TECHNIQUES: CURRENT TRENDS AND DIRECTIONS OF DEVELOPMENT

Streszczenie

Celem rozdziału jest podsumowanie badań z lat 2022-2025 dotyczących systemów komputerowego wspomagania diagnostyki (CAD, *Computer-Aided Diagnosis*) opartych na głębokim uczeniu (DL, *Deep Learning*) z komponentem interpretowalności (XAI, *Explainable Artificial Intelligence*). Analizie poddano 76 publikacji spełniających kryteria włączenia. Dominującymi zadaniami badawczymi były klasyfikacja (63,2%), segmentacja (15,8%) i detekcja (14,5%), a w mniejszym stopniu podejścia wielozadaniowe.

[1] e-mail: m.konieczna@dyd.akademiabialska.pl

Najczęściej stosowane modalności obejmowały MRI (30,3%), RTG (28,9%) oraz CT (23,7%). W architekturach dominowały klasyczne konwolucyjne sieci neuronowe (CNN, *Convolutional Neural Network*; np. ResNet, VGG, DenseNet), rośnie jednak udział hybrydowych modeli CNN-Transformer, szczególnie w analizach trójwymiarowych (3D). Wszystkie badania uwzględniały komponent XAI. Najczęściej stosowano Grad-CAM (*Gradient-weighted Class Activation Mapping*) i jego warianty, a rzadziej SHAP (*SHapley Additive exPlanations*), LIME (*Local Interpretable Model-agnostic Explanations*) czy *Integrated Gradients*. Narzędzia te poprawiały transparentność modeli, zwiększając zaufanie klinicystów. Typowe metryki skuteczności obejmowały dokładność (Accuracy), pole pod krzywą ROC (AUC, *Area Under the Curve*), miarę F1-score oraz współczynnik podobieństwa Dice'a (Dice, *Dice Similarity Coefficient*); większość modeli osiągała wysokie wyniki (AUC > 0,90, Dice > 0,85). Główne bariery praktycznego wdrożenia to brak standaryzacji metryk, ograniczona liczba badań wieloośrodkowych oraz wysokie koszty obliczeniowe modeli trójwymiarowych (3D) i multimodalnych. Najważniejsze kierunki rozwoju obejmują integrację danych multimodalnych, uczenie federacyjne (FL, *Federated Learning*), standaryzację protokołów walidacyjnych oraz zaawansowane narzędzia XAI. Systemy CAD oparte na DL i XAI mają potencjał, by zwiększyć dokładność oraz transparentność diagnostyki obrazowej, jednak ich pełna implementacja wymaga dalszych badań i walidacji klinicznej.

Słowa kluczowe: głębokie uczenie, sztuczna inteligencja wyjaśnialna, systemy wspomagania diagnostyki, obrazowanie medyczne, modele hybrydowe

Abstract

This chapter summarises research published between 2022 and 2025 on computer-aided diagnosis (CAD) systems based on deep learning (DL) with an explainability component (XAI, Explainable Artificial Intelligence). A total of 76 publications meeting the inclusion criteria were analysed. The dominant research tasks were classification (63.2%), segmentation (15.8%), and detection (14.5%), with a smaller share of multitask approaches. The most frequently used imaging modalities included MRI (30.3%), X-ray (28.9%), and CT (23.7%). Classical convolutional neural networks (CNNs; e.g., ResNet, VGG, DenseNet) have dominated the architecture landscape, although the share of hybrid CNN-Transformer models has been increasing, particularly in three-dimensional (3D) analyses. All studies included an XAI component. The most commonly applied explainability techniques were Grad-CAM (Gradient-weighted Class Activation Mapping) and its variants, while SHAP (Shapley Additive exPlanations), LIME (Local Interpretable Model-agnostic Explanations), and Integrated Gradients were used less frequently. These tools enhanced model transparency and increased clinicians' trust. Typical performance

metrics included Accuracy, the area under the ROC curve (AUC, Area Under the Curve), F1-score, and the Dice coefficient (Dice Similarity Coefficient); most models achieved high performance (AUC > 0.90, Dice > 0.85). The main barriers to practical implementation include the lack of metric standardisation, the limited number of multicenter studies, and the high computational costs of 3D and multimodal models. The most important directions for future research include integrating multimodal data, adopting federated learning (FL), standardising validation protocols, and advancing XAI tools. CAD systems based on DL and XAI have the potential to improve both the accuracy and transparency of medical image diagnosis; however, their full implementation requires further research and clinical validation.

Keywords: deep learning, explainable artificial intelligence, diagnostic support systems, medical imaging, hybrid models

Wstęp

W ostatnich latach głębokie uczenie (ang. deep learning, DL) odegrało kluczową rolę w analizie obrazów medycznych, wspierając proces automatyzacji diagnostyki. Techniki DL umożliwiają wykrywanie, segmentację oraz klasyfikację zmian patologicznych w obrazach pochodzących z różnych modalności, takich jak radiografia (RTG), tomografia komputerowa (CT), rezonans magnetyczny (MRI), ultrasonografia (USG) oraz pozytonowa tomografia emisyjna (PET). Modele oparte na konwolucyjnych sieciach neuronowych (CNN), architekturach typu U-Net oraz nowoczesnych podejściach wykorzystujących mechanizmy uwagi – takich jak Vision Transformers czy Swin Transformer – osiągają wysoką skuteczność w zadaniach diagnostycznych i coraz częściej są integrowane z systemami wspomagania decyzji klinicznych (CAD) (Yamashita i in., 2018; Iqbal i in., 2023; Thanoon i in., 2023; Yeasmin i in., 2024).

Analiza literatury opublikowanej w latach 2022-2025 wskazuje, że modele DL charakteryzują się nie tylko wysoką trafnością diagnostyczną, ale także potencjałem w zakresie zwiększania dostępności usług medycznych oraz skracania czasu diagnostyki, co ma szczególne znaczenie w kontekście rosnącego obciążenia systemów opieki zdrowotnej. Nowoczesne systemy CAD oparte na DL wykazują wysoką skuteczność w detekcji zmian nowotworowych, złamań kostnych, ognisk

niedokrwiennych, zwłóknień oraz patologii naczyniowych – również w przypadku obrazów trudnych do jednoznacznej interpretacji (Shobayo, Saatchi, 2025; Berghout, 2025; Shelmerdine i in., 2022).

Coraz większego znaczenia nabierają również interpretowalność modeli oraz poziom zaufania klinicznego do ich decyzji. Narzędzia z zakresu wyjaśnialnej sztucznej inteligencji (XAI), takie jak Grad-CAM, SHAP czy LIME, odgrywają kluczową rolę w procesie akceptacji technologii DL przez środowiska medyczne, szczególnie w zastosowaniach o podwyższonym ryzyku klinicznym (Hou i in., 2024; Selvaraju i in., 2017; Lundberg, Lee, 2017; Ribeiro i in., 2016).
Celem niniejszego przeglądu jest syntetyczna analiza aktualnego stanu wiedzy na temat zastosowań głębokiego uczenia w analizie obrazów medycznych w latach 2022-2025. W szczególności przeanalizowano:

- nowoczesne architektury i podejścia hybrydowe w dziedzinie DL,
- zastosowania modeli DL w analizie obrazów pochodzących z różnych modalności (RTG, CT, MRI, USG, PET),
- metody oceny skuteczności modeli (m.in. AUC, mAP, IoU, współczynnik Dice),
- narzędzia interpretowalności i ich znaczenie kliniczne,
- aktualne wyzwania oraz potencjalne kierunki dalszego rozwoju technologii.

Metodologia przeglądu

Strategia wyszukiwania i źródła danych

W celu przeprowadzenia przeglądu literatury zastosowano systematyczne wyszukiwanie publikacji naukowych opublikowanych w latach 2022-2025 dotyczących zastosowań głębokiego uczenia w analizie obrazów medycznych. Szczególny nacisk położono na prace zawierające komponent diagnostyczny (CAD), segmentacyjny oraz interpretacyjny (XAI).

Wyszukiwania przeprowadzono w pięciu uznanych bazach danych: PubMed, Scopus, IEEE Xplore, ScienceDirect (Elsevier) oraz SpringerLink. Użyto kombinacji słów kluczowych w języku angielskim, takich jak: *deep learning, medical imaging, computer-aided diagnosis, CNN, XAI, CT, MRI, USG, PET, CAD, segmentation, classification, explainability*. Zakres wyszukiwania ograniczono do

publikacji w języku angielskim, opublikowanych w okresie od 1 stycznia 2022 do 1 lipca 2025. W celu zawężenia wyników zastosowano operatory logiczne AND oraz OR, umożliwiające precyzyjne dopasowanie do założonych kryteriów tematycznych.

Kryteria selekcji publikacji

Do analizy włączono wyłącznie publikacje spełniające następujące warunki:

- artykuły oryginalne (empiryczne) zawierające wyniki badań z wykorzystaniem technik głębokiego uczenia,
- zastosowanie modeli DL w analizie obrazów pochodzących z jednej lub więcej modalności (RTG, CT, MRI, USG, PET),
- obecność komponentu diagnostycznego, klasyfikacyjnego, segmentacyjnego lub wspomagania decyzji klinicznych (CAD),
- uwzględnienie narzędzi interpretowalności (XAI) bądź ich analiza,
- pełny tekst publikacji dostępny w języku angielskim.

Z analizy wykluczono prace spełniające następujące kryteria:

- artykuły przeglądowe, komentarze redakcyjne oraz abstrakty bez dostępu do pełnej treści,
- badania koncentrujące się wyłącznie na danych niemedycznych (np. tekstowych, genomowych),
- publikacje nieuwzględniające zastosowań głębokiego uczenia,
- prace bez powiązania z obrazowaniem medycznym lub diagnostyką kliniczną.

Proces selekcji obejmował niezależną ocenę tytułów i streszczeń, a następnie analizę pełnych tekstów. Publikacje spełniające wszystkie kryteria włączenia zostały uwzględnione w analizie końcowej. Ze względów objętościowych bibliografia końcowa ma charakter reprezentatywny i obejmuje wyłącznie pozycje kluczowe dla głównych nurtów badań oraz metod XAI; pełny korpus 76 publikacji stanowił podstawę analizy syntetycznej autora.

Wyniki przeglądu

Charakterystyka zbioru badań: typy zadań, modalności obrazowania i metody interpretowalności

Analiza 76 publikacji zakwalifikowanych do przeglądu (zgodnie z opisanymi kryteriami selekcji) wykazała, że dominującym kierunkiem badań nad systemami CAD opartymi na głębokim uczeniu pozostaje klasyfikacja obrazów medycznych, 63,2% (48/76). Znacznie rzadziej pojawiają się badania ukierunkowane wyłącznie na segmentację, 15,8% (12/76), lub detekcję, 14,5% (11/76). Zauważalny jest również udział projektów wielozadaniowych, łączących klasyfikację z segmentacją, 3,9% (3/76), lub klasyfikację z detekcją, 2,6% (2/76). Struktura ta sugeruje wyraźny trend rozwoju metod integrujących wiele zadań, co pozostaje zgodne z wnioskami nowszych przeglądów dotyczących CAD i analizy obrazów medycznych (Iqbal i in., 2023; Yeasmin i in., 2024).

Pod względem stosowanych modalności obrazowania największy udział mają badania oparte na rezonansie magnetycznym, 30,3% (23/76), oraz radiografii, 28,9% (22/76). Tomografia komputerowa występuje w 23,7% (18/76) prac, natomiast ultrasonografia w 6,6% (5/76). Pozostałe publikacje wykorzystują dane multimodalne (np. połączenie MRI z CT lub PET), jednak ich liczba pozostaje ograniczona, 3,9% (3/76). Struktura ta odzwierciedla utrwaloną rolę MRI w analizach wymagających wysokiej rozdzielczości i danych trójwymiarowych oraz popularność RTG w diagnostyce chorób płuc i urazów kostnych.

Architektury głębokiego uczenia i podejścia 2D/3D

Analiza 76 publikacji wykazała dominację klasycznych konwolucyjnych sieci neuronowych (CNN), spośród których najczęściej stosowaną pozostaje rodzina ResNet, 44,7% (34/76). Znaczący udział odnotowano także dla architektur VGG, 19,7% (15/76), i DenseNet, 18,4% (14/76). Architektura EfficientNet pojawiła się w 13,2% (10/76) prac, natomiast rzadziej stosowane były sieci Inception/Inc-ResNet, MobileNet i Xception, po 6,6% (5/76). Wskazuje to na stopniowe przechodzenie od prostych konwolucji do bardziej zoptymalizowanych modeli sieci szkieletowych.

Obserwuje się także rosnące zainteresowanie architekturami opartymi na mechanizmach uwagi i transformatorach, które wystąpiły w 10,5% (8/76) badań. W wielu przypadkach transformatory nie były stosowane samodzielnie, lecz w konfiguracjach hybrydowych CNN+Transformer, stanowiących 59,2% (45/76) analizowanych rozwiązań. Trend ten wskazuje na potrzebę integracji tradycyjnych sieci konwolucyjnych z mechanizmami sekwencyjnymi i mechanizmami uwagi. Architektury z rodziny U-Net, w tym warianty z mechanizmami uwagi i w wersji trójwymiarowej, były stosowane w 7,9% (6/76) badań i dominowały w zadaniach segmentacyjnych – od złamań miednicy w CT, przez segmentację guzów wątroby, aż po zmiany w stwardnieniu rozsianym. Klasyczne ramy detekcyjne, takie jak Faster R-CNN, pojawiły się jedynie w 1,3% (1/76) publikacji, co wskazuje na ich ograniczone zastosowanie w badaniach z ostatnich lat.

Z perspektywy wymiaru modelowania zdecydowanie dominują podejścia 2D, 92,1% (70/76), podczas gdy architektury 3D (np. 3D CNN, 3D U-Net) wystąpiły w 7,9% (6/76), głównie w kontekście obrazów CT i MRI. Wysokie koszty obliczeniowe oraz wymagania dotyczące danych ograniczają ich szersze wykorzystanie.

Mechanizmy uwagi i wyjaśnialna sztuczna inteligencja

Wszystkie publikacje zakwalifikowane do niniejszego przeglądu zawierały komponent wyjaśnialnej sztucznej inteligencji (XAI). Najczęściej stosowaną grupę metod stanowiły techniki wizualizacji map aktywacji, w tym Grad-CAM i jego warianty, które dominowały w 65,8% (50/76) badań. Popularność tych metod wynika z ich intuicyjności oraz łatwości implementacji, co znajduje odzwierciedlenie w ich dominacji w analizowanym korpusie (Selvaraju i in., 2017; Hou i in., 2024; Tabela 1).

Metody model-agnostic, takie jak SHAP i LIME, stosowano w 13,2% (10/76) publikacji. SHAP był używany m.in. w diagnostyce chorób neurodegeneracyjnych czy mammografii, natomiast LIME w dermoskopii, histopatologii i diagnostyce chorób płuc. Choć bardziej złożone obliczeniowo, metody te umożliwiają szczegółową analizę istotności cech (Lundberg, Lee, 2017; Ribeiro i in., 2016).

Mechanizmy uwagi integrowano z architekturami segmentacyjnymi (np. U-Net z blokami uwagi) oraz z modelami hybrydowymi CNN-Transformer,

występując w 10,5% (8/76) badań. Rozwiązania te nie tylko poprawiały jakość ekstrakcji cech, lecz także wspierały interpretację kliniczną. Rzadziej stosowano metody alternatywne, takie jak Integrated Gradients czy SmoothGrad, (5,3% (4/76)) oraz Guided Backpropagation. Oferują one wyższą rozdzielczość interpretacji, ale wiążą się z większą złożonością obliczeniową (Sundararajan i in., 2017).

Tabela 1. Częstość stosowania metod XAI w analizowanych publikacjach (n=76)

Metoda XAI	Liczba publikacji (n=76)	Procent (%)
Grad-CAM i warianty	50	65,8
SHAP	6	7,9
LIME	4	5,3
Mechanizmy uwagi	8	10,5
Integrated Gradients / SmoothGrad	4	5,3
Guided Backpropagation	2	2,6

Źródło: opracowanie własne.

Zastosowania kliniczne

Analiza 76 publikacji wykazała, że systemy CAD oparte na głębokim uczeniu znajdują szerokie zastosowanie w różnych modalnościach obrazowania medycznego. Najczęściej wykorzystywane są radiografia (35,5%, 27/76) oraz rezonans magnetyczny (26,3%, 20/76). Tomografia komputerowa obejmuje 21,1% (16/76) prac, natomiast ultrasonografia – 14,5% (11/76). Publikacje multimodalne, łączące MRI z CT lub PET, stanowią jedynie niewielki odsetek – po 1,3% (1/76 każda). Każda z modalności charakteryzuje się odmiennym profilem zastosowań klinicznych, doborem architektur sieciowych oraz zakresem stosowanych metryk skuteczności (Tabela 2). Radiografia (RTG) stanowi jedną z najczęściej analizowanych modalności, szczególnie w diagnostyce chorób płuc, gruźlicy oraz złamań kostnych. W badaniach z lat 2022-2025 dominują modele oparte na konwolucyjnych sieciach neuronowych (ResNet, EfficientNet) oraz na architekturach typu Vision Transformer (ViT), które osiągają wysoką skuteczność klasyfikacyjną (AUC 0,92-0,96, ACC 92-96%).

Tabela 2. Zastosowania systemów CAD z komponentem XAI w różnych modalnościach obrazowania medycznego (2022-2025)

Modalność	Liczba badań (%)	Najczęstsze zadania	Typowe architektury	Zakres metryk
RTG	27 (35,5%)	Klasyfikacja, detekcja (gruźlica, zapalenie płuc, złamania)	CNN (ResNet, EfficientNet), ViT	AUC 0,92-0,96; ACC 92-96%; SEN >90%
MRI	20 (26,3%)	Klasyfikacja, segmentacja (guzy, choroby neurodegeneracyjne)	3D CNN, U-Net, Transformers	AUC 0,95-0,98; ACC >93%; Dice >0,85
CT	16 (21,1%)	Detekcja, segmentacja (złamania, krwotoki, guzy)	3D U-Net, CNN–Transformer, YOLO	SEN 88-95%; IoU >0,80; AUC 0,94-0,97
USG	11 (14,5%)	Klasyfikacja, segmentacja (rak piersi, tarczyca, obrazowanie prenatalne)	CNN, U-Net	AUC 0,90-0,95; ACC 89-94%; Dice >0,82
MRI + CT	1 (1,3%)	Analizy multimodalne (guzy, segmentacja)	Hybrydowe CNN+Transformer	AUC ~0,95; Dice >0,83
MRI + PET	1 (1,3%)	Klasyfikacja i segmentacja (nowotwory, przerzuty)	Hybrydowe (Transformer + CNN)	AUC 0,96-0,98; Dice 0,80-0,86

Źródło: opracowanie własne.

Rezonans magnetyczny (MRI) jest najczęściej wykorzystywaną modalnością w badaniach nad CAD, zwłaszcza w diagnostyce nowotworów mózgu i chorób neurodegeneracyjnych. Zastosowanie znajdują tu głównie sieci 3D CNN, architektury U-Net oraz różne warianty Transformerów, które uzyskują wyniki o wysokiej precyzji (AUC 0,95-0,98; Dice > 0,85).

Tomografia komputerowa (CT) pozostaje podstawowym narzędziem w ocenie urazów czaszkowo-mózgowych, złamań i zmian nowotworowych. Architektury 3D (U-Net, CNN-Transformer, YOLO) osiągają czułość na poziomie 88-95% oraz wysokie wartości IoU (>0,80), co potwierdza ich użyteczność w zadaniach detekcji i segmentacji (Shan i in., 2021).

Ultrasonografia (USG) odgrywa rosnącą rolę w diagnostyce raka piersi, chorób tarczycy oraz w obrazowaniu prenatalnym. Modele oparte na CNN i U-Net zapewniają dokładność w granicach 89-94% oraz współczynniki Dice przekraczające 0,82, co potwierdza potencjał tej modalności mimo ograniczonej liczby badań.

Publikacje multimodalne, choć rzadkie (po 1,3% dla MRI+CT i MRI+PET), wskazują na rosnące zainteresowanie integracją danych obrazowych z różnych źródeł. Modele hybrydowe, zwłaszcza konfiguracje CNN+Transformer, uzyskują w nich bardzo wysoką skuteczność (AUC >0,95; Dice >0,83), co podkreśla znaczenie fuzji danych w złożonych zadaniach diagnostycznych (Lipkova i in., 2022; Li i in., 2024).

Podsumowując, skuteczność systemów CAD różni się w zależności od modalności i rodzaju zadania. RTG i CT dominują w badaniach nad złamaniami oraz chorobami płuc, MRI w diagnostyce nowotworów i chorób ośrodkowego układu nerwowego, USG w diagnostyce onkologicznej i prenatalnej, natomiast PET w ocenie nowotworów i przerzutów. Wspólnym trendem jest rozwój modeli hybrydowych i multimodalnych, które umożliwiają poprawę dokładności diagnostycznej oraz zwiększenie interpretowalności decyzji klinicznych.

Metryki oceny i skuteczność modeli

Analiza 76 publikacji wskazuje, że w badaniach nad systemami CAD opartymi na głębokim uczeniu stosowano szerokie spektrum metryk oceny skuteczności. Najczęściej raportowaną miarą była dokładność (Accuracy), obecną w ponad 60% prac, przy czym typowe wartości mieściły się w zakresie 90-99%, co potwierdza wysoką skuteczność modeli w zadaniach klasyfikacyjnych, zwłaszcza dla obrazów RTG i MRI.

AUC (Area Under the ROC Curve) stosowano w około połowie publikacji, głównie w zadaniach klinicznie istotnych, takich jak diagnostyka krwotoków wewnątrzczaszkowych, nowotworów mózgu czy choroby Alzheimera. Wartości AUC często przekraczały 0,90, a w kilku przypadkach osiągały 0,99-1,00, co wskazuje na wyjątkową odporność modeli na zmiany progu decyzyjnego.

Precision, Recall (czułość i swoistość) były wykorzystywane w około 40% badań, zwłaszcza w kontekście detekcji i klasyfikacji zmian patologicznych.

W połączeniu z F1-score (ok. 35% prac) zapewniały bardziej zrównoważoną ocenę modeli w warunkach niezbalansowanych zbiorów danych, np. w dermoskopii, histopatologii i detekcji przerzutów nowotworowych.

Współczynnik Dice i IoU (Jaccard index) dominowały w badaniach segmentacyjnych (ok. 20% i 10% prac). Typowe wartości Dice mieściły się w zakresie 0,85-0,91, a IoU przekraczało 0,80, co wskazuje na wysoką dokładność modeli przy delineacji struktur anatomicznych w CT i MRI.

Rzadziej stosowane były miary dodatkowe, takie jak Cohen's Kappa, MCC, NPV czy mAP (szczególnie w detekcji obiektów, np. złamań). Choć pojawiały się sporadycznie, pozwalały na dodatkowe ujęcie jakości modeli w kontekście wiarygodności diagnostycznej i równowagi klas.

Podsumowując, badania z lat 2022-2025 charakteryzują się wysokimi wartościami kluczowych wskaźników (AUC > 0,90; Dice > 0,85; Accuracy > 90%). Wskazuje to na duży potencjał systemów CAD opartych na głębokim uczeniu w praktyce klinicznej. Jednocześnie widoczny jest brak pełnej standaryzacji raportowania metryk, co utrudnia bezpośrednie porównywanie wyników między badaniami i instytucjami. Standaryzacja w tym zakresie stanowi istotny kierunek dalszych prac.

Interpretowalność i zaufanie kliniczne

Jednym z kluczowych czynników warunkujących praktyczne wdrożenie systemów CAD do diagnostyki klinicznej jest poziom ich interpretowalności oraz związane z nim zaufanie użytkowników końcowych – lekarzy i radiologów. Mimo wysokiej skuteczności modeli głębokiego uczenia w zadaniach klasyfikacyjnych, detekcyjnych i segmentacyjnych, wciąż istnieje obawa przed ich wykorzystaniem jako „czarnych skrzynek", co rodzi wątpliwości dotyczące bezpieczeństwa i odpowiedzialności klinicznej.

W badaniach przeprowadzonych w ostatnich latach podkreśla się, że wizualizacje generowane przez modele (np. Grad-CAM) mogą zwiększać akceptację lekarzy tylko wtedy, gdy wskazywane obszary są zgodne z wiedzą ekspercką i rzeczywistymi lokalizacjami zmian patologicznych. Jednak liczne prace zwracają uwagę na problem nadmiernego uproszczenia – mapy uwagi bywają zbyt ogólne i nie zawsze odzwierciedlają proces decyzyjny modelu.

Dla praktyki klinicznej istotne jest, aby narzędzia XAI były nie tylko poprawne technicznie, ale również użyteczne diagnostycznie. W analizowanych publikacjach wskazywano, że akceptacja modeli rośnie, gdy interpretacje są prezentowane w sposób zbliżony do rutynowej praktyki (np. nakładanie map uwagi na obrazy RTG lub CT, możliwość interaktywnej analizy istotności cech). Z kolei w zadaniach związanych z chorobami rzadkimi lub nowotworami, gdzie stawką są decyzje o dużym znaczeniu klinicznym, często proponuje się stosowanie metod takich jak SHAP czy LIME, mimo ich złożoności obliczeniowej, ze względu na możliwość bardziej precyzyjnej interpretacji (Hou i in., 2024; Lundberg, Lee, 2017; Ribeiro i in., 2016).

Kwestia zaufania klinicznego ma również istotny wymiar regulacyjny i etyczny, ponieważ we wdrożeniach AI w medycynie rośnie znaczenie transparentności, interpretowalności oraz możliwości audytu modeli medycznych. Oznacza to, że brak interpretowalności może stanowić barierę nie tylko psychologiczną dla lekarzy, ale również organizacyjną i wdrożeniową dla producentów systemów CAD.

Podsumowując, lata 2022-2025 przynoszą wyraźne przesunięcie w stronę badań, które akcentują nie tylko skuteczność diagnostyczną, ale także kliniczną użyteczność i przejrzystość modeli. Implementacja narzędzi XAI – od prostych map aktywacji po zaawansowane metody analizy cech – staje się istotnym warunkiem akceptacji przez użytkowników końcowych i jednym z kluczowych wymiarów jakości współczesnych systemów CAD (Hou i in., 2024).

Dyskusja

Analiza zgromadzonej literatury wskazuje, że dominującą rolę w obszarze interpretowalności systemów CAD opartych na głębokim uczeniu odgrywają techniki wizualizacji map aktywacji, w szczególności Grad-CAM oraz jego warianty. Popularność tych metod wynika z ich intuicyjności i łatwości wizualizacji, które pozwalają klinicystom na szybkie zrozumienie podstaw działania modelu. Niemniej jednak ograniczona zdolność Grad-CAM do szczegółowej i ilościowej analizy cech obrazu istotnie obniża jego przydatność w zastosowaniach wymagających precyzyjnej interpretacji lokalnych struktur anatomicznych.

Równolegle zauważalny jest stopniowy wzrost zainteresowania metodami niezależnymi od architektury modelu (model-agnostic), takimi jak SHAP i LIME, które pozwalają na bardziej szczegółową analizę wpływu poszczególnych cech na decyzję klasyfikacyjną. Przykłady obejmują m.in. badania nad klasyfikacją zmian neurodegeneracyjnych, mammografią oraz cytologią szyjki macicy. W dermatologii i histopatologii metody te wspierają transparentność modeli analizujących obrazy dermoskopowe i preparaty tkankowe. Ich główną wadą pozostają jednak wysokie koszty obliczeniowe i brak dostosowania do pracy w czasie rzeczywistym, co ogranicza ich użyteczność kliniczną.

Szczególne znaczenie zyskują mechanizmy uwagi (attention mechanisms), które stanowią kluczowe ogniwo łączące wzrost skuteczności modeli z poprawą ich interpretowalności. Architektury oparte na Transformerach oraz hybrydowych sieciach CNN-Transformer wykazują wysoki potencjał nie tylko w zakresie zwiększania jakości predykcji, lecz także w generowaniu map uwagi odzwierciedlających obszary istotne z punktu widzenia ekspertów klinicznych.

Pomimo dynamicznego rozwoju badań w zakresie interpretowalności, wciąż niewystarczająco rozwinięta pozostaje problematyka dotycząca modeli 3D oraz podejść multimodalnych (np. CT łączone z danymi klinicznymi, MRI z PET). Rozwiązania te lepiej odzwierciedlają złożoność struktur anatomicznych i procesów patologicznych, a ich wyniki eksperymentalne są obiecujące. Brak systematycznych analiz i walidacji klinicznej w tym obszarze stanowi jednak istotną barierę dla translacji tych modeli do praktyki klinicznej.

Podsumowując, w latach 2022-2025 obserwuje się przesunięcie akcentów badawczych z prostych technik wizualizacji na bardziej zaawansowane metody interpretowalności i modele hybrydowe. Równocześnie utrzymują się istotne ograniczenia badawcze, szczególnie w zakresie zastosowań multimodalnych, modeli 3D oraz standaryzacji metod oceny interpretowalności – które wyznaczają kierunki dalszych badań nad klinicznym wykorzystaniem systemów CAD.

Wyzwania i perspektywy rozwoju

Pomimo dynamicznego rozwoju systemów CAD opartych na głębokim uczeniu i metodach XAI, wciąż istnieje szereg barier ograniczających ich pełne wdrożenie kliniczne. Kluczowe problemy obejmują brak standaryzacji metryk

i protokołów walidacyjnych, ograniczoną dostępność danych wieloośrodkowych, wysokie koszty obliczeniowe modeli 3D oraz trudności z przenoszalnością wyników do realnych warunków diagnostycznych.

Brak jednolitych standardów oceny (np. AUC, Dice, IoU) utrudnia porównywalność wyników i budowę benchmarków, natomiast zróżnicowanie populacyjne i sprzętowe ogranicza uogólnialność opracowanych rozwiązań. Modele 3D i multimodalne, mimo wysokiej dokładności i bogatego kontekstu diagnostycznego, wymagają znacznych zasobów obliczeniowych i infrastrukturalnych, co utrudnia ich wdrożenie w środowiskach klinicznych o ograniczonych możliwościach technicznych.

Dodatkową barierą pozostaje luka pomiędzy skutecznością eksperymentalną a realną implementacją kliniczną – modele osiągające wysokie wyniki w badaniach kontrolowanych często tracą wydajność przy zastosowaniu na nowych danych. W tym kontekście szczególnego znaczenia nabiera potrzeba adaptacyjnych systemów DL, zdolnych do ciągłego uczenia się i walidacji w różnych populacjach pacjentów.

W odpowiedzi na powyższe ograniczenia, kierunki dalszego rozwoju koncentrują się na kilku kluczowych obszarach:

1. Zaawansowane architektury hybrydowe – integrujące CNN z mechanizmami uwagi i Transformerami, szczególnie w analizie obrazów 3D i danych o wysokiej zmienności klinicznej.
2. Integracja multimodalnych źródeł danych – łączenie obrazów CT, MRI, PET i USG z danymi klinicznymi oraz raportami tekstowymi, co pozwala na bardziej kompleksową analizę przypadków.
3. Uczenie federacyjne i prywatność danych – rozwój rozproszonych metod trenowania, umożliwiających współpracę wielu ośrodków bez wymiany danych pacjentów, przy jednoczesnym zwiększeniu ogólności modeli.
4. Personalizacja i modele adaptacyjne – dostosowanie predykcji do indywidualnych cech pacjenta, co może zwiększyć trafność diagnostyczną i ograniczyć liczbę wyników fałszywie dodatnich.
5. Real-time CAD i wdrożenia kliniczne – optymalizacja architektur i infrastruktury obliczeniowej, umożliwiająca analizę obrazów w czasie rzeczywistym, zwłaszcza w diagnostyce urazowej i zabiegowej.

6. Ewolucja narzędzi XAI – przejście od intuicyjnych metod (np. Grad-CAM) ku bardziej formalnym i ilościowym podejściom, takim jak SHAP, LIME czy Integrated Gradients, które mogą zwiększyć zaufanie kliniczne poprzez spójną interpretację wyników.

Wnioski

Analiza 76 publikacji z lat 2022-2025 potwierdza, że systemy CAD oparte na głębokim uczeniu i interpretowalności stanowią jeden z najbardziej dynamicznie rozwijających się kierunków współczesnej diagnostyki obrazowej.

Modele głębokiego uczenia osiągają wysoką skuteczność w klasyfikacji, detekcji i segmentacji obrazów medycznych, często przewyższając tradycyjne metody analizy. Szczególnie obiecujące są architektury hybrydowe oraz modele 3D, które lepiej odzwierciedlają złożoność danych klinicznych i umożliwiają bardziej precyzyjną diagnostykę.

Interpretowalność modeli stanowi warunek akceptacji klinicznej – dominują metody oparte na wizualizacji map aktywacji (Grad-CAM i warianty), które, mimo intuicyjności, wymagają uzupełnienia o bardziej formalne narzędzia (SHAP, LIME, Integrated Gradients).

Kierunki dalszego rozwoju obejmują integrację multimodalnych źródeł danych, wdrażanie uczenia federacyjnego oraz rozwój precyzyjnych metod interpretacji i systematycznej walidacji klinicznej, które stanowią warunek translacji wyników badań do praktyki diagnostycznej.

Podsumowując, systemy CAD oparte na głębokim uczeniu i metodach XAI mają potencjał, aby znacząco poprawić dokładność, przejrzystość i efektywność diagnostyki obrazowej. Kluczowym wyzwaniem pozostaje jednak osiągnięcie równowagi między skutecznością modeli, ich interpretowalnością i zaufaniem klinicznym, co stanowi fundament ich przyszłej implementacji w opiece zdrowotnej.

Literatura

1. Berghout, T. (2025). The Neural Frontier of Future Medical Imaging: A Review of Deep Learning for Brain Tumour Detection. *Journal of Imaging, 11*(1), 2. https://doi.org/10.3390/jimaging11010002

2. Hou, J., Liu, S., Bie, Y., Wang, H., Tan, A., Luo, L., Chen, H. (2024). *Self-eXplainable AI for Medical Image Analysis: A Survey and New Outlooks*. arXiv. https://doi.org/10.48550/arXiv.2410.02331
3. Iqbal, S., Qureshi, A.N., Li, J., Mahmood, T. (2023). On the Analyses of Medical Images Using Traditional Machine Learning Techniques and Convolutional Neural Networks. *Archives of Computational Methods in Engineering, 30*(5), 3173-3233. https://doi.org/10.1007/s11831-023-09899-9
4. Li, Y., Daho, M.E.H., Conze, P.-H., Zeghlache, R., Le Boité, H., Tadayoni, R., Cochener, B., Lamard, M., Quellec, G. (2024). A Review of Deep Learning-Based Information Fusion Techniques for Multimodal Medical Image Classification. *Computers in Biology and Medicine, 177,* 108635. https://doi.org/10.1016/j.compbiomed.2024.108635
5. Lipkova, J., Chen, R.J., Chen, B., Lu, M.Y., Barbieri, M., Shao, D., Vaidya, A.J., Chen, C., Zhuang, L., Williamson, D.F.K., Shaban, M., Chen, T.Y., Mahmood, F. (2022). Artificial Intelligence for Multimodal Data Integration in Oncology. *Cancer Cell, 40*(10), 1095-1110. https://doi.org/10.1016/j.ccell.2022.09.012
6. Lundberg, S.M., Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. W: *Advances in Neural Information Processing Systems 30,* 4765-4774. https://dl.acm.org/doi/10.5555/3295222.3295230
7. Ribeiro, M.T., Singh, S., Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. W: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* 1135-1144. https://doi.org/10.1145/2939672.2939778
8. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D. (2017). Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. W: *Proceedings of the IEEE International Conference on Computer Vision (ICCV),* 618-626. https://doi.org/10.1109/ICCV.2017.74
9. Shelmerdine, S.C., White, R.D., Liu, H., Arthurs, O.J., Sebire, N.J. (2022). Artificial Intelligence for Radiological Paediatric Fracture Assessment: A Systematic Review. *Insights into Imaging, 13,* 94. https://doi.org/10.1186/s13244-022-01234-3
10. Shan, W., Guo, J., Mao, X., Zhang, Y., Huang, Y., Wang, S., Li, Z., Meng, X., Zhang, P., Wu, Z., Wang, Q., Liu, Y., He, K., Wang, Y. (2021). Automated Identification of Skull Fractures With Deep Learning: A Comparison Between Object Detection and Segmentation Approach. *Frontiers in Neurology, 12,* 687931. https://doi.org/10.3389/fneur.2021.687931

11. Shobayo, O., Saatchi, R. (2025). Developments in Deep Learning Artificial Neural Network Techniques for Medical Image Analysis and Interpretation. *Diagnostics, 15*(9), 1072. https://doi.org/10.3390/diagnostics15091072
12. Sundararajan, M., Taly, A., Yan, Q. (2017). Axiomatic Attribution for Deep Networks. W: *Proceedings of the 34th International Conference on Machine Learning, 70,* 3319-3328. https://dl.acm.org/doi/10.5555/3305890.3306024
13. Thanoon, M.A., Zulkifley, M.A., Mohd Zainuri, M.A.A., Abdani, S.R. (2023). A Review of Deep Learning Techniques for Lung Cancer Screening and Diagnosis Based on CT Images. *Diagnostics, 13*(16), 2617. https://doi.org/10.3390/diagnostics13162617
14. Yamashita, R., Nishio, M., Do, R.K.G., Togashi, K. (2018). Convolutional Neural Networks: An Overview and Application in Radiology. *Insights into Imaging, 9*(4), 611-629. https://doi.org/10.1007/s13244-018-0639-9
15. Yeasmin, M.N., Al Amin, M., Joti, T.J., Aung, Z., Azim, M.A. (2024). Advances of AI in Image-Based Computer-Aided Diagnosis: A Review. *Array, 23,* 100357. https://doi.org/10.1016/j.array.2024.100357

Michał Świątczak[1], Tomasz Grudniewski[2]

Akademia Bialska im. Jana Pawła II, Biała Podlaska

https://doi.org/10.29316/9788368103304_5

AUTOMATYCZNA KLASYFIKACJA OBRAZÓW MIKROSKOPOWYCH OKRZEMEK Z WYKORZYSTANIEM BIBLIOTEK PYTHONA I TRANSFER LEARNINGU: BADANIE PORÓWNAWCZE EFFICIENTNETB0, RESNET50 I VGG16

AUTOMATIC CLASSIFICATION OF MICROSCOPIC IMAGES OF DIATOMS USING PYTHON LIBRARIES AND TRANSFER LEARNING: A COMPARATIVE STUDY OF EFFICIENTNETB0, RESNET50, AND VGG16

Streszczenie

W rozdziale przedstawiono zastosowanie głębokich konwolucyjnych sieci neuronowych do binarnej klasyfikacji obrazów mikroskopowych na dwie klasy: „okrzemki" oraz „inne obiekty". Zbudowano i porównano trzy modele oparte na architekturach EfficientNetB0,

[1] student, e-mail: swiatczak21536@stud.akademiabialska.pl

[2] https://orcid.org/0000-0003-3394-8992, e-mail: gisbourne2@gmail.com

ResNet50 oraz VGG16, wykorzystując podejście transfer learning z wagami wstępnie wyuczonymi na zbiorze ImageNet. Do badań użyto autorskiego zbioru 3236 obrazów mikroskopowych w jasnym polu, równomiernie podzielonych na klasy diatoms i non_diatoms, z zastosowaniem normalizacji, dedykowanego preprocessingu oraz augmentacji danych. Analizowano również wpływ ważenia klas (parametr class_weight) na skuteczność klasyfikacji, ze szczególnym uwzględnieniem trudniejszej i bardziej zróżnicowanej klasy non_diatoms. Najlepsze wyniki uzyskano dla modeli ResNet50 i VGG16, które przy braku lub niewielkim ważeniu klas osiągnęły 100% dokładności i F1-score równy 1,0 na zbiorze testowym, natomiast EfficientNetB0 – przy odpowiednim doborze class_weight – uzyskał dokładność rzędu 89%. Omówiono zalety i ograniczenia poszczególnych architektur w kontekście wymagań obliczeniowych oraz potencjalnych zastosowań w monitoringu środowiskowym wód. Zaprezentowane wyniki potwierdzają, że nowoczesne modele CNN mogą istotnie przyspieszyć i zautomatyzować wykrywanie okrzemek, stanowiąc użyteczne narzędzie wspomagające analizy biologiczne.

Słowa kluczowe: okrzemki, klasyfikacja obrazów mikroskopowych, głębokie sieci konwolucyjne

Abstract

This chapter presents the use of deep convolutional neural networks for binary classification of microscopic images into two classes: “diatoms” and “other objects”. Three models based on the EfficientNetB0, ResNet50 and VGG16 architectures were built and compared using a transfer learning approach with weights pre-trained on the ImageNet dataset. The study employed a custom bright-field microscopic dataset of 3,236 images, evenly split into diatom and non-diatom classes, with appropriate normalisation, architecture-specific preprocessing, and data augmentation. The impact of class weighting (class_weight parameter) on classification performance was analysed, with particular attention to the more heterogeneous and complex non_diatoms class. The best results were obtained for ResNet50 and VGG16, which, with no or only mild class weighting, achieved 100% accuracy and an F1-score of 1.0 on the test set. In contrast, EfficientNetB0, with a suitably tuned class weight, achieved an accuracy of approximately 89%. The advantages and limitations of each architecture are discussed in terms of computational cost and potential applications in environmental water monitoring. The results demonstrate that modern CNN models can substantially accelerate and automate diatom detection, providing a valuable tool to support biological and ecological analyses.

Keywords: diatoms, microscopic image classification, deep convolutional neural networks

Wstęp

Okrzemki (Bacillariophyta) to jednokomórkowe glony otoczone charakterystycznym krzemionkowym pancerzykiem (frustulą), powszechnie występujące w wodach słodkich i morskich. Ze względu na swoją wrażliwość na zmiany środowiskowe i różnorodność gatunkową stanowią cenione bioindykatory jakości wód – skład gatunkowy i liczebność okrzemek są wykorzystywane do oceny stanu ekologicznego zbiorników wodnych oraz monitorowania zanieczyszczeń (Kireta i in. 2012). Tradycyjna identyfikacja okrzemek odbywa się pod mikroskopem optycznym i wymaga wysokich kompetencji taksonomicznych oraz doświadczenia. Proces ten jest jednak czasochłonny i podatny na subiektywne błędy ludzkie. Zapotrzebowanie na usprawnienie tych analiz zrodziło liczne próby automatyzacji rozpoznawania okrzemek już na początku XXI wieku – przykładowo w ramach projektu ADIAC opracowano systemy komputerowej analizy obrazów okrzemek, by przyspieszyć ich taksonomiczną klasyfikację i uniezależnić ją od ekspertów (Du, Bayer, 2002).

Ostatnie lata przyniosły rewolucyjny postęp w dziedzinie rozpoznawania obrazów dzięki głębokim sieciom neuronowym typu konwolucyjnego (CNN). Przełomowa praca Krizhevsky'ego i współpracowników pokazała, że głębokie CNN mogą znacznie przewyższyć klasyczne metody w zadaniach klasyfikacji obrazów (np. rozpoznawanie obiektów w zbiorze ImageNet) (Krizhevsky i in. 2017), co zapoczątkowało gwałtowny rozwój deep learningu w widzeniu komputerowym (LeCun i in., 2015). W ślad za sukcesami w informatyce, techniki głębokich sieci zaczęto stosować także w naukach biologicznych i środowiskowych do automatycznej identyfikacji organizmów na podstawie obrazów mikroskopowych. Klasyfikacja okrzemek również stała się przedmiotem licznych badań: Pedraza i in. wykorzystali konwolucyjne sieci neuronowe do rozpoznawania gatunków okrzemek, demonstrując przewagę uczenia głębokiego nad podejściami opartymi na cechach ręcznych (Pedraza i in., 2017). Kloster i in. opracowali zautomatyzowany system taksonomiczny oparty na CNN, uzyskując F1-score ok. 97% dla rozpoznawania zróżnicowanych morfologicznie okrzemek ze skanów mikroskopowych, przy czym zauważyli oni, że nawet starsze architektury (np. VGG16) mogą osiągać najwyższą skuteczność w tym zadaniu (Kloster i in., 2020). Z kolei Pu i współpracownicy zaproponowali podejście z wykorzystaniem

transfer learningu oraz augmentacji danych, poprawiając efektywność klasyfikacji przy ograniczonej liczbie obrazów okrzemek (Pu i in., 2023). Najnowsze prace skupiają się również na optymalizacji złożoności modeli – Gunduz i Gunal przedstawili lekki model DiatomNet, oparty na uproszczonej sieci konwolucyjnej, który osiąga wysoką trafność przy niewielkiej liczbie parametrów, co jest istotne z perspektywy wdrożeń na urządzenia o ograniczonej mocy obliczeniowej (Gunduz, Gunal, 2024).

Pomimo obiecujących wyników tych badań, wiele z nich dotyczy klasyfikacji wieloklasowej (do poziomu rodzaju lub gatunku okrzemek) i wymaga obszernych zbiorów danych treningowych. W praktyce monitoringu środowiskowego przydatne może być również prostsze zadanie binarnej klasyfikacji: odróżnienie obrazów zawierających okrzemki od obrazów przedstawiających inne obiekty (np. fragmenty osadu, inne glony lub artefakty). Taki system mógłby służyć do szybkiego przesiewu próbek mikroskopowych, automatycznie wskazując obrazy potencjalnie zawierające okrzemki do dalszej analizy eksperckiej. W literaturze podkreśla się, że transfer learning – wykorzystanie sieci wstępnie nauczonych na dużych zbiorach (np. ImageNet) – istotnie ułatwia trenowanie modeli na małych, specjalistycznych zbiorach danych, pozwalając zaadaptować uniwersalne cechy wizualne do konkretnego problemu (Zhuang i in., 2021). Również techniki radzenia sobie z niezrównoważeniem danych (np. dysproporcją klas) odgrywają kluczową rolę, gdyż modele głębokie mają tendencję do faworyzowania klasy dominującej; rozwiązaniem jest m.in. ważenie klas podczas trenowania (class weighting) (Buda i in., 2018).

Celem niniejszej pracy jest opracowanie i porównanie skuteczności trzech wybranych architektur CNN w zadaniu automatycznej klasyfikacji obrazów mikroskopowych na dwie klasy: „okrzemki" vs „inne obiekty". Analizie porównawczej poddano modele EfficientNetB0, ResNet50 oraz VGG16, reprezentujące różne generacje sieci konwolucyjnych – od relatywnie nowej i zoptymalizowanej pod kątem efektywności obliczeniowej (EfficientNetB0) (Tan i in., 2019), przez głęboką sieć rezydualną powszechnie stosowaną w praktyce (ResNet50) (He i in., 2016), po klasyczną, bardzo głęboką sieć konwolucyjną o prostej strukturze sekwencyjnej (VGG16) (Simonyan i in., 2014). Wszystkie modele wykorzystano w podejściu transfer learning (z wagami wstępnie wyuczonymi na ImageNet) i dostrojono do specyfiki zadania. W rozdziale przedstawiono implementację

rozwiązania w języku Python z użyciem bibliotek TensorFlow/Keras oraz scikit-learn, a także wyniki eksperymentów na rzeczywistym zbiorze obrazów mikroskopowych. Zbadano wpływ wybranych hiperparametrów (w szczególności parametrów class_weight związanych z wagą klas w funkcji straty) na jakość klasyfikacji poszczególnych modeli.

Rozdział ma charakter przeglądowo-wdrożeniowy: w *Metodach* opisano dane oraz zastosowane algorytmy i techniki uczenia maszynowego, w *Wynikach* zaprezentowano i porównano efektywność poszczególnych modeli, natomiast *Dyskusja* omawia uzyskane wyniki w kontekście literatury i zastosowań praktycznych, wskazując zalety i ograniczenia badanej metody. We *Wnioskach* końcowych podsumowano najważniejsze obserwacje oraz nakreślono kierunki dalszych badań nad automatycznym rozpoznawaniem okrzemek.

Metody

Dane i przygotowanie danych

Do trenowania i testowania modeli wykorzystano autorski zbiór zdjęć mikroskopowych w jasnym polu, obejmujący dwie kategorie: (1) obrazy zawierające co najmniej jedną okrzemkę (klasa *diatoms*) oraz (2) obrazy nie zawierające okrzemek, lecz inne struktury (klasa *non_diatoms*, np. cząstki detrytusu, inne glony, zanieczyszczenia tła). Źródłem danych były fotografie uzyskane z preparatów środowiskowych, w tym część obrazów pochodziła z dostępnych kolekcji, a część wygenerowano w sposób syntetyczny (dokładniej, dla klasy *non_diatoms* stworzono kilka obrazów z losowo rozmieszczonymi obiektami na wzór zanieczyszczeń tła, aby zwiększyć reprezentację trudnych przypadków). Finalnie zebrano 1618 obrazów okrzemek i 1618 obrazów nie-okrzemek do uczenia modeli. Dane zostały podzielone na zbiory: treningowy (po 1618 obrazów każdej klasy) oraz walidacyjny i testowy – po 350 obrazów z każdej klasy w każdym z tych zbiorów. Każdy obraz zeskalowano do wymiarów 224×224 piksele (zgodnych z wymaganiami wybranych architektur CNN) i zapisano w formacie PNG bez kompresji stratnej. Przed podaniem do sieci wartości pikseli znormalizowano do zakresu [0,1]; ponadto zastosowano standardowe przekształcenie cech wejściowych dedykowane dla danej architektury (funkcje preprocessingowe z biblioteki Keras,

odpowiadające m.in. odjęciu średniej i skalowaniu według statystyk zbioru ImageNet dla zapewnienia kompatybilności cech) (Kloster i in., 2020). Dla zapewnienia wysokiej jakości danych przeprowadzono ręczną inspekcję losowych próbek każdej klasy, usuwając obrazy rozmyte, zaszumione lub błędnie oznaczone.

Architektury sieci i transfer learning. Do klasyfikacji wykorzystano trzy konwolucyjne sieci neuronowe o architekturach zaprojektowanych pierwotnie do wieloklasowej klasyfikacji dużych zbiorów danych (ImageNet). Wykorzystano następujące modele bazowe (bez wierzchnich warstw klasyfikacyjnych): EfficientNetB0 (model z rodziny EfficientNet, zawierający zoptymalizowane skalowanie głębokości i szerokości sieci w celu zwiększenia efektywności; ~5 mln parametrów) (Tan i in., 2019), ResNet50 (głęboka sieć 50-warstwowa z połączeniami rezydualnymi; ~25 mln parametrów) (He i in., 2016) oraz VGG16 (16-warstwowa sieć konwolucyjna o sekwencyjnej budowie; ~138 mln parametrów) (Simonyan i in., 2014). Parametry wszystkich tych modeli inicjalizowano wagami wstępnie wytrenowanymi na zbiorze ImageNet (1,3 mln obrazów naturalnych, 1000 klas). Dzięki temu wykorzystano uniwersalne cechy niskopoziomowe wyekstrahowane z dużego zbioru, co przyspieszyło konwergencję i poprawiło skuteczność klasyfikacji na ograniczonym zbiorze specjalistycznym (transfer learning) [9]. Na bazie wczytanych architektur zbudowano modele sieci hybrydowych dostosowane do klasyfikacji binarnej, dodając następujące warstwy:

- warstwa spłaszczająca cechy konwolucyjne (GlobalAveragePooling2D lub Flatten),
- jedna lub dwie warstwy w pełni połączone (Dense) z aktywacją ReLU (128-256 neuronów),
- warstwa wyjściowa z jednym neuronen (Dense (1)) i aktywacją sigmoid (zwraacająca prawdopodobieństwo klasy *diatoms*).

Powyższa konfiguracja odzwierciedla typowe podejście polegające na wykorzystaniu wstępnie wyuczonych feature extractorów CNN i dołączeniu nowego klasyfikatora dostrojonego do zadania docelowego. We wszystkich modelach zastosowano funkcję straty binary cross-entropy, odpowiednią dla problemu dwuklasowego. Optymalizację przeprowadzono metodą Adam (z domyślnym krokiem uczenia się 0,001) (Buda i in., 2018). Trenowanie odbywało się w dwóch etapach: początkowo zamrożono wszystkie warstwy bazowe modelu (trenując jedynie dodane warstwy gęste) przez ok. 5-10 epok, po czym odblokowano częściowo

niższe warstwy bazowe (stopniowo odmrażając kilka ostatnich bloków konwolucyjnych) i przeprowadzono dalsze dostrajanie (fine-tuning) przez kolejne 10-15 epok ze zmniejszonym learning rate (rzędu 1e-4). Takie dwufazowe podejście zapobiega destabilizacji wczesnych, ogólnych cech podczas inicjalnej fazy uczenia i pozwala na lepsze dostosowanie głębokich warstw w drugiej fazie (Zhuang i in., 2021).

Augmentacja danych i ważenie klas

Aby zwiększyć różnorodność danych treningowych i zmniejszyć ryzyko przeuczenia, zastosowano techniki augmentacji danych za pomocą generatora obrazów (Keras *ImageDataGenerator*). Przy każdym epizodzie treningowym obrazy treningowe poddawano losowym transformacjom:

- obrót o losowy kąt (do 20°),
- przesunięcie w pionie i poziomie (do 20% wymiarów),
- skalowanie (zoom do ±15%),
- odbicie lustrzane w poziomie.

Dzięki tym przekształceniom model uczył się niezmienniczości wobec drobnych zmian orientacji czy skalowania okrzemek, co poprawia jego uogólnienie. Obrazy walidacyjne i testowe nie były augmentowane – poddawano je jedynie takiej samej normalizacji (reskalowaniu intensywności), aby ocena jakości była miarodajna.

Analiza wstępna wykazała, że modele mają trudność z poprawnym rozpoznawaniem klasy *non_diatoms* – przy niezbalansowanych danych lub dużej różnorodności negatywnych przykładów algorytm może ignorować tę klasę, maksymalizując dokładność poprzez klasyfikowanie większości obiektów jako okrzemki. W celu przeciwdziałania temu efektowi wprowadzono ważenie klas w funkcji kosztu. Parametr class_weight ustala relatywną wagę błędu dla danej klasy podczas obliczania straty: wyższa waga dla klasy *non_diatoms* zmusza model do przywiązywania większej uwagi do poprawnej klasyfikacji próbek nie-okrzemek. Dla każdej z testowanych architektur przeprowadzono serię eksperymentów, zmieniając wartość class_weight przypisaną klasie *non_diatoms* (przy założeniu wagi klasy *diatoms* równej 1.0). Zakres testowanych wartości obejmował w zależności od modelu od 1.0 (brak ważenia – obie klasy traktowane

równorzędnie) do nawet 25.0 dla klasy *non_diatoms*. Dzięki temu zbadano wpływ różnych poziomów balansowania klas na ostateczne wyniki klasyfikacji.

Procedura trenowania i ewaluacji

Wszystkie modele trenowano na tej samej partii danych treningowych, przy zachowaniu deterministyczności operacji (ustawienie odpowiednich flag w TensorFlow, np. TF_DETERMINISTIC_OPS=1 w celu porównywalności eksperymentów). Monitorowano przebieg uczenia za pomocą zestawu callbacków Keras: stosowano mechanizm ModelCheckpoint zapisujący najlepszy model (o najwyższej dokładności na zbiorze walidacyjnym) oraz EarlyStopping przerywający trenowanie po kilku epokach bez poprawy wyniku walidacyjnego. Dodatkowo callback ReduceLROnPlateau automatycznie obniżał wartość learning rate o czynnik 0,5 przy stagnacji metryki walidacyjnej, co ułatwiało dopracowanie minimum funkcji kosztu. Po zakończeniu trenowania wybrany najlepszy model był oceniany na niezależnym zbiorze testowym, niewykorzystywanym w procesie uczenia. Jako miary skuteczności klasyfikacji raportowano dokładność (accuracy), a także precyzję, czułość oraz F1-score dla każdej z klas (obliczone na podstawie macierzy pomyłek) (Kloster i in., 2020). W szczególności F1-score okazał się użyteczną miarą, zwłaszcza dla klasy mniejszościowej, gdyż łączy precyzję i czułość w jedną wartość liczbową. Pomiary te posłużyły do porównania efektywności rozpatrywanych modeli CNN oraz oceny wpływu ważenia klas na ich działanie.

Charakterystyka porównywanych modeli

Tabela 1 zestawia kluczowe cechy trzech zaimplementowanych architektur sieci neuronowych. Zawarto przybliżoną liczbę parametrów każdego modelu, co odzwierciedla złożoność sieci (mająca wpływ na wymagania obliczeniowe i potencjalne ryzyko przeuczenia), a także podstawowe informacje o strukturze danej architektury i jej pochodzeniu. Warto podkreślić, że EfficientNetB0 wyróżnia się najmniejszą liczbą parametrów dzięki zastosowaniu automatycznego skalowania głębokości, szerokości i rozdzielczości sieci (Tan i in., 2019); ResNet50 jest pośredni pod względem wielkości, zaś VGG16 posiada anachronicznie dużą liczbę parametrów (model ten nie korzysta z żadnych skrótowych połączeń ani warstw

redukujących wymiarowość cech wewnętrznych, stąd jego duża rozbudowa) (Simonyan i in., 2014). Wszystkie trzy modele osiągały w oryginalnych zadaniach wysokie wyniki – dla przykładu, topologia ResNet50 zdobyła w 2015 roku pierwsze miejsce w konkursie ImageNet (top-5 accuracy ~96%) (He i in., 2016), zaś EfficientNetB0 trenowany na ImageNet uzyskuje ~77% top-1 accuracy przy znacznie mniejszej liczbie parametrów niż porównywalne modele (Tan i in., 2019).

Tabela 1. Podsumowanie wybranych architektur CNN użytych w badaniu

Model	Rok publikacji	Liczba parametrów ~	Charakterystyka architektury
EfficientNetB0	2019	5,3 mln	Zoptymalizowana sieć CNN (compound scaling); najmniejszy model z rodziny EfficientNet; głębokość ~16 warstw konwolucyjnych.
ResNet50	2016	25,6 mln	Głęboka sieć rezydualna (residual learning) składająca się z ~50 warstw; wykorzystuje bloki ResNet z połączeniami skrótowymi redukującymi zanik gradientu.
VGG16	2014	138,3 mln	Bardzo głęboka sieć sekwencyjna (16 warstw konwolucyjnych) o prostym układzie; duża liczba parametrów z uwagi na rozbudowane warstwy w pełni połączone w oryginalnej architekturze.

Źródło: opracowanie własne.

Wyniki

Eksperymenty wstępne (zbiór ograniczony)

Na początkowym etapie przeprowadzono testy z ograniczonym zbiorem danych, aby zweryfikować poprawność implementacji i ocenić wpływ parametrów treningu. W eksperymencie tym wykorzystano niewielką próbkę 120 obrazów (po 60 z każdej klasy) – taki rozmiar zbioru celowo wybrano w celu zasymulowania sytuacji skrajnie małej liczby danych. Sieć EfficientNetB0 została nauczona w dwóch wariantach: (a) bez ważenia klas oraz (b) z zastosowaniem

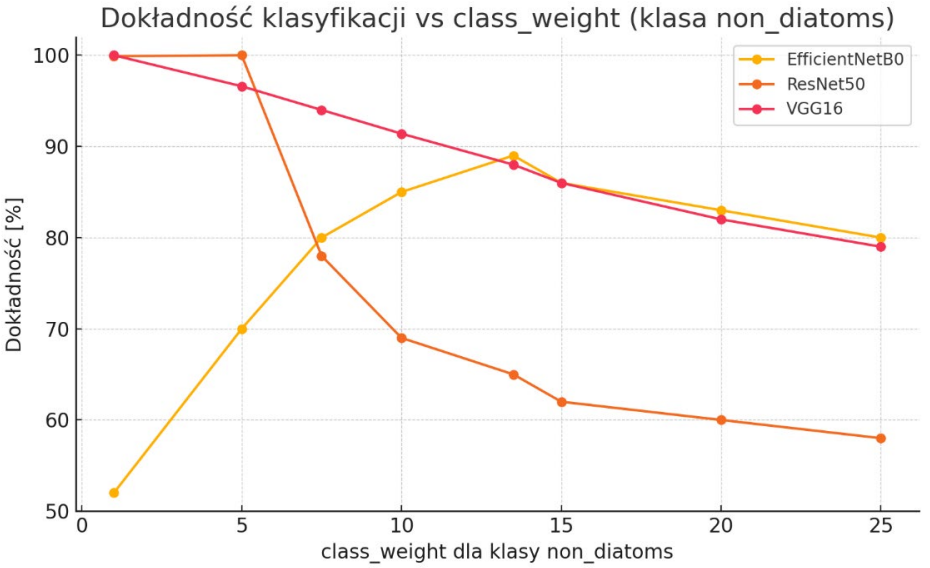
wagi class_weight = 1.3 dla klasy *non_diatoms*. Otrzymane wyniki uwidoczniły wyraźny problem niezrównoważenia: bez ważenia klas model osiągał blisko 100% czułość dla klasy okrzemek kosztem niemal całkowitego pomijania klasy nie-okrzemek (F1-score dla klasy *non_diatoms* wynosił zaledwie 0,09 przy ~52% ogólnej dokładności). Po wprowadzeniu niewielkiej modyfikacji wag klas sytuacja uległa diametralnej poprawie – model zaczął rozpoznawać obie klasy z porównywalną skutecznością (dokładność wzrosła powyżej 90%, a miary F1 obu klas wyrównały się na poziomie ~0,9). Wskazuje to, że nawet przy bardzo ograniczonym zbiorze, odpowiednie zbalansowanie funkcji kosztu potrafi znacząco wpłynąć na wyniki klasyfikacji. Obserwacja ta utwierdziła nas w konieczności analizy wpływu parametru class_weight również dla pełnego zbioru danych.

Eksperymenty na pełnym zbiorze – wpływ class_weight

Główne eksperymenty przeprowadzono na pełnym zestawie 3236 obrazów (1618+1618 trening, 700 test), trenując i testując po kolei każdy z trzech modeli CNN. Dla każdego modelu zbadano różne wartości parametru class_weight przypisanego klasie *non_diatoms*, tak aby znaleźć optymalny balans między precyzją wykrywania okrzemek i nie-okrzemek. Rysunek 1 ilustruje zależność dokładności testowej od wartości wagi klasy *non_diatoms* dla EfficientNetB0, ResNet50 i VGG16. Wykres przedstawia procent poprawnych klasyfikacji na zbiorze testowym w funkcji wartości parametru class_weight (relatywnej wagi klasy „nie-okrzemka" w funkcji kosztu) dla trzech architektur: EfficientNetB0, ResNet50 i VGG16. Z kolei Tabela 2 zestawia najlepsze osiągnięte wyniki (maksymalną dokładność i odpowiadający jej F1-score) dla każdej architektury wraz z informacją, przy jakiej wartości class_weight zostały uzyskane.

Na podstawie powyższych wyników można zauważyć wyraźnie odmienne zachowanie poszczególnych modeli. EfficientNetB0 bez ważenia klas (class_weight = 1.0) osiągał bardzo niską skuteczność ogólną (ok. 52% – wynik zbliżony do losowego, potwierdzający, że model w tej konfiguracji praktycznie nie rozpoznawał klasy *non_diatoms*). Stopniowe zwiększanie wagi klasy nie-okrzemek przynosiło poprawę – Rysunek 1 pokazuje, że dokładność rosła wraz z class_weight do pewnego punktu, po czym zaczęła spadać. Optimum dla EfficientNetB0 osiągnięto przy class_weight ≈ 13,5, gdzie dokładność testowa wyniosła 89,0%,

a F1-score dla obu klas przekroczył 0,88 (Tabela 2). Dalsze zwiększanie wagi (np. do 15) skutkowało już pogorszeniem wyników, zapewne na skutek przewartościowania klasy *non_diatoms* i nadmiernego karania pomyłek w tej klasie (prowadzącego do spadku trafności dla klasy *diatoms*).



Rysunek 1. Dokładność klasyfikacji w zależności od wagi klasy non-diatoms dla poszczególnych modeli CNN

Źródło: opracowanie własne.

W przypadku ResNet50 tendencja była w dużej mierze odwrotna. Ten model już przy braku ważenia (class_weight = 1.0) poradził sobie znakomicie, osiągając prawie 100% dokładności (F1 = 1.00 dla obu klas). Nawet niewielkie zwiększenie wagi do 5,0 dało perfekcyjną klasyfikację na zbiorze testowym (100% poprawnych decyzji). Jednak dalsze zwiększanie class_weight powodowało gwałtowne załamanie efektywności – przy class_weight = 7.5 dokładność spadła do ~78%, a przy 10.0 – do ~69% (Rysunek 1). Widzimy zatem, że ResNet50 był bardzo wrażliwy na zbyt wysokie wagi klas, co objawiało się przeuczeniem modelu na klasę *non_diatoms* i utratą ogólnej zdolności generalizacji. Należy podkreślić, że optimum dla ResNet50 leżało w zakresie niskich wartości class_weight – praktycznie najlepsze wyniki uzyskano bez żadnego ważenia (model ten naturalnie

poradził sobie z rozpoznaniem obu klas, być może dzięki swojej dużej głębokości i zdolności do uchwycenia subtelnych różnic wizualnych).

VGG16 wykazał zbliżoną skuteczność do ResNet50 w zakresie niższych wag klas. Przy class_weight = 1.0 (brak balansu) model ten również osiągnął 100% dokładności testowej, prawidłowo klasyfikując wszystkie obrazy z obu klas. Świadczy to o znakomitym dopasowaniu do danych – co może jednak budzić obawy o ewentualne przeuczenie, szczególnie zważywszy na dużą pojemność sieci VGG16. W odróżnieniu od ResNet50, sieć VGG16 nie była aż tak wrażliwa na umiarkowane zwiększenie wagi klasy mniejszości: dla class_weight = 5 dokładność wyniosła 96,6%, a dla 10 – 91,4%. Dopiero wyższe wartości (15-25) spowodowały istotny spadek efektywności (do ~79% przy class_weight = 25, Rysunek 1). Oznacza to, że VGG16 cechowała się pewną odpornością na nadmierne dociążenie klasy *non_diatoms* – spadek następował stopniowo, wolniej niż w przypadku ResNet50. Najlepszy wynik (100% trafności) odnotowano jednak przy braku ważenia, podobnie jak dla ResNet50.

Tabela 2. Najlepsze osiągnięte wyniki klasyfikacji na zbiorze testowym dla poszczególnych modeli (metryki dla optymalnej wartości parametru class_weight, w nawiasie podano tę wartość)

Model	class_weight (optimum)	Dokładność (Accuracy)	F1-score (diatoms)	F1-score (non_diatoms)
EfficientNetB0	13.5	89.0%	0.90	0.88
ResNet50	5.0	100%	1.00	1.00
VGG16	1.0	100%	1.00	1.00

Źródło: opracowanie własne.

Analizując powyższą tabelę (Tabela 2), można stwierdzić, że najwyższą skuteczność (w sensie dokładności i F1) uzyskano dla modeli ResNet50 i VGG16 – oba te modele osiągnęły *perfekcyjne wyniki* na zbiorze testowym (brak pomyłek w 700 obrazach testowych). EfficientNetB0, mimo że nie dorównał im pod względem bezwzględnej dokładności, również zdołał poprawnie sklasyfikować blisko 9 na 10 obrazów testowych przy odpowiednim dostrojeniu wag klas. Warto jednak zwrócić uwagę, że rezultat 100% może świadczyć o pewnej łatwości zadania dla bardzo pojemnych modeli (ResNet, VGG) – w dyskusji omawiamy, czy wynika to z rzeczywistej generalizacji modelu, czy też z ograniczeń różnorodności danych

testowych. Istotne jest również porównanie zachowania modeli: EfficientNetB0 wymagał intensywnego ważenia klasy, aby nauczyć się rozpoznawać obiekty niebędące okrzemkami, podczas gdy ResNet50 i VGG16 radziły sobie z tym *ex aequo* już bez dodatkowych zabiegów. Świadczy to o tym, że mniej złożony model (EfficientNetB0) miał tendencję do ignorowania klasy trudniejszej (non_diatoms) i dopiero silna regularyzacja kosztu zmusiła go do zrównoważenia predykcji, natomiast modele głębsze (ResNet, VGG) dysponowały wystarczającą mocą reprezentacyjną, by od razu uczyć się cech odróżniających obie klasy. Z drugiej strony, przesadne zwiększenie wagi klasy minorowej szkodziło modelom głębokim (zwłaszcza ResNet50), co sugeruje, że w ich przypadku mechanizm residualny i duża głębia samoczynnie zapewniają pewną odporność na niezbalansowanie, a nadmierna ingerencja w postaci dużych wag może rozregulować proces trenowania.

Dyskusja

Uzyskane wyniki potwierdzają, że głębokie sieci neuronowe z powodzeniem można zastosować do automatycznego rozpoznawania okrzemek na zdjęciach mikroskopowych. Wszystkie trzy przetestowane architektury (EfficientNetB0, ResNet50, VGG16) wykazały bardzo wysoką skuteczność klasyfikacji – w najlepszych konfiguracjach przekraczającą 89% dokładności, a w dwóch przypadkach dochodzącą nawet do 100%. Oznacza to, że modele te były w stanie nauczyć się charakterystycznych cech odróżniających okrzemki (np. symetria i ornamentyka pancerzyka) od innych struktur obecnych w obrazie. Jest to wynik zgodny z trendami obserwowanymi w literaturze: już wcześniejsze prace wskazywały, że głębokie CNN potrafią efektywnie rozpoznawać nawet drobne różnice morfologiczne organizmów, przewyższając metody tradycyjne. Co więcej, uzyskane F1-score rzędu 0,9-1,0 dla klasy okrzemek sugerują, że podejście oparte na transfer learningu może znacząco usprawnić analizy ekologiczne – automatyczny klasyfikator jest w stanie z dużą pewnością wskazać obrazy zawierające okrzemki, co z punktu widzenia monitoringu środowiskowego ma duże znaczenie praktyczne (np. przyspieszenie wstępnego etapu oceny próbek).

Wyniki ujawniły jednak istotne różnice między architekturami, szczególnie w kontekście czułości na parametr class_weight. Model EfficientNetB0

początkowo miał wyraźną tendencję do faworyzowania klasy okrzemek kosztem nie-okrzemek – bez odpowiedniego ważenia klas niemal wszystkie próbki negatywne były klasyfikowane błędnie. Jest to zrozumiałe, biorąc pod uwagę naturę danych: klasa *non_diatoms* obejmuje obiekty bardzo zróżnicowane (różne rodzaje zanieczyszczeń, fragmenty innych organizmów, artefakty tła), podczas gdy klasa *diatoms* skupia względnie jednorodne, łatwo rozpoznawalne struktury (promieniste lub symetryczne kształty o wyrazistych teksturach). Mniej złożony model, jak EfficientNetB0, dysponujący ograniczoną liczbą filtrów, mógł skoncentrować się na łatwiejszej do uchwycenia klasie okrzemek, ignorując „szum" klasy przeciwnej. Dopiero podniesienie wagi błędu dla nie-okrzemek w funkcji kosztu skutecznie wymusiło na modelu nauczenie się cech charakterystycznych również dla klasy negatywnej – w efekcie osiągnął on blisko 90% dokładności, poprawnie identyfikując znaczną część wcześniej pomijanych obiektów. Z punktu widzenia praktycznego oznacza to, że w zastosowaniach z małymi lub niejednorodnymi zbiorami danych konieczne może być stosowanie technik balansowania klas (jak class weighting czy nadpróbkowanie danych), aby uniknąć silnego biasu modelu w kierunku dominującej/łatwiejszej klasy (Buda i in., 2018).

W przeciwieństwie do EfficientNetB0, modele ResNet50 i VGG16 okazały się dużo bardziej „robustowe" – bez jakiegokolwiek ważenia klas potrafiły one prawidłowo sklasyfikować praktycznie wszystkie próbki testowe obu klas. Można to interpretować następująco: architektury te, będąc znacznie większe (odpowiednio ~25 mln i ~138 mln parametrów), są w stanie uchwycić subtelne różnice pomiędzy obrazami okrzemek i nie-okrzemek nawet wtedy, gdy przykłady klasy negatywnej są różnorodne i potencjalnie trudniejsze do nauczenia. ResNet50, dzięki głębokiej strukturze z blokami rezydualnymi, efektywnie uczy się złożonych reprezentacji – przypuszczalnie zawarł on w swoich głębokich warstwach opis wielu rodzajów obiektów tła, co pozwoliło na niemal bezbłędne ich odróżnianie od okrzemek. Podobnie VGG16, mimo prostej architektury, ma ogromną liczbę parametrów, co umożliwia mu zapamiętanie nawet drobnych detali wszystkich obrazów treningowych. Nasze wyniki sugerują, że dla prostego zadania binarnego duże modele mogą łatwo osiągnąć maksymalną skuteczność, jeśli tylko dysponują wystarczającymi danymi do treningu. W naszym przypadku zarówno ResNet50, jak i VGG16 osiągnęły 100% trafności na zbiorze testowym – co wskazuje, że problem został przez nie praktycznie rozwiązany w całości dla dostępnych

danych. Podobną obserwację odnotowali Kloster i in. (2020), w którym model VGG16 również wykazał się najwyższą skutecznością klasyfikacji okrzemek na pewnym zestawie danych. Warto jednak zaznaczyć, że tak wysoka dokładność może być częściowo złudna – istnieje ryzyko, że model mógł nauczyć się pewnych artefaktów specyficznych dla naszego zbioru (np. charakterystycznego tła preparatu czy sposobu akwizycji obrazu), co ograniczałoby jego uogólnienie do nowych danych. Innymi słowy, perfekcyjny wynik na teście może oznaczać pełne dopasowanie do próbek z badanego korpusu, lecz w warunkach rzeczywistych (inne laboratoria, inne mikroskopy) skuteczność może być nieco niższa. Z tego względu, pomimo znakomitych wyników ResNet50 i VGG16, w praktyce warto zachować ostrożność i dalej walidować model na szerszym spektrum danych.

Interesujące spostrzeżenie z naszych eksperymentów to różnice w odporności na przeuczenie i wrażliwości na parametry pomiędzy modelami. ResNet50 okazał się bardzo stabilny przy niskich wartościach class_weight (osiągał wysoką skuteczność bez balansu), ale zarazem gwałtownie tracił dokładność przy zbyt dużych wagach dla klasy minorowej. Sugeruje to, że model ten miał wbudowaną zdolność radzenia sobie z pewnym niezbalansowaniem, a zbyt agresywne wymuszanie uwagi na klasie trudniejszej zaburzało jego proces uczenia – być może prowadząc do nadmiernego dostrojenia filtrów na specyficzne cechy nie-okrzemek kosztem utraty ogólnych właściwości dyskryminacyjnych. Z kolei VGG16 wydaje się odporny na przeuczenie w tym sensie, że nawet dla sporych wag klas nie dochodziło do całkowitego załamania wyników (spadek był stopniowy). Można spekulować, że ze względu na architekturę sekwencyjną bez połączeń rezydualnych VGG16 uczy się bardziej „hierarchicznie" – zwiększenie wagi klasy *non_diatoms* stopniowo modyfikowało decyzje ostatnich warstw, ale wcześniejsze warstwy nadal wyłapywały szeroki zakres cech, zachowując umiarkowaną efektywność dla obu klas. EfficientNetB0 cechował się natomiast największą wrażliwością na ustawienia hiperparametrów – wymagał starannego doboru class_weight oraz innych parametrów treningu (np. zakresu augmentacji, harmonogramu uczenia), aby osiągnąć wynik porównywalny z pozostałymi modelami. Jest to zgodne z intuicją: mniejszy model ma mniejszy margines błędu w doborze parametrów, gdyż łatwiej może nie zdążyć nauczyć się wszystkiego przy niewłaściwej konfiguracji. Z drugiej strony, EfficientNetB0 jest modelem bardzo efektywnym obliczeniowo – trenował się znacznie szybciej i zużywał mniej pamięci GPU niż ResNet50 czy

VGG16, co w praktyce może być istotną zaletą (szczególnie przy ograniczonych zasobach sprzętowych lub w zastosowaniach *edge computing*).

Podsumowując, nasze badania pokazują, że w zadaniu binarnej klasyfikacji obrazów mikroskopowych okrzemek każdy z rozważanych modeli może osiągnąć bardzo dobre wyniki, ale wybór optymalnego rozwiązania powinien uwzględniać kontekst zastosowania. Jeśli priorytetem jest maksymalna skuteczność i dostępne są odpowiednie zasoby obliczeniowe, modele pokroju ResNet50 czy VGG16 zapewniają najwyższą jakość predykcji (choć potencjalnie kosztem ryzyka przeuczenia, które należy kontrolować poprzez walidację na zewnętrznych danych). Z drugiej strony, jeśli kluczowa jest szybkość działania, niska zajętość pamięci i łatwość wdrożenia (np. na urządzeniu przenośnym w terenie), EfficientNetB0 stanowi atrakcyjną alternatywę – oferuje on zbalansowaną precyzję przy ułamku liczby parametrów VGG16. W naszym eksperymencie najlepsza konfiguracja EfficientNetB0 (z class_weight=13.5) osiągnęła wprawdzie niższą dokładność niż pozostałe modele, ale wciąż na poziomie ~89%, co w wielu zastosowaniach mogłoby być wynikiem akceptowalnym, zwłaszcza biorąc pod uwagę korzyści w postaci szybkości i mniejszych wymagań sprzętowych. W literaturze wskazuje się, że często nowsze, zminiaturyzowane architektury są w stanie dorównać klasycznym, ciężkim modelom dzięki lepszemu wykorzystaniu parametrów (Tan i in., 2019). W naszym przypadku różnica w skuteczności między EfficientNetB0 a ResNet50/VGG16 jest zauważalna, jednak można przypuszczać, że na trudniejszych, wieloklasowych zadaniach różnice te mogłyby się zmniejszyć (większe modele mogłyby szybciej popaść w przeuczenie, podczas gdy EfficientNet generalizowałby lepiej). To pozostaje jednak kwestią do zbadania w odrębnych eksperymentach.

Porównując nasze wyniki z innymi badaniami, należy wspomnieć, że większość dotychczasowych prac skupiała się na klasyfikacji gatunków okrzemek (problem wieloklasowy) albo detekcji i segmentacji okrzemek na obrazie. Zaproponowane podejście upraszcza zagadnienie wykrycia obecności okrzemek w obrazie. Można je traktować jako moduł w szerszym systemie – np. pierwszy krok pipeline'u, po którym nastąpiłaby identyfikacja gatunkowa znalezisk. Taka dwuetapowa strategia (detekcja → identyfikacja) bywa stosowana w systemach analizy planktonu i okrzemek, co pozwala rozdzielić zadania i zastosować wyspecjalizowane modele na kolejnych etapach. Wyniki naszego binarnego

klasyfikatora sugerują, że etap detekcji okrzemek można zrealizować z dużą pewnością przy użyciu stosunkowo prostego modelu CNN. Co więcej, transfer learning okazał się kluczowy – bez niego, dysponując jedynie ~3 tysiącami obrazów, prawdopodobnie nie uzyskalibyśmy zbliżonych wyników (trening głębokiej sieci od zera na tak ograniczonym zbiorze mógłby prowadzić do słabej konwergencji lub silnego overfittingu). Warto również zaznaczyć, że użycie wag klas jest prostą, lecz skuteczną metodą poprawy miar dla klasy rzadkiej/trudniejszej – w dziedzinie bioinformatyki obrazowej często zmagamy się z niezrównoważonymi danymi (np. rzadkie gatunki w próbie), stąd obserwacje dotyczące doboru class_weight (Rysunek 1) mogą być przydatne przy trenowaniu innych modeli do detekcji rzadkich obiektów biologicznych.

Ograniczenia badań

Należy wspomnieć o pewnych ograniczeniach przeprowadzonego eksperymentu. Po pierwsze, zbiór danych, choć stosunkowo duży jak na zadanie niszowe, może nie odzwierciedlać pełni zmienności spotykanej w praktyce. Obrazy pochodziły z określonych źródeł i były częściowo przetworzone (np. przez wygładzenie tła, usunięcie oczywistych artefaktów), co mogło ułatwić modelom zadanie. W przyszłości warto rozszerzyć zbiór o zdjęcia z różnych mikroskopów, o odmiennych parametrach akwizycji i z różnych lokalizacji geograficznych, by sprawdzić, czy modele utrzymają wysoką skuteczność (zagadnienie zdolności generalizacji domenowej). Po drugie, nasza ocena ograniczyła się do klasyfikacji całych obrazów z pojedynczym obiektem – w realnych próbkach często na jednym zdjęciu znajduje się wiele okrzemek różnych gatunków lub okrzemki przemieszane z innymi detrytusami. Nasz model w obecnej postaci nie lokalizuje okrzemek na obrazie, a jedynie stwierdza ich obecność. Z praktycznego punktu widzenia przydatne byłoby połączenie klasyfikatora z modułem detekcji/segmentacji, który zaznaczałby położenie okrzemek na większych obrazach (np. przy użyciu sieci typu U-Net lub detektorów obiektów YOLO/Faster R-CNN) (Kloster i in., 2020). Po trzecie, skupiliśmy się na trzech wybranych architekturach CNN, jednak rozwój w tej dziedzinie jest bardzo dynamiczny. Istnieją nowsze podejścia, jak chociażby sieci transformerowe (Vision Transformers) czy bardziej zaawansowane modele konwolucyjne (DenseNet, EfficientNet w nowszych wersjach,

MobileNetV3 itp.), które mogą oferować lepszy kompromis między złożonością a skutecznością. W przyszłości celowe byłoby przetestowanie także tych rozwiązań w kontekście okrzemek. Mimo powyższych ograniczeń przeprowadzone badanie dostarczyło cennych informacji o zachowaniu klasycznych modeli CNN w zadaniu mikroskopowej klasyfikacji i stanowi solidną podstawę do dalszych prac rozszerzających funkcjonalność systemu.

Podsumowanie i wnioski

Przedstawione w pracy wyniki jednoznacznie pokazują, że głębokie sieci konwolucyjne stanowią skuteczne narzędzie do automatycznej klasyfikacji obrazów mikroskopowych okrzemek. Wykorzystanie architektur CNN wraz z techniką transfer learningu umożliwiło uzyskanie bardzo wysokiej trafności detekcji – najlepsze modele (ResNet50, VGG16) osiągnęły na niezależnym zbiorze testowym dokładność rzędu 99-100%, zaś lżejszy EfficientNetB0, przy odpowiednim dostrojeniu, zbliżył się do 90%. Zastosowanie pretrenowanych modeli znacząco przyspieszyło proces uczenia i pozwoliło efektywnie wykorzystać ograniczony zbiór danych, potwierdzając korzyści transfer learningu w dziedzinie analizy obrazów biologicznych. Wprowadzając ważenie klas w funkcji straty, zdołaliśmy zrównoważyć wpływ poszczególnych klas na uczenie – zmniejszając liczbę błędów przy rozpoznawaniu trudniejszych przypadków (nie-okrzemek) nawet o rząd wielkości.

Badania wykazały, że nie istnieje jeden uniwersalnie najlepszy model – każdy z testowanych ma swoje zalety. ResNet50 okazał się najstabilniejszy i łatwy w trenowaniu (osiągał doskonałe wyniki bez specjalnych zabiegów, choć wymaga umiarkowanego ważenia, by nie rozregulować jego uczenia). VGG16 mimo starszej konstrukcji nadal zapewnia znakomitą skuteczność i wykazał odporność na przeuczenie przy niewielkim niezbalansowaniu danych, jednak jest modelem bardzo dużym, co może utrudniać jego praktyczne wykorzystanie w środowiskach o ograniczonej mocy obliczeniowej. EfficientNetB0 z kolei wyróżniał się najwyższą efektywnością obliczeniową i kompaktowością – przy optymalnych ustawieniach osiągnął wysoką precyzję detekcji okrzemek, ustępując nieco dokładnością większym sieciom, ale oferując potencjalnie lepszą integrację w aplikacjach działających w czasie rzeczywistym czy na urządzeniach mobilnych. Wybór modelu

powinien być zatem dostosowany do specyfiki danych i wymagań aplikacji: jeżeli dysponujemy zróżnicowanymi danymi i zależy nam na maksymalnej skuteczności – głębsza sieć może być właściwym wyborem; jeśli natomiast kluczowa jest szybkość i prostota wdrożenia – efektywna architektura pokroju EfficientNetB0 może okazać się wystarczająca.

Z praktycznego punktu widzenia wyniki te dowodzą, że zautomatyzowane systemy rozpoznawania okrzemek mogą realnie wspomóc pracę biologów i ekologów. Zastosowanie modelu CNN pozwala przyspieszyć analizę mikroskopowych preparatów – komputer może w kilka sekund ocenić setki obrazów, podczas gdy manualne sprawdzenie zajęłoby specjalistom wiele godzin. Co więcej, automatyczna klasyfikacja eliminuje czynnik subiektywny, zapewniając powtarzalność ocen. Tego typu narzędzia mogą znaleźć zastosowanie w monitoringu jakości wód (szybkie wykrywanie zmian w zespołach okrzemek wskazujących na degradację środowiska) czy w badaniach paleoekologicznych (masowa analiza próbek osadów pod kątem zawartości okrzemek). Oczywiście, w obecnej formie model rozpoznaje jedynie fakt obecności okrzemki, nie zastąpi więc eksperta w identyfikacji gatunków – jednak stanowi wartościowy element wspomagający, filtrując dane i kierując uwagę człowieka tam, gdzie jest to potrzebne.

Planowane dalsze prace obejmują kilka kierunków. Po pierwsze, rozbudowę i urozmaicenie zbioru danych – zgromadzenie większej liczby zdjęć okrzemek z różnych źródeł, a także trudniejszych przykładów (np. obrazów z silnym szumem, różnym tłem, różnymi gatunkami okrzemek), co pozwoli jeszcze lepiej ocenić uogólnienie modeli. Po drugie, integrację modułu detekcji obiektów lub segmentacji okrzemek, tak aby system potrafił zlokalizować okrzemki na większych obrazach zawierających wiele organizmów. Wykorzystanie w tym celu np. segmentacyjnej sieci U-Net lub detektorów opartych o CNN (jak YOLO, Faster R-CNN) umożliwi analizę bardziej złożonych zdjęć mikroskopowych, z jakimi spotykają się praktycy. Po trzecie, planowane jest przetestowanie nowszych architektur głębokiego uczenia, w tym sieci DenseNet, EfficientNet w wyższych wersjach (B3-B7), a także modeli opartych na transformatory wizji (Vision Transformers), które ostatnio osiągają rewelacyjne wyniki w wielu zadaniach klasyfikacji obrazów. Czwartym aspektem jest zwiększenie interpretowalności modelu – zastosowanie metod eXplainable AI (np. wizualizacji aktywacji typu Grad-CAM) mogłoby dostarczyć wglądu, na jakie cechy okrzemek zwraca uwagę sieć, co nie

tylko buduje zaufanie do modelu, ale i może pomóc biologom w odkryciu nowych diagnostycznych cech morfologicznych.

Reasumując, niniejsza praca dowiodła, że nowoczesne metody głębokiego uczenia umożliwiają automatyzację rozpoznawania okrzemek z wysoką skutecznością. Stanowi to ważny krok w kierunku pełnego cyfrowego wspomagania analiz ekologicznych – od akwizycji obrazu, przez detekcję i identyfikację organizmów, po interpretację wyników. Wykorzystanie tego typu rozwiązań w praktyce może przełożyć się na przyspieszenie i obiektywizację monitoringu środowiska wodnego, dostarczając narzędzi bioinformatycznych wspierających ochronę ekosystemów wodnych i badania bioróżnorodności.

Literatura

1. Buda, M., Maki, A., Mazurowski, M. A. (2018). A systematic study of the class imbalance problem in convolutional neural networks. *Neural Netw, 106*, 249-259. https://doi.org/10.1016/j.neunet.2018.07.011
2. du Buf, H., Bayer, M. M. (red.). (2002). *Automatic Diatom Identification*. World Scientific, Singapore. https://doi.org/10.1142/4907
3. Gunduz, H., Gunal, S. (2024). A lightweight convolutional neural network model for diatom classification: DiatomNet. *PeerJ Comput Sci, 10*, e1970. https://doi.org/10.7717/peerj-cs.1970
4. He, K., Zhang, X., Ren, S., Sun, J. (2016). Deep residual learning for image recognition. *Proc. CVPR*, 770-778. https://doi.org/10.1109/CVPR.2016.90
5. Kireta, A. R., Reavie, E. D., Sgro, G. V. (2012). Planktonic and periphytic diatoms as indicators of stress on great rivers of the United States: Testing water quality and disturbance models. *Ecol Indic, 13*(1), 222-231. https://doi.org/10.1016/j.ecolind.2011.06.006
6. Kloster, M., Langenkämper, D., Zurowietz, M., Beszteri, B., Nattkemper, T. W. (2020). Deep learning-based diatom taxonomy on virtual slides. *Sci Rep, 10*(1), 14416. https://doi.org/10.1038/s41598-020-71165-w
7. Krizhevsky, A., Sutskever, I., Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Commun ACM*, *60*(6), 84-90. https://doi.org/10.1145/3065386
8. LeCun, Y., Bengio, Y., Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436-444. https://doi.org/10.1038/nature14539

9. Pedraza, A., Bueno, G., Deniz, O. (2017). Automated Diatom Classification (Part B): A Deep Learning Approach. *Appl Sci (Basel), 7*(5), 460. https://doi.org/10.3390/app7050460
10. Pu, S. Y., Zhang, F., Shu, Y. X., Fu, W. Q. (2023). Microscopic image recognition of diatoms based on deep learning. *J Phycol, 59*(5), 1166-1178. https://doi.org/10.1111/jpy.13390
11. Simonyan, K., Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint* arXiv, *1409*, 1556. https://doi.org/10.48550/arXiv.1409.1556
12. Tan, M., Le, Q. V. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. *Proc. 36th ICML*, 6105-6114. https://doi.org/10.48550/arXiv.1905.11946
13. Zhuang, F., Qi, Z., Duan, K. (2021). A comprehensive survey on transfer learning. *Proc IEEE*, *109*(1), 43-76. https://doi.org/10.1109/JPROC.2020.3004555

Krzysztof Łuczka[1]
Uniwersytet Ekonomiczny w Poznaniu
https://doi.org/10.29316/9788368103304_6

ZMIANY I INNOWACJE W OCENIE ZDOLNOŚCI KREDYTOWEJ

CHANGES AND INNOVATIONS IN CREDITWORTHINESS ASSESSMENT

Streszczenie

Celem rozdziału jest przedstawienie ewolucji metod oceny zdolności kredytowej oraz roli nowoczesnych technologii w tym procesie. Analiza obejmuje zarówno klasyczne podejścia statystyczne, jak i współczesne metody oparte na uczeniu maszynowym, sztucznej inteligencji oraz dużych modelach językowych (LLM). W pierwszej części omówiono historyczne ujęcie problemu, począwszy od wczesnych form oceny opartej na reputacji i subiektywnej analizie kredytodawców, aż po rozwój systemu punktowego FICO i standaryzację credit scoringu. Następnie zaprezentowano najnowsze rozwiązania, takie jak modele hybrydowe (np. TCN-DilateFormer), algorytmy zespołowe (Random Forest, XGBoost, TabNet) oraz zastosowanie metod głębokiego uczenia w prognozowaniu ryzyka kredytowego. Praca omawia również polskie uwarunkowania systemowe - w szczególności funkcjonowanie Biura Informacji Kredytowej oraz znaczenie Rekomendacji T w zarządzaniu ryzykiem detalicznym. Wnioski wskazują, że choć metody tradycyjne pozostają fundamentem oceny kredytowej, to integracja algorytmów uczenia maszynowego

[1] https://orcid.org/0009-0003-6898-7155, e-mail: krzysztofluczka@outlook.com

pozwala na zwiększenie dokładności i efektywności prognoz. Jednocześnie podkreślono konieczność zapewnienia interpretowalności modeli oraz zachowania równowagi między automatyzacją decyzji a czynnikiem ludzkim.

Słowa kluczowe: ocena zdolności kredytowej, scoring kredytowy, uczenie maszynowe, modele predykcyjne, Biuro Informacji Kredytowej, technologie finansowe, algorytmy klasyfikacyjne

Abstract

The aim of this chapter is to present the evolution of creditworthiness assessment methods and the role of modern technologies in this process. The analysis encompasses both classical statistical approaches and contemporary methods based on machine learning, artificial intelligence, and large language models (LLMs). The first part discusses the historical approach to the problem, from early forms of assessment based on reputation and subjective lender analysis, to the development of the FICO scoring system and the standardisation of credit scoring. This is followed by a presentation of recent solutions, including hybrid models (e.g., TCN-DilateFormer), ensemble algorithms (Random Forest, XGBoost, TabNet), and the application of deep learning methods to credit risk forecasting. The work also discusses the Polish systemic conditions – in particular, the operation of the Polish Credit Information Bureau (BIK) and the importance of "Rekomendacja T" in retail risk management. The conclusions indicate that while traditional methods remain the foundation of credit assessment, integrating machine learning algorithms increases forecast accuracy and efficiency. The need to ensure model interpretability and maintain a balance between decision automation and the human factor is also emphasised.

Keywords: creditworthiness assessment, credit scoring, machine learning, predictive models, Biuro Informacji Kredytowej, financial technologies, classification algorithms

Wstęp

Nowoczesne rozwiązania technologiczne pozwalają nie tylko szybciej gromadzić i przetwarzać ogromne ilości danych, ale również odkrywać wzorce, których człowiek samodzielnie nie byłby w stanie dostrzec. Dzięki technologiom opartym na sztucznej inteligencji, uczeniu maszynowym czy analizie big data możliwe staje się bardziej precyzyjne prognozowanie ryzyka kredytowego i dopasowanie oferty do indywidualnych możliwości klienta. Technologia zmienia

sposób, w jaki instytucje finansowe podchodzą do samej definicji zdolności kredytowej – pozwala ją rozszerzać o nowe źródła danych (takie jak dane cyfrowe, ślady w Internecie, zachowania konsumenckie), a przez to czyni ocenę bardziej kompleksową, dynamiczną i dostosowaną do współczesnych realiów gospodarki cyfrowej.

Zdolność kredytowa to pojęcie bardzo szerokie, które nie ogranicza się jedynie do matematycznego wskaźnika czy prostego porównania przychodów i zobowiązań. W niniejszej pracy będzie oznaczała całościową ocenę możliwości wywiązania się przez klienta z przyszłych zobowiązań finansowych, przy czym obejmuje ona zarówno aspekty finansowe (dochody, wydatki, stabilność źródeł przychodu), jak i elementy behawioralne (historia spłat, wiarygodność, przewidywalność zachowań ekonomicznych), a także szerszy kontekst społeczno-gospodarczy. Jest to wielowymiarowy konstrukt, który łączy w sobie obiektywne dane liczbowe z bardziej jakościowymi informacjami o osobie lub podmiocie ubiegającym się o kredyt.

Rozdział ten stanowi opisowy przegląd literatury Scopus / Web of Science / Google Scholar – nie posiada on ściśle zdefiniowanych kryteriów doboru artykułów, zatem nie jest on ani systematycznym, ani metaanalitycznym opracowaniem całego dorobku badawczego w tym zakresie. Jego celem jest przede wszystkim syntetyczne przedstawienie wybranych publikacji, które w sposób szczególnie wyrazisty ukazują ewolucję metod oceny zdolności kredytowej oraz rolę nowoczesnych technologii w tym procesie. Nacisk został położony na ogólne kierunki rozwoju i główne wnioski badawcze, a nie na szczegółową rekonstrukcję procedur metodologicznych. Za przyjęciem takiej formy stoi fakt, że ze względu na szybkość zmian w obszarze technologii finansowych, przedstawione ustalenia mogą stosunkowo szybko ulec dezaktualizacji. Mimo tych ograniczeń narracyjny charakter przeglądu pozwala uchwycić kluczowe trendy oraz sformułować wnioski istotne z perspektywy dalszych badań nad wpływem technologii na ocenę zdolności kredytowej.

Podstawowe metody oceny zdolności kredytowej

Zdolność kredytowa odzwierciedla wiarygodność finansową jednostki lub przedsiębiorstwa i jest oceniana na podstawie szeregu czynników ilościowych

oraz jakościowych. Współcześnie kluczowe znaczenie mają elementy takie jak historia kredytowa, poziom dochodów, stabilność zatrudnienia, wskaźniki zadłużenia, a także dodatkowe informacje z alternatywnych źródeł danych, np. zachowania konsumenckie, lokalne wskaźniki ekonomiczne czy nawet aktywność w mediach społecznościowych. Ocena ta obecnie wspierana jest zaawansowanymi modelami statystycznymi i uczenia maszynowego. Jednakże wcześniej proces ten miał jednak zupełnie inny charakter – bazował głównie na reputacji klienta, referencjach osobistych, ocenie majątku trwałego (np. ziemi czy nieruchomości) oraz subiektywnym osądzie kredytodawcy.

Badanie Brennecke (2016), oparte na raportach R. G. Dun & Co., ukazuje wczesny etap rozwoju instytucjonalnej oceny historii kredytowej w XIX-wiecznych Stanach Zjednoczonych – przejście od lokalnej, osobistej reputacji kupca do systematycznego, opartego na danych monitorowania wiarygodności finansowej. W latach 1850-1860 informacja kredytowa zaczęła nabierać charakteru zorganizowanego dobra rynkowego: raporty o firmach były regularnie aktualizowane i dostępne dla subskrybentów, co pozwalało tworzyć wczesne profile ryzyka. Jednak, jak pokazuje analiza zachowań użytkowników, informacje te wykorzystywano głównie w reakcji na pogorszenie warunków rynkowych lub negatywne sygnały o dłużnikach, a nie w celu stałej preselekcji. W rezultacie proces oceny kredytowej w omawianym okresie miał charakter reaktywny i kontracykliczny – opierał się na weryfikacji po zdarzeniach, a nie na prewencyjnym screeningu. Ta ewolucja pokazuje, że historia kredytowa w połowie XIX w. była raczej zapisem epizodów niewypłacalności niż ciągłą miarą reputacji finansowej, a dopiero rozwój systematycznego raportowania (i późniejsze ustandaryzowanie punktacji) przekształcił ją w narzędzie predykcyjne używane do oceny ryzyka przed udzieleniem kredytu.

Historyczne badanie Duranda z 1941 roku położyło fundamenty pod nowoczesny credit scoring, wprowadzając po raz pierwszy systematyczne, oparte na danych metody analizy ryzyka kredytowego. Zastąpił subiektywne oceny obiektywną analizą statystyczną, co było rewolucją w branży finansowej lat 40. XX wieku i stało się podstawą dla wszystkich późniejszych rozwojów w dziedzinie oceny ryzyka kredytowego. Autor wskazuje, że kluczowym czynnikiem determinującym ryzyko niewypłacalności jest poziom dochodu kredytobiorcy, przy czym znaczenie mają również takie elementy jak wysokość kredytu, okres spłaty, wkład

własny czy wartość zabezpieczenia. Z czynników niefinansowych największą rolę odgrywają stabilność zatrudnienia i miejsca zamieszkania. Durand opracował narzędzia statystyczne w postaci współczynników efektywności, które pozwalają ocenić wartość informacyjną poszczególnych zmiennych i priorytetyzować je w procesie decyzyjnym. Ważnym wkładem są także proste formuły rankingowe, które mimo swojej prostoty dają wyniki porównywalne ze złożonymi metodami. Zatem Durand pokazuje tym samym, że systematyczna analiza statystyczna przewyższa subiektywną ocenę kredytodawców i może znacząco ograniczyć straty instytucji finansowych, wyznaczając kierunek dla dalszego rozwoju metod ilościowych w ocenie ryzyka kredytowego.

System FICO powstał w drugiej połowie lat 50. XX wieku jako rezultat prac inżynierów i matematyków amerykańskiej firmy Fair, Isaac and Company, którzy przenieśli metody statystyczne i operacyjne, stosowane wcześniej w wojsku i ubezpieczeniach, do sfery masowego kredytowania konsumenckiego (Poon, 2012). Był to moment przełomowy w historii oceny kredytowej, ponieważ po raz pierwszy wprowadzono sformalizowany, ilościowy model punktowy, który pozwalał instytucjom finansowym podejmować decyzje kredytowe w sposób zautomatyzowany, porównywalny i skalowalny. W odróżnieniu od wcześniejszych, subiektywnych metod oceny opartej na reputacji lokalnej, rozmowach z klientem czy intuicji urzędnika, FICO stanowiło próbę standaryzacji zaufania w gospodarce opartej na danych. Jak pokazuje Poon (2012), przełomowość systemu nie wynikała jednak z jego rzekomej obiektywności technologicznej, lecz z faktu, że łączył techniczne modelowanie ryzyka z organizacyjnym uczeniem się instytucji finansowych. Scorecardy FICO były nie tylko narzędziem analitycznym, ale też mechanizmem koordynacji i dyscyplinowania praktyk kredytowych – umożliwiały masowe rozszerzenie rynku konsumenckiego kredytu przy jednoczesnym utrzymaniu kontroli nad ryzykiem.

Z kolei Myers i Forgy (1963) dokonali jednego z pierwszych systematycznych porównań metod ważenia predyktorów w numerycznych systemach oceny ryzyka kredytowego. Analiza została przeprowadzona na próbie 300 wniosków kredytowych (150 „dobrych" i 150 „złych") oraz na odrębnej próbie testowej. Autorzy zestawili skuteczność zarówno metod bardziej zaawansowanych, takich jak klasyczna analiza dyskryminacyjna czy krokowa regresja wieloraka, jak i podejść prostszych, np. przypisania równych wag wszystkim 21 zmiennym. Wyniki wskazały na

porównywalną zdolność rozróżniania kredytobiorców przez obie grupy metod, co dowodzi, że złożone modele nie zawsze przewyższają proste rozwiązania.

W realiach poprzedniego wieku, przy ograniczonych możliwościach obliczeniowych i skromnych zasobach danych, trudno byłoby o lepszą i bardziej systematyczną analizę. To właśnie te badania najlepiej pokazują, jak stopniowo odchodzono od intuicyjnych, subiektywnych ocen kredytodawców w stronę podejścia ilościowego, które stało się standardem w następnych dekadach. Dlatego proponuję traktować je jako punkt wyjścia – zamykający etap wczesnych prób analitycznych i otwierający drzwi do późniejszych badań, opartych już na coraz bardziej zaawansowanych metodach statystycznych i uczenia maszynowego.

Przegląd najnowszych rozwiązań

W ciągu ostatnich czterech dekad obserwujemy wyraźne fale popularności różnych metod analitycznych i uczenia maszynowego. W latach 80. i 90. XX wieku dominowały klasyczne metody statystyczne, takie jak liniowa analiza dyskryminacyjna (LDA), wykorzystywane głównie do klasyfikacji i redukcji wymiarowości danych. W pierwszej dekadzie XXI wieku nastąpił wzrost zainteresowania metodami opartymi na uczeniu maszynowym, takimi jak SVM i Random Forest, które pozwalały na bardziej elastyczne i skuteczne modelowanie złożonych zależności w danych. Jednakże to ostatnie lata, szczególnie od 2020 roku, przyniosły przełom w obszarze dużych modeli językowych (LLM), które rewolucjonizują przetwarzanie języka naturalnego i wspierają szeroki zakres zastosowań od automatyzacji treści po systemy wspomagania decyzji. Te ewolucje metodologiczne znacząco wpływają na współczesne życie, dobrze wpasowując się do systemów finansowych i bankowych. W celu zrozumienia rozwoju i skuteczności różnych metod w kontekście oceny zdolności kredytowej, poniższy przegląd literatury koncentruje się na kluczowych technikach na przestrzeni ostatniej dekady.

W badaniu Shen i Wu (2025) nad predykcją ryzyka kredytowego przedsiębiorstw szczególną uwagę zwraca propozycja hybrydowego modelu TCN-DilateFormer, łączącego zalety sieci konwolucyjnych temporalnych (TCN), które dobrze wyłapują krótkoterminowe wzorce w danych, oraz architektur typu transformer (DilateFormer), które uwzględniają zależności długoterminowe. Autorzy opracowali rozwiązanie ukierunkowane na poprawę dokładności prognozowania

niewypłacalności oraz przezwyciężenie problemu silnie niezbalansowanych danych (stosunek przypadków wypłacalnych do niewypłacalnych wynosił ok. 49:1).

Badanie oparto na danych finansowych 3762 spółek notowanych na rynku chińskim (lata 2012-2022), z czego wykorzystano 132 wskaźniki z sześciu obszarów: wypłacalności, operacyjności, rentowności, rozwoju, struktury wskaźników oraz przepływów pieniężnych. Zastosowano zarówno techniki augmentacji próbek pozytywnych (powielanie, dodawanie szumu Gaussa), jak i downsamplingu próbek negatywnych, co znacząco ograniczyło ryzyko stronniczości modelu. Rezultaty eksperymentalne wskazują, że TCN-DilateFormer przewyższył zarówno tradycyjne metody uczenia maszynowego, jak i inne modele hybrydowe (m.in. CNN-LSTM, CNN-Transformer, ConvTimeNet), osiągając parametry:

- precision = 0.9681 – około 97% firm oznaczonych przez model jako ryzykowne były faktycznie ryzykowne,
- recall = 0.8689 – około 87% faktycznych bankructw zostało wykrytych przez model,
- specificity = 0.9809 – około 98% poprawnie zostało zakwalifikowanych jako bezpieczne.

Test Wilcoxona potwierdził istotność statystyczną wyników, a analiza wrażliwości dodatkowo wykazała stabilność modelu przy zmianach parametrów. W ujęciu praktycznym rozwiązanie to oferuje instytucjom finansowym i inwestorom bardziej precyzyjne narzędzie do oceny ryzyka kredytowego i zarządzania portfelem. Autorzy wskazują na możliwość rozszerzenia podejścia na inne kategorie ryzyka (rynkowe, operacyjne) oraz integrację z dodatkowymi źródłami danych, np. makroekonomicznymi i tekstowymi.

W podobny sposób Xing i in. (2024) przedstawili model zespołowy (ensemble) do przewidywania oceny kredytowej, integrujący trzy różne algorytmy uczenia maszynowego: Random Forest, XGBoost oraz TabNet. Autorzy podjęli problem niezbalansowanych danych w klasyfikacji ryzyka kredytowego, stosując szereg technik przetwarzania, w tym m.in. oversampling oraz metodę SMOTE-ENN – czyli tworzenie nowych sztucznych próbek klasy mniejszościowej (tutaj klasy poor) o cechach pośrednich, jednocześnie czyszcząc zbiór poprzez usunięcie takich punktów, które wyglądają na błędne. Kluczową innowacją było zastosowanie modelowania zespołowego w podejściu stackingowym, w którym Random Forest pełnił rolę meta-modelu integrującego wyniki poszczególnych klasyfikatorów.

Badanie przeprowadzono na zbiorze danych Credit Score Classification (100 000 rekordów, trzy poziomy oceny kredytowej: good, standard, poor). Wyniki empiryczne wykazały, że model zespołowy znacząco przewyższył wydajnością modele pojedyncze, osiągając po zastosowaniu SMOTE-ENN wskaźniki F1-score = 0.7968 (średnia między precision a recall) oraz ROC AUC = 0.9172 (umiejętność modelu w odróżnianiu klasy pozytywnej od negatywnej). Poprawa ta potwierdziła skuteczność zarówno samego podejścia zespołowego, jak i zaawansowanych technik balansowania danych. Autorzy podkreślili, że model zapewnia równowagę między precision i recall, co ma istotne znaczenie w ocenie ryzyka kredytowego, redukując jednocześnie zarówno błędy fałszywie pozytywne, jak i fałszywie negatywne.

Stąpor (2016) przeprowadziła porównanie maszyn wektorów nośnych (z jądrem Gaussa), zestawiając je z heteroskedastyczną analizą dyskryminacyjną (HLDA). Tradycyjna LDA zakłada, że dane w każdej grupie mają rozkład normalny i że wszystkie grupy mają taką samą macierz kowariancji. W praktyce to założenie często jest fałszywe - w grupie osób z problemami finansowymi dane są dużo bardziej „rozproszone" niż w grupie osób stabilnych finansowo, a heteroskedastyczność pozwala to ograniczenie złamać. Badanie oparto na zbiorze danych kredytowych obejmującym 1 000 obserwacji (700 klientów wypłacalnych i 300 niewypłacalnych) oraz 20 cech opisujących profil kredytobiorców. Wyniki wskazały na zbliżoną skuteczność obu podejść: HLDA osiągnęła średnią dokładność 75,10%, podczas gdy SVM uzyskał 76,10%. Różnice te nie były statystycznie istotne, co sugeruje porównywalny potencjał obu metod w ocenie ryzyka kredytowego. Jednocześnie podkreślono, że SVM wiąże się ze znacznie większym kosztem obliczeniowym, ze względu na przykładowo intensywne strojenie hiperparametrów lub większą liczbę wektorów wspierających, podczas gdy HLDA zapewnia prostsze i mniej zasobochłonne rozwiązanie. Autorka rekomenduje HLDA w scenariuszach wymagających efektywności obliczeniowej i prostoty wdrożenia, natomiast SVM jest korzystnym wyborem w warunkach dostępności większych zasobów obliczeniowych i potrzeby maksymalizacji jakości klasyfikacji.

Norling i Abdu (2023) przeanalizowali zastosowanie regresji logistycznej do prognozowania zdolności kredytowej w oparciu o dane ze szwedzkiego instytutu kredytowego, obejmujące około 15 tys. wniosków kredytowych z lat 2020-2022. Modele były budowane dla różnych horyzontów spłaty (6-21 miesięcy),

z wykorzystaniem inżynierii cech (m.in. dyskretyzacja zmiennych ciągłych, selekcja zmiennych wsteczna) oraz kontroli kolinearności. Wyniki pokazały, że regresja logistyczna przewyższa klasyfikator losowy, lecz charakteryzowała się niską czułością wobec klientów niewypłacalnych (AUC ok. 0,290,31). Istotne predyktory zmieniały się w zależności od horyzontu czasowego – krótkoterminowo największe znaczenie miały wcześniejsze zaległości, a długoterminowo informacje o zatwierdzonych kredytach konsumenckich i dochodach.

Z kolei Mestiri (2024) prezentuje systematyczne porównanie tradycyjnych modeli statystycznych, technik uczenia maszynowego oraz metod głębokiego uczenia. Autor uwzględnił sześć podejść: LDA, regresję logistyczną, drzewa decyzyjne, Random Forest (RF), SVM oraz głębokie sieci neuronowe (DNN). Analiza została oparta na próbie 688 obserwacji opisanych dwunastoma zmiennymi finansowymi klientów. Wyniki badania jednoznacznie wskazały przewagę algorytmów uczenia maszynowego i głębokiego uczenia nad klasycznymi technikami statystycznymi. Szczególnie wyróżniły się RF, SVM oraz DNN, które wykazały wyższą zdolność predykcyjną i lepsze dopasowanie do danych.

Bardzo podobne wnioski popiera praca Mbonigaba (2025). Przeprowadził badanie nad skutecznością różnych technik regresji statystycznej w ocenie ryzyka kredytowego w finansach konsumenckich. Autor wykorzystał dane wtórne z lat 2020-2024 i zastosował ilościowy projekt badawczy, porównując klasyczne metody statystyczne z bardziej zaawansowanymi podejściami uczenia maszynowego. W zakresie metodologii uwzględniono regresję logistyczną, regresję liniową wielokrotną, SVM oraz sieci neuronowe. Analiza wykazała, że choć regresja logistyczna pozostaje solidnym narzędziem w tradycyjnym scoringu kredytowym (dokładność 85,3%, AUC = 0,78), to zaawansowane modele osiągnęły istotnie wyższą skuteczność predykcyjną – SVM (92,3%) i sieci neuronowe (91,1%). Badanie pozwoliło także zidentyfikować kluczowe predyktory zdolności kredytowej: długość historii kredytowej (korelacja z niewypłacalnością -0,74), poziom dochodów (-0,52) oraz wiek (-0,26). Ogólny współczynnik korelacji (-0,62) potwierdził silną odwrotną zależność między zmiennymi predykcyjnymi a ryzykiem kredytowym. Wnioski podkreślają, że choć metody tradycyjne zachowują swoją użyteczność, to modele hybrydowe i integracja uczenia maszynowego wyraźnie poprawiają dokładność prognoz. Autor rekomenduje m.in. włączenie alternatywnych źródeł danych kredytowych, regularne aktualizowanie modeli oraz

stosowanie strategii przeciwdziałania stronniczości, aby zwiększyć sprawiedliwość decyzji kredytowych.

Z kolei praca autorstwa Golec i Alabduljalil (2025) stanowi pierwszą próbę kompleksowego ujęcia zastosowań dużych modeli językowych (LLM) w kontekście szacowania ryzyka kredytowego. Nie nadają się do oceny ryzyka kredytowego, ponieważ zostały zaprojektowane do przetwarzania i generowania języka naturalnego, a nie do precyzyjnej analizy danych finansowych i szacowania prawdopodobieństwa niewypłacalności. Jednakże autorzy spojrzeli na problem z perspektywy analizy dokumentów finansowych. Bazując na metodologii PRISMA, dokonali analizy 60 publikacji z lat 2020-2025, co pozwoliło im wyodrębnić cztery główne wymiary klasyfikacji:

- architektury modeli – jaki rodzaj modelu należy zastosować?
- modalności danych – jaki rodzaj danych? Przykładowo: tradycyjne dane jak zarobki czy historia kredytowa, teksty finansowe takie jak raporty czy dokumenty firm oraz dane o zachowaniu klienta w postaci formy korzystania z konta bankowego.
- mechanizmy interpretowalności – jak wytłumaczyć decyzję modelu?
- domena zastosowań – do czego można wykorzystywać LLMy?

W obszarze architektur modeli zidentyfikowano zarówno rozwiązania bazujące na klasycznych enkoderach (np. FinBERT), jak i dekoderach generatywnych (GPT-3.5, GPT-4), a także podejścia hybrydowe i specjalistyczne modele finansowe (FinGPT, FinLLaMA). Analiza modalności danych wskazała na stopniowe odchodzenie od wyłącznie danych strukturalnych na rzecz nieustrukturyzowanych tekstów finansowych czy szeregów czasowych. W zakresie interpretowalności dominują metody post hoc (takie jak SHAP, LIME), jednak rośnie zainteresowanie wyjaśnieniami generowanymi bezpośrednio w procesie wnioskowania oraz projektowaniem modeli wewnętrznie transparentnych. Jeśli chodzi o domeny zastosowań, prace koncentrują się na ocenie zdolności kredytowej, analizie nastrojów rynkowych, a także wykrywaniu oszustw i przeciwdziałaniu praniu pieniędzy. Autorzy wskazują na rosnącą popularność hybryd LLM-ów z klasycznymi metodami uczenia maszynowego oraz dominację metod post hoc. Jednocześnie autorzy identyfikują istotne luki badawcze:

- brak rozwiniętych technik interpretowalności przyczynowej – modele funkcjonują jak „czarne skrzynki", tj. trudno zrozumieć jak dokładnie podejmują decyzje,
- ograniczoną reprodukowalność wyników – dane najczęściej są małe i bardzo zróżnicowane, by móc odtworzyć te same wyniki,
- podatność modeli na stronniczość i halucynacje – niektóre modele mogą być stronnicze wobec pewnych grup społecznych,
- niedobory w zakresie ujednoliconego benchmarkingu – brakuje wspólnych standardów do porównywania wyników oceny zdolności kredytowej.

Figini i Giudici (2017) zaproponowali zastosowanie Bayesowskiego Łączenia Wiedzy (BMA) w ocenie ryzyka kredytowego jako alternatywy dla tradycyjnej regresji logistycznej opartej na pojedynczym modelu. Autorzy wskazali, że w warunkach braku jednoznacznych wskazówek teoretycznych dotyczących doboru zmiennych predykcyjnych, BMA umożliwia formalne uwzględnienie niepewności modelowej poprzez łączenie predykcji wielu konkurencyjnych modeli w średnią ważoną ich prawdopodobieństw posteriori. Wyniki empiryczne potwierdziły, że podejście to zapewnia lepszą kalibrację i stabilność prognoz, a także nieznacznie wyższą wartość AUC w porównaniu z najlepszym pojedynczym modelem regresji logistycznej.

Ocena zdolności kredytowej w Polsce

Biuro Informacji Kredytowej (BIK) zostało utworzone w 1997 roku z inicjatywy Związku Banków Polskich i 21 największych banków w kraju jako odpowiedź na problemy z niewystarczającym przepływem informacji w polskim sektorze bankowym po transformacji ustrojowej. W pierwszej połowie lat 90. brak dostępu do danych o historii kredytowej klientów prowadził do wysokiego udziału należności zagrożonych, sięgającego często ponad 25%, co osłabiało stabilność systemu finansowego. Celem powołania BIK było ograniczenie ryzyka kredytowego, zwiększenie bezpieczeństwa obrotu pieniężnego oraz zapobieganie nadmiernemu zadłużeniu poprzez systematyczne gromadzenie, przetwarzanie i udostępnianie informacji o zobowiązaniach kredytowych klientów banków i instytucji finansowych. Jak podkreśla Związek Banków Polskich (2013), utworzenie BIK umożliwiło wprowadzenie nowoczesnych mechanizmów oceny

zdolności kredytowej i znacząco poprawiło efektywność oraz bezpieczeństwo krajowego rynku kredytowego.

Według danych Biura Informacji Kredytowej (2019) średni wynik punktowy BIK w Polsce na koniec czerwca 2019 roku wynosił 79 punktów, co wskazuje na relatywnie wysoki poziom wiarygodności kredytowej wśród polskich konsumentów. Ponad połowa kredytobiorców (ok. 7 mln osób) uzyskała ocenę w najwyższym przedziale 80-100 punktów, co oznacza niskie ryzyko kredytowe i wysokie prawdopodobieństwo terminowej spłaty zobowiązań. Wynik BIK opiera się na czterech podstawowych komponentach: terminowości spłat, stopniu wykorzystania limitów kredytowych, liczbie zapytań o kredyt oraz długości historii kredytowej (wraz ze zgodą na przetwarzanie danych). Analizy przestrzenne wskazują na koncentrację najwyższych ocen w dużych miastach oraz w regionach o stabilnej strukturze gospodarczej, m.in. na Podkarpaciu (Rzeszów – 82,2; Krosno – 81,9). Wskaźnik scoringowy można było obliczyć dla ok. 86% aktywnych klientów (13,4 mln osób), co potwierdza wysoki poziom upowszechnienia systemu oceny kredytowej w Polsce. Dane te dowodzą, że scoring BIK stał się kluczowym narzędziem oceny ryzyka kredytowego i elementem infrastruktury finansowej kraju.

Ważnym elementem do poruszenia jest również Rekomendacja T stanowi ona dokument regulacyjny opracowany przez Komisję Nadzoru Finansowego (KNF), którego celem jest ujednolicenie i usystematyzowanie zasad zarządzania ryzykiem detalicznych ekspozycji kredytowych w polskim sektorze bankowym. Została wprowadzona w lutym 2010 roku jako reakcja na globalny kryzys finansowy i niewystarczające zabezpieczenia systemów oceny ryzyka kredytowego wynikające z wcześniejszych ram Bazylea II. Jej aktualna wersja (KNF, 2013) określa dobre praktyki w zakresie identyfikacji, pomiaru, monitorowania i kontroli ryzyka kredytowego dla banków działających w Polsce, ze szczególnym uwzględnieniem kredytów konsumenckich i innych detalicznych form zadłużenia.

Zgodnie z dokumentem KNF (2013), Rekomendacja T nakłada na banki obowiązek prowadzenia pisemnej polityki ryzyka detalicznego, określania apetytu na ryzyko oraz limitu wskaźnika długu do dochodu. Wymaga również rozdziału funkcji sprzedażowych od analitycznych, wprowadzenia systematycznego monitoringu portfela kredytowego, testów warunków skrajnych, a także korzystania z zewnętrznych baz danych, takich jak Biuro Informacji Kredytowej (BIK)

i biura informacji gospodarczej. Dokument promuje zasadę konserwatywnej oceny zdolności kredytowej, przejrzystych relacji z klientem oraz proporcjonalności procedur do skali działalności banku. Jak zauważa Strycharczyk (2012), wprowadzenie Rekomendacji T było odpowiedzią na obserwowane w latach poprzednich zbyt liberalne praktyki kredytowe, takie jak wydłużanie okresów spłaty, zaniżanie wymogów dotyczących zabezpieczeń czy udzielanie kredytów osobom o niskiej zdolności kredytowej. Autorka podkreśla, że wdrożenie Rekomendacji przyniosło istotne korzyści w postaci poprawy jakości portfeli kredytowych, wzrostu przejrzystości procesów decyzyjnych i ochrony konsumentów przed nadmiernym zadłużeniem, ale jednocześnie wiązało się z pewnymi zagrożeniami. Do najczęściej wskazywanych należały: wzrost kosztów wdrożenia nowych procedur, ograniczony dostęp do kredytu dla gospodarstw o niższych dochodach oraz ryzyko przenoszenia części popytu na kredyty do sektora parabankowego.

W efekcie zarówno Biuro Informacji Kredytowej, jak i Rekomendacja T stanowią kluczowe narzędzia wspierające odpowiedzialne udzielanie kredytów, ograniczanie ryzyka finansowego i zwiększanie stabilności polskiego sektora bankowego. Z kolei w polskiej literaturze naukowej zagadnienie oceny zdolności kredytowej rozwijało się równolegle z trendami światowymi, choć często w większym stopniu uwzględniało lokalne uwarunkowania gospodarcze i instytucjonalne. Zatem w tym rozdziale zostaną omówione wybrane polskie prace naukowe, które pokazują, w jaki sposób definiuje się i bada zdolność kredytową w krajowym kontekście oraz jakie rozwiązania są proponowane w praktyce.

W pracy Juszczyk (2008) podjęto próbę systematycznego ujęcia oceny zdolności kredytowej małych i średnich przedsiębiorstw w praktyce banków komercyjnych i spółdzielczych w Polsce. Badanie oparto na wywiadach z kadrą kierowniczą i merytoryczną banków oraz analizie literatury i uregulowań prawnych. Autorzy wskazali, że zdolność kredytowa obejmuje zarówno aspekt formalno-prawny, jak i merytoryczny, przy czym ten drugi dzieli się na czynniki personalne (zaufanie do kredytobiorcy) i ekonomiczne (takie jak sytuacja finansowa, zabezpieczenia). Szczególną uwagę poświęcono różnorodnym metodom oceny: opisowej (listy kontrolne i klasy ryzyka), punktowej (system wag i oceny wskaźników), finansowej (analiza rentowności, płynności, zadłużenia i sprawności działania), dyskryminacyjnej (m.in. model Altmana), a także podejściom jakościowym, jak CAMPARI ICE czy 5C/6C. Wskazano również na znaczenie oceny

projektów inwestycyjnych z wykorzystaniem NPV, IRR czy progów rentowności. Wyniki podkreślają, że mali i średni przedsiębiorcy stanowią niejednorodną grupę kredytobiorców, co wymaga elastycznego doboru metod – od prostszych analiz w przypadku mniejszych kredytów po pełne procedury wskaźnikowe przy większych przedsięwzięciach. Autorzy rekomendują łączenie metod ilościowych i jakościowych, dostosowanie analizy do specyfiki branżowej oraz regularną aktualizację modeli, co stanowi fundament współczesnych praktyk oceny ryzyka kredytowego w Polsce.

Istotny wkład w dyskusję nad samą problematyką oceny zdolności kredytowej wniosła praca Nowaka (2017), której celem była identyfikacja głównych nieprawidłowości w tym procesie w odniesieniu do klientów detalicznych. Autor, opierając się na analizie raportów organów nadzoru, literatury oraz regulacji prawnych, wskazał na liczne uchybienia zarówno po stronie banków, jak i kredytobiorców. Po stronie instytucji finansowych najczęściej powtarzającymi się problemami były m.in. niedoszacowanie ryzyka stopy procentowej, zaniżanie kosztów utrzymania gospodarstw domowych, nadmierne wydłużanie okresu kredytowania czy zbyt liberalne zasady przyznawania kredytów. Z kolei konsumenci nierzadko wykazywali brak rzetelności (np. poprzez zatajenie istotnych informacji) oraz deficyty kompetencji finansowych, czego przejawem była m.in. niechęć do czytania umów kredytowych przed ich podpisaniem. Autor zwrócił także uwagę na konieczność działań naprawczych, zarówno po stronie regulatorów (np. rekomendacje KNF i EBA), jak i rynku, w tym edukacji finansowej konsumentów. Wnioski tej pracy ukazują, że proces oceny zdolności kredytowej ma charakter systemowy i wymaga kompleksowych działań, obejmujących jednocześnie banki, konsumentów oraz organy nadzoru.

W literaturze polskiej na uwagę zasługuje praca Czech (2021), która podjęła próbę oceny ryzyka kredytowego Polski w okresie pandemii COVID-19 na podstawie ubezpieczenia SCDS na wypadek niewypłacalności państwa. Badanie objęło okres od marca 2019 do lutego 2021 i wykorzystało zarówno analizę dynamiki, jak i modele regresji liniowej z uwzględnieniem czynników makroekonomicznych oraz nieekonomicznych. Wyniki wskazały, że w początkowej fazie pandemii spready SCDS wzrosły, jednak z czasem powróciły do trendu spadkowego, mimo wzrostu długu publicznego i pogorszenia wskaźników gospodarczych. Szczególnie interesujące jest odkrycie, iż determinanty ryzyka

kredytowego różniły się przed i w trakcie pandemii – w okresach stabilnych dominowały zmienne makroekonomiczne, natomiast w czasie kryzysu istotną rolę odgrywały również czynniki związane z przebiegiem pandemii, takie jak liczba zgonów czy zakażeń. Autorka dochodzi do wniosku, że spready SCDS dobrze odzwierciedlają poziom ryzyka kredytowego w okresach równowagi, jednak ich użyteczność w warunkach niepewności generowanej przez czynniki pozaekonomiczne jest ograniczona.

Polski rynek kredytowy i jego badanie ukazują złożoność zagadnienia oceny zdolności kredytowej zarówno na poziomie mikro, jak i makro. Badacze podkreślają, że ocena ta wymaga podejścia wielowymiarowego, łączącego analizę ilościową i jakościową, a także uwzględniającego czynniki instytucjonalne i koniunkturalne. Z jednej strony wskazuje się na potrzebę stosowania różnorodnych metod i dostosowania ich do specyfiki kredytobiorcy, z drugiej na ryzyka wynikające z nieprawidłowości proceduralnych czy asymetrii informacji. Jednocześnie najnowsze badania dowodzą, że tradycyjne narzędzia oceny ryzyka kredytowego mogą tracić skuteczność w warunkach kryzysów o charakterze pozaekonomicznym, co potwierdza konieczność ciągłej aktualizacji i poszerzania metodologii.

Podsumowanie i wnioski

Według przeglądu literatury Markova i in. (2022) literatura nie dostarcza jednoznacznej odpowiedzi na to, która metoda jest najbardziej efektywna. Wykazano natomiast, że w liczbie zastosowań dominuje głównie regresja logistyczna, a dopiero na kolejnych miejscach znajdują się SVM i sieci neuronowe. Z niniejszego przeglądu literatury również nie można jednoznacznie określić, która metoda oceny zdolności kredytowej jest najskuteczniejsza, choć z analizy przytoczonych badań wynika, że regresja logistyczna, mimo iż pozostaje solidnym narzędziem bazowym w credit scoringu, jest generalnie mniej skuteczna niż zaawansowane metody uczenia maszynowego i głębokiego uczenia.

Mimo dynamicznego rozwoju metod ilościowych i technologii AI, wciąż istnieje wiele obszarów wymagających pogłębionych badań. Po pierwsze, konieczne jest lepsze zrozumienie wpływu nietypowych i nieustrukturyzowanych źródeł danych – takich, jak aktywność w mediach społecznościowych czy dane behawioralne z urządzeń mobilnych – na trafność i uczciwość oceny kredytowej,

zwłaszcza w kontekście prywatności i ochrony danych. Po drugie, brakuje spójnych ram do oceny ryzyka algorytmicznego i modelowego, w tym standardów interpretowalności oraz mechanizmów wykrywania i korygowania stronniczości czy „halucynacji" w modelach uczenia maszynowego. Trzecim niewyjaśnionym obszarem pozostaje integracja metod predykcyjnych z systemami decyzyjnymi banków w sposób, który zapewni zarówno zgodność z regulacjami (np. Rekomendacja T, Basel III), jak i elastyczność operacyjną. Wreszcie, wymaga się badań nad skutecznością hybrydowych modeli człowiek-maszyna, które określą, jak najlepiej łączyć intuicyjne, jakościowe oceny ekspertów z precyzyjnymi, ale złożonymi algorytmami, aby maksymalizować zarówno trafność decyzji kredytowych, jak i ochronę konsumentów.

W Polsce ocena zdolności kredytowej wciąż pozostaje hybrydą metod tradycyjnych i nowoczesnych. Z jednej strony mocno zakorzenione są rozwiązania regulacyjne i jakościowa ocena klienta, z drugiej strony coraz wyraźniej widać wpływ technologii i modeli ilościowych. W artykule podkreślono, że największym wyzwaniem nie jest dziś brak narzędzi, ale odpowiedzialne ich stosowanie oraz równoległa edukacja finansowa społeczeństwa, bo tylko wtedy ocena zdolności kredytowej będzie rzetelna i wiarygodna. Należy zaznaczyć, że nie można całkowicie wyeliminować czynnika ludzkiego – doświadczenie, intuicja i możliwość uwzględnienia kontekstu poza surowymi danymi nadal są nieocenione. Zaleca się, zatem model hybrydowy człowiek-maszyna, w którym system podejmuje decyzję kredytową wyłącznie w sytuacjach jednoznacznych, gdy wszystkie kluczowe wskaźniki wskazują na ten sam kierunek. W przypadkach granicznych lub niespójnych ostateczne rozstrzygnięcie pozostaje w gestii człowieka.

Przytoczone argumenty wskazują jednoznacznie, że ogólny wpływ technologii na ocenę zdolności kredytowej jest zdecydowanie pozytywny. Umożliwia szybsze, bardziej kompleksowe i dokładniejsze prognozowanie ryzyka, a także włączenie nowych źródeł danych do analizy. Jednocześnie zwrócono uwagę, że nadal pozostaje ogromne pole do rozwoju, szczególnie w zakresie przejrzystości modeli i lepszej interpretowalności wyników. Powinniśmy dążyć do systemu, w którym kluczowe znaczenie zachowa historia kredytowa i dane twarde, ponieważ to one w największym stopniu odzwierciedlają rzeczywistą wiarygodność finansową jednostki.

Literatura

1. Brennecke, C. (2016). Information Acquisition in Antebellum U.S. Credit Markets: Evidence from Nineteenth-Century Credit Reports. *FDIC Center for Financial Research Working Paper No. 2016-04.*
2. Czech, M. (2021). Assessment of the credit risk of Poland based on sovereign credit default swap spreads during the COVID-19 pandemic. *Ekonomia i Prawo. Economics and Law, 20*(3), 497-511. https://doi.org/10.12775/EiP.2021.030
3. Durand, D. (1941). *Risk elements in consumer instalment financing*. New York: National Bureau of Economic Research.
4. Figini, S., Giudici, P. (2017). Credit risk assessment with Bayesian model averaging. *Communications in Statistics – Theory and Methods, 46*(19), 9507-9517. https://doi.org/10.1080/03610926.2016.1212070
5. Golec, M., AlabdulJalil, M. (2025). Interpretable LLMs for credit risk: A systematic review and taxonomy (arXiv:2506.04290v2) [Preprint]. *arXiv*. https://doi.org/10.48550/arXiv.2506.04290
6. Juszczyk, S. (2008). Zdolność kredytowa w ocenie banku na przykładzie małych i średnich przedsiębiorstw. *Ekonomiczne i Organizacyjne Problemy Bankowości, 66*, 17-27. Pobrane z: https://sj.wne.sggw.pl/pdf/EIOGZ_2008_n66_s17.pdf
7. Markov, A., Seleznyova, Z., Lapshin, V. (2022). Credit scoring methods: Latest trends and points to consider. *The Journal of Finance and Data Science, 8,* 180-201. https://doi.org/10.1016/j.jfds.2022.07.002
8. Mbonigaba, C. (2025). Evaluating credit scoring models with statistical regression techniques in consumer finance (Version March 01, 2025). *SSRN*. https://doi.org/10.2139/ssrn.5188111
9. Mestiri, S. (2024). Credit scoring using machine learning and deep learning-based models. *Data Science in Finance and Economics, 4*(2), 236-248. https://doi.org/10.3934/DSFE.2024009
10. Myers, J. H., Forgy, E. W. (1963). The development of numerical credit evaluation systems. *Journal of the American Statistical Association, 58*(303), 799-806. https://doi.org/10.2307/2282727
11. Nowak, P. (2017). The incorrectness in the assessment procedure of creditworthiness. *Journal of Finance and Financial Law, 2*(14), 53-65. https://doi.org/10.18778/2391-6478.2.14.05

12. Shen, C., Wu, J. (2025). Research on credit risk of listed companies: A hybrid model based on TCN and DilateFormer. *Scientific Reports, 15*, 2599. https://doi.org/10.1038/s41598-025-86371-7
13. Stąpor, K. (2016). *A critical comparison of discriminant analysis and SVM-based approaches to credit scoring*. Katowice: Wydawnictwo Uniwersytetu Ekonomicznego w Katowicach. Pobrane z: http://sbc.org.pl/Content/264089/PDF/05.pdf
14. Strycharczyk, A. (2012). Rekomendacja T – szanse i zagrożenia. *Problemy Zarządzania, Finansów i Marketingu, 27*, 545-554. Pobrane z: https://bazhum.muzhp.pl/media/texts/problemy-zarzadzania-finansow-i-marketingu/2012-tom-27/problemy_zarzadzania_finansow_i_marketingu-r2012-t27-s545-554.pdf
15. Xing, Q., Yu, C., Huang, S., Zheng, Q., Mu, X., Sun, M. (2024). Enhanced credit score prediction using an ensemble deep learning model. *arXiv*. https://doi.org/10.48550/arXiv.2410.00256

Netografia

1. Biuro Informacji Kredytowej. (2019). *BIK: Prawie 7 mln Polaków ma doskonały scoring*. Informacja prasowa. Pobrane z: https://media.bik.pl/informacje-prasowe/454131/bik-prawie-7-mln-polakow-ma-doskonaly-scoring (dostęp: 01.09.2025).
2. Komisja Nadzoru Finansowego. (2013). *Rekomendacja T dotycząca dobrych praktyk w zakresie zarządzania ryzykiem detalicznych ekspozycji kredytowych*. Pobrane z: https://www.knf.gov.pl/knf/pl/komponenty/img/rekomendacja_t_%2814_09_2018%29_63160.pdf (dostęp: 01.09.2025).
3. Norling, J., Abdu, S. (2023). From data to decision: Using logistic regression to determine creditworthiness (Bachelor's thesis, KTH Royal Institute of Technology). *KTH Royal Institute of Technology.* Pobrane z: https://kth.diva-portal.org/smash/record.jsf?pid=diva2:1833653 (dostęp: 01.09.2025).
4. Poon, M. A. (2012). What lenders see – A history of the Fair Isaac scorecard [Doctoral dissertation, University of California, San Diego]. *eScholarship.* Pobrane z: https://escholarship.org/uc/item/7n1369x2 (dostęp: 01.09.2025).
5. Związek Banków Polskich. (2013). *Jak powstało BIK*. Biuro Informacji Kredytowej. Pobrane z: https://www.skef.pl/wp-content/uploads/2013/07/Jak-powstało-BIK.pdf (dostęp: 01.09.2025).

Ewelina Melaniuk[1]

Akademia Bialska im. Jana Pawła II, Biała Podlaska

https://doi.org/10.29316/9788368103304_7

WYKORZYSTANIE TECHNOLOGII AI W KREACJI WIRTUALNYCH INFLUENCERÓW

THE USE OF AI TECHNOLOGY IN THE CREATION OF VIRTUAL INFLUENCERS

Streszczenie

Wirtualni influencerzy, czyli cyfrowe postacie aktywne w mediach społecznościowych, stanowią jeden z najbardziej dynamicznie rozwijających się trendów w marketingu cyfrowym. Rosnąca popularność tego zjawiska wynika z postępu w dziedzinie sztucznej inteligencji (AI), grafiki 3D oraz generatywnych sieci neuronowych (GAN). W rozdziale dokonano analizy procesu tworzenia wirtualnych influencerów, przedstawiono przykłady globalne i polskie oraz zidentyfikowano kluczowe technologie i wyzwania etyczne związane z tym zjawiskiem. Wnioski płynące z badań potwierdzają duży potencjał marketingowy wirtualnych postaci, podkreślając przy tym znaczenie spójnej narracji i świadomości wyzwań etycznych.

Słowa kluczowe: wirtualny influencer, sztuczna inteligencja (AI), media społecznościowe, marketing cyfrowy

[1] https://orcid.org/0009-0002-9434-6258, e-mail: e.melaniuk@akademiabialska.pl

Abstract

Virtual influencers, or digital characters active on social media, are among the fastest-growing trends in digital marketing. The growing popularity of this phenomenon is attributed to advances in artificial intelligence (AI), 3D graphics, and generative adversarial networks (GANs). The chapter analyses the process of creating virtual influencers, presents global and Polish examples, and identifies the key technologies and ethical challenges associated with this phenomenon. The research conclusions confirm the significant marketing potential of virtual characters, while emphasising the importance of coherent narration and awareness of ethical challenges.

Keywords: virtual influencer, artificial intelligence (AI), social media, digital marketing

Wstęp

Wirtualni influencerzy, czyli postacie generowane komputerowo, stanowią jeden z najbardziej dynamicznie rozwijających się trendów w marketingu cyfrowym. Rosnąca popularność tego zjawiska wynika z postępu w dziedzinie sztucznej inteligencji (AI), grafiki 3D oraz generatywnych sieci neuronowych (GAN) (Berryman, Kavka, 2018).

Początki fenomenu cyfrowych twarzy sięgają eksperymentów z silnikami graficznymi i animacją komputerową, które jeszcze w pierwszej dekadzie XXI wieku wykorzystywano głównie w branży rozrywkowej, filmach animowanych czy grach komputerowych. Rewolucja nadeszła wraz z udoskonaleniem metod syntezy obrazu, technik motion capture oraz rozwoju generatywnych sieci neuronowych (GAN), umożliwiających tworzenie postaci niemal nieodróżnialnych od prawdziwych ludzi (Goodfellow i in., 2014; Marwick, 2018).

Dzięki ciągłemu postępowi w dziedzinie AI wirtualni influencerzy mogą nie tylko prezentować realistyczny wizerunek i ruch, ale również dynamicznie reagować na odbiorców, prowadzić interakcję w czasie rzeczywistym czy personalizować treści pod kątem określonych grup docelowych (Floridi, 2020). Takie rozwiązania przekładają się na wzrost skuteczności strategii marketingowych oraz pozwalają twórcom i markom budować spójne, w pełni kontrolowane narracje w przestrzeni cyfrowej (Koles, 2024).

Jednym z czynników napędzających rosnące zainteresowanie wirtualnymi influencerami jest potrzeba innowacyjności w coraz bardziej konkurencyjnym środowisku mediów społecznościowych. Cyfrowe postacie nie wymagają przerw, są odporne na skandale czy wahania nastrojów, a ich działalność może być prowadzona 24/7 na wielu rynkach równocześnie. To również odpowiedź na rosnące oczekiwania odbiorców wobec personalizacji, autentyczności i różnorodności komunikacji.

Wyjątkowo interesujące jest to, że wirtualni influencerzy są w stanie aktywnie uczestniczyć w kształtowaniu wartości, norm społecznych czy nawet kreować specyficzne poczucie bliskości z odbiorcą, którego nie sposób osiągnąć za pomocą klasycznych technik reklamowych. Jednocześnie towarzyszą temu liczne pytania etyczne: gdzie przebiega granica pomiędzy rzeczywistością a fikcją? W jaki sposób obecność wirtualnych twarzy wpływa na autentyczność przekazu, zaufanie odbiorców oraz ich dobrostan psychiczny (Floridi, 2020; Gunkel, Wales, 2021)?

W Polsce rynek wirtualnych influencerów pozostaje młody, jednak wyraźnie widoczny jest rosnący potencjał innowacyjnych projektów, takich jak Sara Kosmos czy inicjatywy agencji Nakatomi. Sugeruje to, że rodzima branża digitalowa coraz śmielej podąża ścieżką wyznaczaną przez globalnych liderów.

Celem niniejszego rozdziału jest kompleksowa analiza procesu tworzenia oraz wykorzystania wirtualnych influencerów w praktyce marketingowej i komunikacyjnej, z uwzględnieniem uwarunkowań technologicznych, społeczno-kulturowych oraz etycznych, a także wskazanie kierunków rozwoju nowoczesnych technologii na świecie i w Polsce.

Metodologia

Przeprowadzono badania jakościowe, obejmujące:

1. Analizę literatury przedmiotu, raportów branżowych oraz publikacji naukowych z lat 2018-2025 dotyczących AI, influencerów i marketingu cyfrowego (Floridi, 2020).
2. Studium przypadków globalnych (Lil Miquela, Shudu, Imma), uwzględniających strategie komunikacyjne, estetykę treści i sposób interakcji z odbiorcą.

3. Studium przypadków polskich (Sara Kosmos, projekty Nakatomi), czyli ocenę obecności i odbioru wirtualnych postaci na rynku krajowym (SpidersWeb, 2023; Nakatomi, 2023).
4. Analizę porównawczą wyników krajowych i międzynarodowych wraz z aktualnymi trendami marketingowymi (metaverse, VR/AR, automatyzacja personalizacji treści) (McKinsey & Company, 2022).

Badanie miało charakter interdyscyplinarny, łącząc elementy komunikacji, marketingu, medioznawstwa i etyki technologicznej. Analiza treści została przeprowadzona zgodnie z metodologią jakościową (Krippendorff, 2018), co pozwoliło wyodrębnić kluczowe wzorce i wnioski dotyczące wirtualnych influencerów w skali globalnej i krajowej.

Wyniki i dyskusja

Technologie i możliwości

Grafika 3D stanowi fundament wizualny wirtualnych influencerów, umożliwiając tworzenie postaci o szczegółowych rysach twarzy, realistycznej mimice oraz bogatych teksturach skóry, włosów i ubrań. Zaawansowane modele 3D pozwalają na wysoką wierność wizualną oraz na precyzyjne odwzorowanie indywidualnych cech postaci takich, jak drobne zmarszczki, blizny czy szczególne gesty mimiczne, co zwiększa atrakcyjność postaci w odbiorze społecznościowym (Goodfellow i in., 2014). Dzięki temu odbiorcy mogą łatwiej utożsamiać się z wirtualną postacią, a jej wizerunek staje się bardziej wiarygodny i spójny.

Kolejnym istotnym elementem jest motion capture (mocap) – technika rejestracji ruchu aktorów, pozwalająca przenieść naturalne animacje na cyfrowe postacie. Mocap umożliwia tworzenie realistycznych ruchów ciała i mimiki twarzy, które są płynne, zsynchronizowane i zgodne z zachowaniami człowieka. Dzięki temu postacie wirtualne poruszają się w sposób naturalny i autentyczny, co zwiększa poczucie realności oraz ułatwia identyfikację emocjonalną odbiorców z postacią.

Podstawą generowania twarzy, sylwetek i animacji ruchu są również generative adversarial networks (GAN) – zaawansowane sieci neuronowe zdolne do tworzenia hiperrealistycznych obrazów na podstawie danych treningowych. GAN

umożliwiają dynamiczne kreowanie nowych wariantów wyglądu postaci, eksperymentowanie z różnymi stylami wizualnymi oraz generowanie realistycznych reakcji na interakcje użytkowników (Goodfellow i in., 2014). W praktyce oznacza to, że twórcy mogą łatwo modyfikować postać, wprowadzać nowe stroje, fryzury, mimikę, a wszystko to bez konieczności ręcznego modelowania każdej zmiany.

Integracja wirtualnych influencerów z sztuczną inteligencją (AI) pozwala im analizować reakcje odbiorców w czasie rzeczywistym, przewidywać preferencje użytkowników oraz automatycznie dostosowywać treści do indywidualnych potrzeb (Floridi, 2020; Huang, Rust, 2021). Dzięki AI cyfrowe postacie mogą reagować na komentarze, wiadomości prywatne, polubienia czy trendy w mediach społecznościowych, co zwiększa interaktywność i sprawia, że kontakt z nimi przypomina rozmowę z prawdziwą osobą.

Dodatkowo AI umożliwia personalizację kampanii marketingowych, dopasowując przekaz do wieku, lokalizacji, zainteresowań czy zachowań odbiorców. Jedna postać może prowadzić równolegle kilka spójnych, ale indywidualnie dostosowanych kampanii, co zwiększa efektywność działań marketingowych, optymalizuje zasięgi i jednocześnie redukuje koszty w porównaniu z tradycyjnymi influencerami.

Wirtualni influencerzy coraz częściej korzystają z narzędzi AR (rozszerzonej rzeczywistości) i VR (wirtualnej rzeczywistości), co umożliwia użytkownikom bezpośrednią interakcję w środowisku cyfrowym i zwiększa poziom immersji. Dzięki AR użytkownicy mogą doświadczać wirtualnych postaci w fizycznej przestrzeni, np. poprzez aplikacje mobilne pozwalające na wirtualne przymierzanie ubrań, nakładanie cyfrowych akcesoriów lub udział w interaktywnych kampaniach reklamowych (Javornik, 2016). Funkcja AR pozwala na spersonalizowane doświadczenie, ponieważ systemy mogą dopasowywać rozmiar, kolorystykę lub styl ubrań do indywidualnych cech użytkownika.

Technologie VR pozwalają natomiast na pełne zanurzenie użytkownika w środowisku wirtualnym, w którym może on wchodzić w interakcje z influencerem w czasie rzeczywistym. Przykłady zastosowań obejmują wspólne robienie zdjęć w wirtualnym otoczeniu, uczestnictwo w koncertach, premierach produktów, wydarzeniach online czy wirtualnych doświadczeniach gamifikacyjnych (Hollebeek i in., 2020; Kim i in., 2023). VR umożliwia tworzenie środowisk

całkowicie odmiennych od rzeczywistości, co zwiększa atrakcyjność i pozwala odbiorcom na przeżycie doświadczeń niemożliwych w świecie fizycznym.

Integracja AR i VR z technologiami 3D, motion capture (mocap), generative adversarial networks (GAN) i sztuczną inteligencją (AI) umożliwia wirtualnym influencerom realistyczny wygląd, płynne ruchy i autentyczne reakcje. Na przykład: wirtualna postać może reagować na gesty użytkownika, wyrażać emocje w odpowiedzi na komentarze lub dostosowywać treści kampanii marketingowych w czasie rzeczywistym, co zwiększa poczucie obecności i zaangażowanie (Flavián i in., 2019; Huang, Rust, 2021).

Dzięki tym połączeniom technologii odbiorcy zyskują bardziej angażujące i immersyjne doświadczenie, co wzmaga ich interakcję i przywiązanie do wirtualnego influencera. W rezultacie postacie cyfrowe stają się skutecznym narzędziem marketingowym, pozwalającym na personalizowane kampanie i długotrwałe budowanie lojalności odbiorców, w czym przewyższają tradycyjnych influencerów ograniczonych czasem, dostępnością czy fizycznym wyglądem.

Podsumowując, połączenie tych technologii umożliwia wirtualnym influencerom funkcjonowanie na poziomie zbliżonym do rzeczywistych osób, łącząc atrakcyjny wizerunek wizualny, realistyczne ruchy i dynamiczną interakcję z odbiorcami. Dzięki temu cyfrowe postacie stają się skutecznym i elastycznym narzędziem marketingowym w mediach społecznościowych, zdolnym do tworzenia spersonalizowanych, angażujących i immersyjnych doświadczeń dla użytkowników.

Przykłady globalne

Wirtualni influencerzy zdobywają coraz większą popularność na całym świecie, stając się aktywnymi uczestnikami kampanii marketingowych marek modowych, kosmetycznych, technologicznych i lifestyle'owych. Według raportu (Johnson i in., 2020), na świecie działa już ponad 150 wirtualnych influencerów o zasięgach powyżej 100 tysięcy obserwujących, a rynek ten rośnie w tempie ok. 25% rocznie (Statista, 2023). Dla młodych odbiorców są oni jednocześnie źródłem rozrywki i inspiracji – 41% respondentów w wieku 16-24 lat twierdzi, że ufa wirtualnym postaciom w podobnym stopniu, jak realnym influencerom (Palloff, 2007).

Poniżej przedstawiono najbardziej rozpoznawalne postacie cyfrowe oraz ich działania:

1. Lil Miquela (USA), stworzona przez firmę Brud, jest jedną z pierwszych i najbardziej znanych wirtualnych influencerek. Angażuje się w kampanie reklamowe dużych światowych marek takich, jak Calvin Klein, Prada czy Samsung, publikując treści w formie zdjęć, wideo oraz interaktywnych postów. Lil Miquelę wyróżnia umiejętność tworzenia narracji – postać promuje produkty oraz uczestniczy w dyskusjach społecznych, poruszając tematy różnorodności i inkluzywności, co zwiększa jej autentyczność w odbiorze społeczności. Liczba obserwujących: ~3,2 mln na Instagramie, przychody szacunkowe: 2-3 mln USD rocznie z kampanii marketingowych (Koles, 2024; Nakatomi, 2024).
2. Shudu Gram (Wielka Brytania) jest uważana za pierwszą cyfrową supermodelkę, która współpracuje z luksusowymi markami modowymi, w tym Balmain i Fenty Beauty. Została stworzona przez fotografa i twórcę cyfrowego. Jej wizerunek jest hiperrealistyczny, a postać zachowuje się w sposób charakterystyczny dla branży mody, co pozwala markom tworzyć spójne kampanie reklamowe bez ograniczeń czasowych i logistycznych związanych z udziałem tradycyjnych modeli. Liczba obserwujących: ~200 tys. na Instagramie, przychody szacunkowe: 500 tys. – 1 mln USD rocznie (Marwick, 2021; Nakatomi, 2024).
3. Imma (Japonia) zdobyła popularność dzięki swojemu unikalnemu, realistycznemu wyglądowi i aktywności w mediach społecznościowych. Postać ta pełni funkcję ambasadorki marek technologicznych i modowych, uczestniczy w sesjach fotograficznych oraz kampaniach promujących lifestyle. Imma wyróżnia się zdolnością do interakcji z użytkownikami w czasie rzeczywistym dzięki integracji z AI, co pozwala na spersonalizowane reakcje na komentarze i wiadomości. Liczba obserwujących: ~370 tys. na Instagramie przychody szacunkowe: ~400-600 tys. USD rocznie (Koles, 2024; Nakatomi, 2024).
4. Bermuda (USA) jest znana z prowokacyjnych i humorystycznych postów, które przyciągają młodszą grupę odbiorców. Liczba obserwujących: ~130 tys. na Instagramie, przychody szacunkowe: ~200 tys. USD rocznie. Knox Frost (USA) działa w segmencie lifestyle'u i sportu, prowadząc kampanie reklamowe dla marek odzieżowych i technologicznych. Liczba obserwujących:

~85 tys. na Instagramie, przychody szacunkowe: ~150 tys. USD rocznie. Obie postacie pokazują, że wirtualni influencerzy mogą być dostosowani do różnych grup docelowych i branż (Nakatomi, 2024).

5. Hatsune Miku (Japonia) to wirtualna piosenkarka stworzona w technologii Vocaloid, która zdobyła międzynarodową popularność jako idol muzyczny. Choć początkowo funkcjonowała głównie w przemyśle muzycznym, Miku uczestniczy również w kampaniach marketingowych i reklamowych marek technologicznych, odzieżowych oraz gier komputerowych. Ta wirtualna postać działa w social mediach, a także wchodzi w interakcję z fanami podczas koncertów VR i holograficznych wydarzeń. Liczba obserwujących: ~2,5 mln na platformach społecznościowych i YouTube, przychody szacunkowe: ~5-10 mln USD rocznie (koncerty VR, licencje muzyczne, współprace marketingowe) (Condry, 2013; Nakatomi, 2024).
6. CodeMiko (Korea/USA) jest wirtualną streamerką i influencerką działającą na platformie gamingowej Twitch w Korei oraz w USA. Postać wykorzystuje technologię motion capture w czasie rzeczywistym oraz AI do interakcji z widzami podczas transmisji na żywo. CodeMiko wyróżnia się dużym zaangażowaniem społeczności i kreatywnym wykorzystaniem wirtualnego avatara do gier, wywiadów i interaktywnych wydarzeń online, co pokazuje potencjał wirtualnych influencerów w branży rozrywki cyfrowej. Liczba obserwujących: ~900 tys. na Twitchu przychody szacunkowe: ~1,2-1,5 mln USD rocznie (donacje, subskrypcje, sponsoring) (Nakatomi, 2024).
7. AiAngel (USA) zdobyła popularność w mediach społecznościowych dzięki hiperrealistycznemu wyglądowi i aktywności w kampaniach lifestyle i fashion, opowiadając o zagadnieniach naukowych w przystępnej formie. Postać korzysta z AI do analizy zachowań odbiorców i dopasowywania treści do ich preferencji. AiAngel jest przykładem nowej generacji globalnych, wirtualnych postaci, które integrują marketing, interakcję w czasie rzeczywistym oraz kreację cyfrowego storytellingu. Liczba obserwujących: ~600 tys. na Instagramie, przychody szacunkowe: ~700 tys. – 1 mln USD rocznie (Nakatomi, 2024).

Globalni wirtualni influencerzy stosują szereg spójnych strategii marketingowych, które pozwalają im skutecznie konkurować z twórcami rzeczywistymi oraz budować wysokie zaangażowanie użytkowników. Po pierwsze, kluczowe

znaczenie ma regularne publikowanie spójnych treści wizualnych oraz storytellingowych, tworzących rozpoznawalną tożsamość postaci. Treści te są starannie planowane i kuratorowane, tak aby budować narrację, która wzmacnia poczucie autentyczności oraz ciągłości komunikacji (Laszkiewicz, Kalińska-Kula, 2023).

Drugą istotną praktyką jest udział w kampaniach reklamowych marek luksusowych lub/i masowych, co zwiększa rozpoznawalność wirtualnych postaci, ich prestiż oraz globalny zasięg. Współpraca z wirtualnymi influencerami jest dla marek atrakcyjna, ponieważ umożliwia pełną kontrolę nad wizerunkiem, eliminuje ryzyko zachowań niepożądanych oraz pozwala na tworzenie zaawansowanych, często futurystycznych wizualizacji, niedostępnych w przypadku pracy z osobami rzeczywistymi (Łaszkiewicz, 2024).

Kolejną strategią jest aktywna interakcja z odbiorcami, obejmująca odpowiadanie na komentarze, prowadzenie dialogu, a nawet organizowanie sesji pytań i odpowiedzi. Tego typu działania sprzyjają budowaniu relacji paraspołecznej, w której odbiorcy angażują się emocjonalnie i postrzegają wirtualnego influencera jako bliską osobę, mimo świadomości jego fikcyjności (Góraj, 2024).

Ostatnim elementem jest wykorzystywanie zaawansowanej analityki danych i narzędzi sztucznej inteligencji. Algorytmy AI umożliwiają dopasowanie publikowanych treści do preferencji użytkowników, analizę reakcji społeczności oraz optymalizację kampanii marketingowych. Narzędzia te są wykorzystywane do generowania grafik, animacji i scenariuszy komunikacyjnych, co pozwala szybko reagować na trendy i zmieniające się potrzeby odbiorców (Gerlich, 2023).

Podsumowując, światowe przykłady pokazują, że wirtualni influencerzy nie tylko konkurują z tradycyjnymi osobami w marketingu, ale również oferują markom wyjątkową elastyczność, co staje się coraz bardziej atrakcyjne dla marek i odbiorców na całym świecie. Co istotne, mogą pełnić funkcję nieformalnych nauczycieli, piosenkarzy, graczy, doradców itp., a ich przewagą jest możliwość pełnej kontroli nad wizerunkiem i narracją, możliwość hiperpersonalizacji treści, brak zmęczenia i stała dostępność w mediach społecznościowych 24/7 (Gunkel, Wales, 2021; Kim i in., 2023).

Przykłady w Polsce

W Polsce rynek wirtualnych influencerów dopiero się rozwija. Pierwsze przykłady cyfrowych postaci angażujących się w działania marketingowe i edukacyjne to:

1. Sara Kosmos – to jedna z pierwszych polskich wirtualnych influencerek, stworzona przez agencję Nakatomi. Jej profil na Instagramie i TikToku skupia się na lifestyle, modzie oraz nowoczesnych technologiach. Postać uczestniczy w kampaniach promocyjnych dla marek technologicznych i modowych. Integracja z AI umożliwia analizę reakcji odbiorców i dopasowanie publikowanych treści (Nakatomi, 2023; SpidersWeb, 2023).
2. Lil Miquela Polska (lokalne adaptacje). Niektóre polskie firmy eksperymentują z tworzeniem lokalnych wersji międzynarodowych wirtualnych influencerów, dostosowanych do polskiej kultury i języka, np. w kampaniach reklamowych marek odzieżowych. Takie postacie uczestniczą w działaniach marketingowych online, zwiększając zaangażowanie odbiorców poprzez media społecznościowe (Money.pl, 2023).
3. Wirtualni streamerzy i awatary w branży gamingowej. Polscy twórcy gamingowi coraz częściej korzystają z wirtualnych postaci 3D oraz motion capture do prowadzenia transmisji na żywo, np. na Twitchu czy YouTube. Awatary te pozwalają na większą interakcję z publicznością oraz tworzenie unikalnych doświadczeń immersyjnych (PwC, 2021).
4. Wirtualni asystenci edukacyjni. Polskie uczelnie i platformy edukacyjne eksperymentują z wirtualnymi postaciami wspierającymi proces nauczania, np. w kursach online z programowania czy sztucznej inteligencji. Takie wirtualne postacie pełnią funkcję edukacyjną i marketingową (Koles, 2024).

Podsumowując, liczba polskich wirtualnych influencerów jest obecnie znacznie mniejsza niż w przypadku ich globalnych odpowiedników. Polskie marki stopniowo dostrzegają w nich szansę na wyróżnienie się w konkurencyjnym środowisku mediów społecznościowych. Można zatem stwierdzić, że Polska podąża ścieżką wyznaczoną przez globalnych pionierów, a rozwój tej niszy będzie prawdopodobnie przyspieszał wraz z zainteresowaniem społeczności online i spadkiem kosztów tworzenia zaawansowanych modeli cyfrowych.

Reakcja odbiorców, wyzwania etyczne i trendy wirtualnych influencerów

Wirtualni influencerzy wywołują zróżnicowane reakcje wśród odbiorców. Badania pokazują, że publiczność jest zwykle pozytywnie nastawiona do cyfrowych postaci, zwłaszcza gdy treści są angażujące, estetyczne i spersonalizowane (Koles, 2024). Interaktywność postaci, możliwa dzięki integracji z AI, pozwala na dynamiczną reakcję na komentarze, wiadomości i trendy w mediach społecznościowych, co zwiększa poczucie relacji paraspołecznej pomiędzy użytkownikiem a postacią. Dostosowując treści do indywidualnych preferencji odbiorców, wirtualni influencerzy znacząco podnoszą poziom zaangażowania i lojalności wobec marek (Floridi, 2020).

Jednocześnie pojawiają się istotne wyzwania etyczne. Po pierwsze, kwestia transparentności – użytkownicy mogą nie być świadomi, że mają do czynienia z postacią cyfrową, co rodzi pytania o manipulację i rzetelność przekazu (Gunkel, Wales, 2021). Po drugie, problem ochrony danych – wirtualni influencerzy analizują zachowania użytkowników, co wymaga odpowiedniego zarządzania prywatnością i zgodnością z regulacjami, m.in. RODO. Po trzecie, kwestie kulturowe i społeczne – niektóre wirtualne postacie mogą propagować nierówności lub stereotypy, np. idealizując nierealistyczne wzorce piękna, co może wpływać negatywnie na samoocenę odbiorców (Kaplan, Haenlein, 2019).

W kontekście trendów i kierunków rozwoju, wirtualni influencerzy ewoluują bliżej pełnej immersji i większej personalizacji. Popularność zyskują postacie tworzone w technologii VR i AR, które umożliwiają odbiorcom interakcję w przestrzeni trójwymiarowej oraz uczestnictwo w wirtualnych wydarzeniach i koncertach, np. Hatsune Miku w Japonii (Condry, 2013). Kolejnym trendem jest rozwój wirtualnych streamerów, takich jak CodeMiko, którzy dzięki motion capture i AI angażują społeczności gamingowe w czasie rzeczywistym.

Równie istotny jest rozwój lokalnych wirtualnych influencerów, np. w Polsce, takich jak Sara Kosmos czy Ada z AI Academy. Trend ten wskazuje na rosnącą popularność postaci dostosowanych do specyfiki kulturowej i językowej danego kraju, które mogą prowadzić spersonalizowane kampanie marketingowe i edukacyjne (Nakatomi, 2023; SpidersWeb, 2023). W sektorze edukacji wirtualni influencerzy wspierają proces nauki poprzez interaktywny storytelling i indywidualne dostosowanie treści (Koles, 2024).

Podsumowując, wirtualni influencerzy funkcjonują na granicy innowacji technologicznych, marketingu i etyki. Popularność rośnie dzięki możliwości angażowania odbiorców w sposób spersonalizowany i interaktywny, jednak równocześnie wymaga odpowiedniego podejścia do kwestii etycznych, transparentności, bezpieczeństwa danych i wpływu społecznego. W przyszłości można oczekiwać dalszego rozwoju postaci wirtualnych w kierunku pełnej immersji, integracji z VR/AR oraz lokalizacji kulturowej, co pozwoli markom i instytucjom skuteczniej angażować odbiorców na całym świecie.

Podsumowanie i wnioski

Wirtualni influencerzy to przykład, jak technologia może zmieniać świat komunikacji i marketingu. Dzięki połączeniu grafiki 3D i sztucznej inteligencji marki zyskują możliwość tworzenia postaci, które są zawsze dostępne, spójne w przekazie i niezwykle elastyczne. Mogą one mówić dokładnie to, co zaplanuje twórca, a jednocześnie reagować na zachowania odbiorców w sposób coraz bardziej naturalny (Floridi, 2020; Marwick, 2021).

Polskie przykłady takie jak Sara Kosmos czy Ailunna pokazują, że jesteśmy dopiero na początku tej drogi. Projekty te bardziej testują potencjał technologii niż jeszcze realnie konkurują z tradycyjnymi influencerami. Aby wirtualne postacie mogły zdobyć szerszą publiczność, muszą oferować coś więcej niż tylko atrakcyjną grafikę – potrzebne jest autentyczna narracja, poczucie spójności i emocje, z którymi odbiorcy mogą się utożsamiać (SpidersWeb, 2023; Nakatomi, 2023).

Nie można też pomijać kwestii etycznych. Wirtualne postacie, nawet jeśli z założenia fikcyjne, wywołują realne emocje i budują więź paraspołeczną. To rodzi pytania: czy odbiorcy zawsze wiedzą, że mają do czynienia z AI? Jak takie postacie wpływają na samoocenę i postrzeganie świata przez młodych ludzi? Jeśli braknie przejrzystości i jasnych zasad, łatwo o nadużycia – od dezinformacji po promowanie nierealistycznych wzorców piękna (Moustakas i in., 2020).

W kolejnych latach możemy spodziewać się coraz większej integracji wirtualnych postaci z metaverse oraz technologiami VR i AR. To otworzy drogę do zupełnie nowych form interakcji, gdzie influencer publikuje posty, ale też uczestniczy w wydarzeniach i rozmowach w czasie rzeczywistym, niemal jak

prawdziwy człowiek. Wszystko wskazuje na to, iż wirtualni influencerzy to nie chwilowa moda, ale zapowiedź głębszych zmian w komunikacji cyfrowej. Mogą stać się wartościowym narzędziem marketingu, pod warunkiem, że nie zapomnimy, iż po drugiej stronie ekranu zawsze stoi człowiek – odbiorca, który oczekuje szczerości i sensownego dialogu.

Dlatego przyszłe badania i praktyka powinny iść w dwóch kierunkach: po pierwsze, analizować długoterminowy wpływ wirtualnych postaci na odbiorców, a po drugie – pracować nad etycznymi zasadami ich wykorzystania. Tylko wtedy innowacyjność połączona z odpowiedzialnością pozwoli wykorzystać w pełni potencjał tej technologii.

Literatura

1. Berryman, R., Kavka, M. (2018). Crying on YouTube: Vlogs, self-exposure and the productivity of negative affect. *Convergence, 24*(1), 85-98. https://doi.org/10.1177/1354856517736981
2. Condry, I. (2013). *The soul of anime: Collaborative creativity and Japan's media success story*. Duke University Press.
3. Flavián, C., Orús, C., Sánchez, S. I. (2019). The impact of virtual, augmented and mixed reality technologies on the customer experience. *Journal of Business Research, 100*, 547-560. https://doi.org/10.1016/j.jbusres.2018.10.050
4. Floridi, L. (2020). AI and its new winter: From myths to realities. *Philosophy & Technology*, *33*, 1-3. https://doi.org/10.1007/s13347-020-00396-6
5. Floridi, L. (2023). *The ethics of artificial intelligence*. Oxford University Press.
6. Gerlich, M. (2023). The power of virtual influencers: Impact on consumer behaviour and attitudes in the age of AI. *Administrative Sciences, 13*(8), 178. https://doi.org/10.3390/admsci13080178
7. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y. (2014). Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27, 2672-2680
8. Góraj, K. (2024). The impact of computergenerated influencers on social media advertising. *Social Communication, 25*(1), 137-144. https://doi.org/10.57656/sc-2024-0014
9. Gunkel, D.J., Wales, J.J. (2021). Debate: what is personhood in the age of AI?. *AI & Soc, 36*, 473-486. https://doi.org/10.1007/s00146-020-01129-1

10. Hollebeek, L. D., Clark, M. K., Andreassen, T. W., Sigurdsson, V., & Smith, D. (2020). *Virtual reality through the customer journey: Framework and propositions*. *Journal of Retailing and Consumer Services, 55*, Article 102056. https://doi.org/10.1016/j.jretconser.2020.102056
11. Huang, M.-H., Rust, R. T. (2021). A strategic framework for artificial intelligence in marketing. *Journal of the Academy of Marketing Science, 49*(1), 30-50. https://doi.org/10.1007/s11747-020-00749-9
12. Javornik, A. (2016). Augmented reality: Research agenda for studying the impact of AR on consumer behaviour. *Journal of Retailing and Consumer Services, 30*, 252-261. https://doi.org/10.1016/j.jretconser.2016.02.004
13. Johnson, S. S. (2020). The Future of Work. *American Journal of Health Promotion, 34*(7), 809-812. https://doi.org/10.1177/0890117120943748a
14. Kaplan, A. M., Haenlein, M., (2019). Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. *Business Horizons, 62*(1), 15-25. https://doi.org/10.1016/j.bushor.2018.08.004
15. Kim, E., Kim, D., Shoenberger, H. (2023). The next hype in social media advertising: Examining virtual influencers' brand endorsement effectiveness. *Frontiers in Psychology, 14*, 1089051. https://doi.org/10.3389/fpsyg.2023.1089051
16. Krippendorff, K. (2018). *Content Analysis: An Introduction to Its Methodology (4th ed.)*. Thousand Oaks, CA: Sage.
17. Koles, B., Audrezet, A., Guidry Moulard, J., Ameen, N., McKenna, B. (2024). The authentic virtual influencer: Authenticity manifestations in the metaverse. *Journal of Business Research, 170*, 114325. https://doi.org/10.1016/j.jbusres.2023.114325
18. Laszkiewicz, A., Kalińska-Kula, M. (2023). Virtual influencers as an emerging marketing theory: A systematic literature review. *International Journal of Consumer Studies, 447*(6), 2479-2494. https://doi.org/10.1111/ijcs.12956
19. Lee, H., Shin, M., Yang, J., Chock, T. M. (2025). Virtual Influencers vs. Human Influencers in the Context of Influencer Marketing: The Moderating Role of Machine Heuristic on Perceived Authenticity of Influencers. *International Journal of Human-Computer Interaction*, *41*(10), 6029-6046. https://doi.org/10.1080/10447318.2024.2374100
20. Łaszkiewicz, A. (2024). Artificial intelligence (AI) influencers in influencer marketing. *Zeszyty Naukowe Wyższej Szkoły Informatyki i Zarządzania, 193*(2), 21-36. https://doi.org/10.29119/1641-3466.2024.193.2

21. Marwick, A. E. (2018). The Algorithmic Celebrity: The Future of Internet Fame and Microcelebrity Studies. W: C. Abidin, M. L. Brown (red.), Microcelebrity Around the Globe (s. 161-169). Emerald Publishing. https://doi.org/10.1108/978-1-78756-749-820181015
22. Moustakas, E., Lamba, N., Mahmoud, D., Ranganathan, C. (2020). Blurring lines between fiction and reality: Perspectives of experts on marketing effectiveness of virtual influencers. *2020 International Conference on Cyber Security and Protection of Digital Services* (Cyber Security), 1-6. https://doi.org/10.1109/CyberSecurity49315.2020.9138861
23. Palloff, R.M., Pratt, K. (2007). *Building Online Learning Communities: Effective Strategies for the Virtual Classroom*. San Francisco: Jossey-Bass.

Netografia

1. McKinsey & Company. (2022). *The rise of virtual influencers*. Pobrane z: https://www.mckinsey.com/industries/technology (dostęp: 08.09.2025).
2. Money.pl. (2023). *Mają miliony fanów, choć sami naprawdę nie istnieją – top 10 wirtualnych influencerów*. Pobrane z: https://www.money.pl/gospodarka/maja-miliony-fanow-choc-sami-naprawde-nie-istnieja-top-10-wirtualnych-influencerow-6874529348348640a.html (dostęp: 08.09.2025).
3. Nakatomi. (2023). *Wirtualni influencerzy – tworzenie i promocja*. Pobrane z: https://nakatomi.pl/wirtualni-influencerzy (dostęp: 08.09.2025).
4. Nakatomi. (2024). *Wirtualni influencerzy – kim są, TOP 10 i ich zarobki.* Pobrane z: https://nakatomi.pl/wirtualni-influencerzy (dostęp: 08.09.2025).
5. PwC. (2021). *Consumer trust in AI-driven marketing*. Pobrane z: https://www.pwc.com (dostęp: 08.09.2025).
6. SpidersWeb. (2023). *Sara Kosmos, Lil Miquela – polska próba wirtualnej influencerki*. Pobrane z: https://spidersweb.pl/plus/2023/03/sara-kosmos-lil-miquela (dostęp: 08.09.2025).
7. Statista. (2023). *Number of active virtual influencers worldwide 2019-2023*. Pobrane z: https://www.statista.com/statistics/1171446/virtual-influencers-worldwide (dostęp: 08.09.2025).

Volodymyr Malovichko[1], Roman Rybalka[2], Kostiantyn Honcharov[3]

Ukrainian State University of Science and Technologies, Dniepr

https://doi.org/10.29316/9788368103304_8

PROPOSAL FOR ENHANCING SUSTAINABILITY IN THE DIGITALISATION OF EDUCATION IN THE USUST

PROPOZYCJA ZWIĘKSZENIA ZRÓWNOWAŻONEGO ROZWOJU W CYFRYZACJI EDUKACJI W USUST

Abstract

The purpose of this study is to develop a proposal to enhance the resilience of the education-related network infrastructure at the Ukrainian State University of Science and Technologies by utilising blockchain technology. This study focuses on defining the criteria for selecting the best candidate among existing blockchain frameworks, selecting one, and evaluating its performance in terms of transaction throughput using five regression and three artificial neural network models. The best model is then selected based on the mean-squared error. Findings showed that the current network topology is sufficient to allow deploying and maintaining the Hyperledger Fabric. This study implies that network throughput can be predicted for a given number of peers and endorsers using the selected model.

Keywords: blockchain, regression, radial basis neural network, Hyperledger Fabric

[1] https://orcid.org/0009-0008-2704-5555, e-mail: v.v.malovichko@ust.edu.ua
[2] https://orcid.org/0000-0001-7444-0517, e-mail: r.v.rybalka@ust.edu.ua
[3] https://orcid.org/0000-0002-0432-1077, e-mail: k.v.honcharov@ust.edu.ua

Streszczenie

Celem badania jest stworzenie propozycji zwiększenia odporności infrastruktury sieciowej związanej z edukacją Ukraińskiego Państwowego Uniwersytetu Nauki i Technologii w oparciu o technologię blockchain. Badanie koncentruje się na zdefiniowaniu kryteriów wyboru najlepszego kandydata spośród istniejących frameworków blockchain, wybraniu jednego z nich i ocenie jego wydajności pod kątem przepustowości transakcji za pomocą pięciu modeli regresji i trzech modeli sztucznych sieci neuronowych, przy jednoczesnym wyborze najlepszego modelu z uwzględnieniem błędu średniokwadratowego. Wyniki badań wykazały, że obecna topologia sieci jest wystarczająca do wdrożenia i utrzymania Hyperledger Fabric. Badanie sugeruje, że przepustowość sieci można przewidzieć dla danej liczby peerów i endorserów, biorąc pod uwagę wybrany model.

Słowa kluczowe: blockchain, regresja, radialna sieć neuronowa, Hyperledger Fabric

Introduction

The current stage of societal development is characterised by the rapid proliferation of digital technologies across all areas of human activity. Digitalisation has become a crucial prerequisite for the innovative advancement of the economy, science, and education. In higher education, it enhances the efficiency of educational process management, expands access to academic resources, and fosters new learning formats and interactive engagement between students and instructors. The necessity of digitalising educational processes is driven not only by global trends but also by the need to adapt to contemporary challenges, including data security and resilience against cyber threats.

In Ukraine, the digitalisation of education has become particularly relevant following the COVID-19 pandemic and the full-scale invasion by the Russian Federation (RF). On the one hand, the introduction of digital technologies facilitates the continuity of the educational process regardless of the learners' location; on the other hand, it introduces new challenges in the field of information security, particularly concerning the protection of personal data of students and instructors, as well as the confidentiality of educational course materials and administrative systems (Kravtsov et al., 2021; Pavlysh et al., 2023).

At present, a wide variety of digital tools are utilised across the education sector. In particular, Learning Management Systems (LMS) such as Moodle, Google Classroom, Blackboard, and others provide the ability to create educational courses, upload and integrate various types of content (lecture texts, presentations, videos, audio files, documents, and links), develop interactive assignments and tests, enable communication between students and instructors through forums, chats, and private messages, as well as monitor learners' progress and assess their performance. Video conferencing tools (Zoom, Google Meet, Microsoft Teams) enable lectures and consultations to be held online. Various online learning platforms (Coursera, edX, Udemy, and others) expand opportunities for student self-education. In addition, cloud services (Google Drive, Microsoft OneDrive) are used for collaborative project work and data storage, while administrative information systems help automate the management processes of educational institutions.

In Ukraine, several nationwide digitalisation projects are underway in the education sector. The state online platform "Diia.Osvita" offers free online courses and educational series designed to enhance digital literacy, promote professional development, and cultivate new entrepreneurial, professional, and social skills among citizens (Kyrylenko, Jakimova, 2025). The Unified State Electronic Database on Education (EDEBO) serves as an information system that ensures centralised record-keeping, promotes transparency, and automates key processes within Ukraine's education sector.

For secondary education, various national-level digital solutions have also been developed. The "All-Ukrainian Online School" electronic platform offers students nationwide access to video lessons, tests, and educational materials. Additionally, state electronic grade books and attendance journals, integrated into the software-hardware complex "AIKOM," enable automated tracking of student performance and attendance. Moreover, the educational mobile application "Mriia" is designed to support students and parents by facilitating interactive engagement with the educational process.

At the same time, higher education institutions lack unified state-standardised digitalisation solutions. Each university is compelled to develop its own platforms and tools, taking into account the specific characteristics of the educational process and the available resources (Fernández et al., 2023; Pavlysh et al., 2023). For

example, some institutions implement Learning Management Systems, such as Moodle or Google Classroom, while others develop proprietary electronic portals tailored to the university's internal needs for students and instructors. Higher education institutions also operate internal administrative information systems to manage academic processes, personnel, document circulation, and financial accounting. Such autonomy enables the creation of flexible digital solutions; however, it also complicates efforts toward unification, system integration, and maintaining a high level of cybersecurity (Areshonkov, 2020).

The localisation and organisational structure of the Ukrainian State University of Science and Technologies (USUST) make it a suitable candidate for examining proposals to improve sustainability in the digitalisation process. USUST was established in the city of Dnipro, Ukraine, in 2021 through the merger of the Dnipro National University of Railway Transport named after Academician V. Lazaryan and the National Metallurgical Academy of Ukraine (UDUNT, n.d.). At present, USUST comprises five formerly independent higher education institutions, located across five separate campuses within the city. Each institution has been reorganised into an educational and research institute. The shortest distance between adjacent campuses is the width of a roadway, while the longest distance is approximately 5 km. Operational processes at USUST are divided into intra-institute and inter-institute (university-wide) activities. Consequently, information exchange – whether in electronic or paper form – occurs between geographically distributed institutes.

Digitalised workflows at the USUST are supported through client-server interaction between automated workstations (AW) and the databases (DBs) of the respective institutes. AW functions as a rich-client application, with user access protected by passwords. Access to the AWs from the DB is organised within each institute's local area network (LAN) and is regulated by predefined user roles corresponding to the workstation's hierarchical level. The lowest level consists of departmental AWs, followed by those of the dean's offices and directorates. Independent DB instances are hosted on servers within each institute, making these DBs centralised data repositories at the institute level.

The AW automates a wide range of operational processes at USUST related to student record management. These include logging the progress of individual student study plans, generating grade sheets for recording academic performance

during assessment activities, and preparing academic transcripts and diploma supplements. The AW significantly simplifies the preparation of various student-related documents, such as submissions for student transfers, withdrawals, and reinstatements, as well as the generation of scholarship orders and directives for forming academic groups. Depending on their access rights, AW users can view statistical information, including academic group lists, student performance rankings, the presence of academic debts, affiliation with specific preferential categories, and forms of educational funding, among others. Additionally, AW users have access to operational documentation that supports the educational process, including curricula, educational programs, and related materials.

The current digitalisation support infrastructure at USUST covers the majority of the university's operational processes. However, it remains vulnerable to certain risks, including blackouts (temporary power outages lasting several hours to several days), physical damage to elements of the university's IT infrastructure, and cyber threats. Between October 2022 and September 2024, the armed forces of the RF carried out 1,024 attacks on Ukraine's energy infrastructure using aerial means of destruction (unmanned aerial vehicles and missiles) (Ghalushhenko, 2024). During 2021-2023, the RF conducted approximately 150 cyberattacks per month targeting Ukraine's infrastructure, primarily critical infrastructure systems (Чайка, 2023). Among other consequences, these attacks have resulted in partial or complete power outages in populated areas. Considering the nature and scale of threats posed by the RF, the risks of blackouts, physical destruction (or damage) to critical infrastructure facilities in Ukraine, and cyber threats are expected to persist throughout the ongoing full-scale invasion launched by the RF on 24 February 2022.

Enhancing the current organisation of client-server interaction between the AW and the DB would significantly reduce the risks associated with the aforementioned threats, even without altering the USUST's existing network infrastructure. The presence of several geographically distributed servers within USUST enables the establishment of a distributed data storage system (distributed ledger).

One of the most relevant solutions for the specified problem is blockchain. Among its advantages (Capocasale et al., 2023) are:

- DB decentralisation through redundancy – copies of the DB are stored on multiple servers (each at a respective institution), consequently eliminating a Single Point of Failure (SPoF) and increasing resilience to the failure of a single DB repository (and under certain conditions, even more);
- Immutability of DB records – data is stored in an ordered sequence of blocks that are cryptographically linked;
- Transparency – the ability to view the history of performed actions on the DB (a consequence of immutability);
- Resistance to data forgery due to consensus mechanisms, which ensures the ability to function in a network environment without complete trust in its nodes.

The need to enhance the resilience of USUST's current network infrastructure against the aforementioned threats substantiates the relevance of this research's topic.

The purpose of this research is to formulate a proposal to enhance the resilience of the IT infrastructure supporting the digitalisation of the educational process at USUST against physical threats (e.g., destruction, damage, or temporary incapacitation due to blackouts) and cyber threats.

Methodology

Current network infrastructure of USUST and risk analysis

Electronic communication between the institutes of USUST (labelled as "Organisation" in Figure 1) is carried out via the Internet, while within each institute, it is implemented through LAN. The network infrastructure of each USUST institute is organised as follows (Figure 1): Internet connectivity is provided by at least one service provider; the institute's LAN provides all (N + M) client computers (Clients) with Internet access, among which M computers are configured as AWs, whose access to the corresponding DBs is ensured through the LAN. On each institute's server (Server 1), a DB instance containing records relevant to that institute is deployed. Each AW has assigned permission roles (password-protected), enabling role-based access to the DB for reading and writing operations. Other servers (e.g., Server 2) host additional resources, including the

institute's website, a learning management system (e.g., Moodle-based "Lider"), a platform for developing practical skills for students in relevant specialities (e.g., computer engineering and cybersecurity), a library repository, and other services. On average, each institute operates about 5 servers (only 2 are shown in Figure 1).

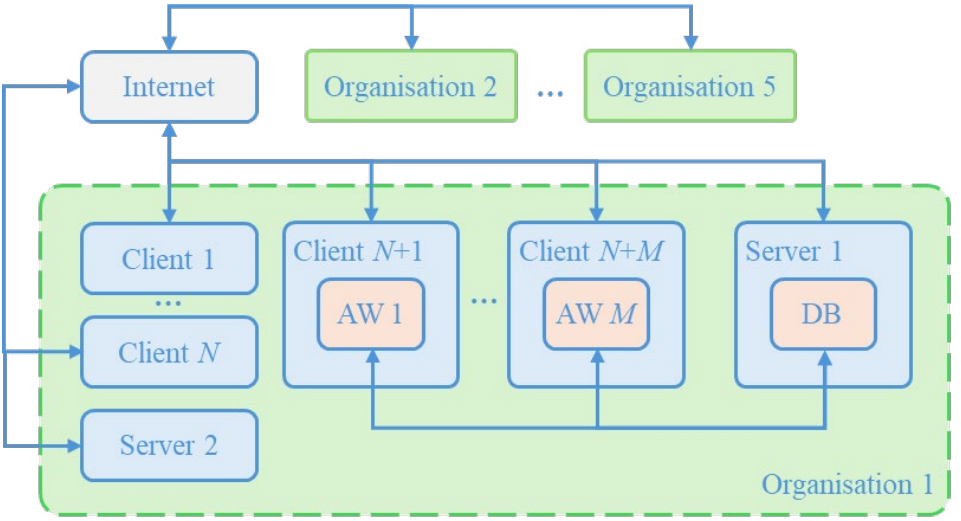


Figure 1. Network infrastructure of the USUST

Source: compiled by the author.

Blackouts can significantly hinder the maintenance of electronic communication within USUST. During blackouts, access to the institutes' DBs may be wholly or partially unavailable in cases of (1) power supply interruption at the institute or (2) power outage affecting the internet service provider's equipment. Under the current USUST network structure (Figure 1), DB access for a given institute is impossible in the first case and limited to the institute's LAN in the second.

The centralised placement of DBs (Figure 1) within each USUST institute has certain drawbacks typical of this network topology. Among these drawbacks (Merlec, In, 2024) (Yuan et al., 2025) (Nasir et al., 2024) are the following: SPoF (Single Point of Failure) – a server failure makes access to the corresponding DB impossible; data tampering – unauthorized modification of DB records; limited data transparency – lack of a built-in capability to trace the sequence of DB modifications; and scalability limitations – reduced performance due to the

potential increase in the volume of requests directed to a single processing point, i.e., the server.

Table 1 presents the results of risk identification and analysis for the current configuration of the USUST network infrastructure. As shown in this table, risk R1 has the highest likelihood and impact scores. Therefore, its processing priority is also the highest.

Table 1. Risk identification & analysis for the current configuration of the USUST network

Risk ID	Description	Source	Likelihood	Impact
R1	Blackouts leading to DB/server downtime	War-related power outages, unstable energy supply	High	High (loss of service continuity, disruption of academic processes)
R2	Internet disconnection between institutes	War-related infrastructure damage, Internet Service Provider issues	High	Medium (loss of inter-institute communication, delays)
R3	Single point of failure of the central server	Hardware failure, misconfiguration, or cyberattack	Medium	High (unavailability of services, data loss)
R4	Data tampering within the centralised DB	Insider misuse or compromised AW credentials	Medium	High (loss of data integrity, reputational damage)
R5	Data breach exposing personal information	External hacking, weak password policies and phishing	Medium	High (legal, ethical, and reputational consequences)
R6	DDoS attack on servers or the network	External adversaries are exploiting weak network protection	Medium	Medium (temporary denial of services)

Source: compiled by the author.

A brief overview of blockchain

The use of distributed ledger technology reduces the risk associated with centralised DB storage. This technology involves storing identical copies of the DB on different network nodes (peers) and maintaining them in a consistent state.

One example of this technology is blockchain. Data in blockchain is represented as ‘blocks’ that are cryptographically linked in pairs to form a ‘chain’. This organisation of data allows blockchain to ensure at least the following (Capocasale et al., 2023): decentralisation – elimination of SPoF by storing DBs on different network nodes; significant complication of the procedure for changing block data – due to the need to change all subsequent blocks, which leads to the practical immutability of blocks; transparency – due to the ability to view all blocks in the chain.

One of the most important attributes of blockchain is the consensus algorithm. It allows consensus to be reached on the distributed ledger without the need for a single decision-making centre (i.e., SPoF is eliminated). It ensures network resilience in the event of faulty or malicious nodes or certain types of attacks (e.g., double-spending, 51% attacks, DDoS). Consensus mechanisms are classified into proof-based and voting-based (Nasir et al., 2024). The former are typically used in permissionless blockchains, while the latter are used in permissioned blockchains. As of 2022, there are approximately 30 consensus algorithms (Guru et al., 2023). Comparing the characteristics of all these algorithms is beyond the scope of this study.

In general, voting-based consensus mechanisms are more energy-efficient and faster than proof-based ones, but they can be less decentralised. The energy efficiency of voting-based consensus mechanisms stems from the lack of competition for computing power to obtain the right to create a block. Unlike, for example, Proof-of-Work (PoW), network validators do not consume significant amounts of energy. The high speed of voting-based consensus is enabled by the absence of anonymous nodes in private (permissioned) networks, which allows the use of faster consensus algorithms. On the other hand, voting-based consensus mechanisms can typically have problems with decentralisation. These consensus mechanisms are used in private (permissioned) networks, where decision-making is limited to a predetermined set of peers. The result is a risk of collusion among these peers, which can lead to network manipulation.

Blockchain selection criteria

The effectiveness of consensus algorithms (and, as a result, network security and performance) depends on how scalability, security, and decentralisation are optimised (the so-called 'Blockchain Trilemma problem'). For instance, the PoW advantage offers high security and decentralisation, but the scalability of business transactions remains insufficient (Nasir et al., 2024). Conversely, in Practical Byzantine Fault Tolerance (PBFT) – a representative instance of Byzantine Fault Tolerance (BFT) – the use of comparatively weaker consensus algorithms results in a lower level of security than that provided by PoW, while achieving higher performance. The scalability of PBFT is, to some extent, superior to that of PoW. However, it is notably constrained by the exponential increase in communication overhead as the number of participating nodes grows. (Tomić, 2021). Choosing a consensus algorithm involves choosing a specific implementation of the Blockchain Trilemma problem solution.

When selecting a blockchain implementation, numerous criteria must be considered, which can be categorised as follows:

- Governance and Architecture: public (permissionless) or private (permissioned); level of decentralisation – the impact of the distribution of decision-making power among peers.
- Performance and Scalability (Nasir et al., 2024): transaction throughput – number of successful transactions per second (TPS); latency – the time interval between the moment a transaction is submitted and its final confirmation; scalability – the ability of blockchain to operate effectively when the number of peers and/or the volume of transactions increases; resource consumption – energy consumption, central processing unit operating time, memory capacity, channel bandwidth; complexity of deployment and maintenance.
- Security, Privacy, and Reliability: resilience to cyber-attacks, e.g., 51%, DoS, Sybil; fault tolerance – Crash Fault Tolerance (CFT), which tolerates failure (due to hardware, software or network-related issues) of f nodes (while the total number of nodes is not less than $2f + 1$) (Zhou et al., 2023), Or BFT, which tolerates even the presence of f nodes with malicious behaviour

(while the total number of nodes is not less than $3f + 1$) (Capocasale et al., 2023); the ability to ensure data and/or user confidentiality (anonymity).

- Functionality and Technical Details: consensus algorithm; programming language support; DB type; interoperability – the ability to exchange data with other blockchains; relationship to cryptocurrency.

To reduce the likelihood of risks R1-R4 (Table 1), we suggest choosing a blockchain implementation that meets the following requirements: is private (permissioned); has a moderate level of decentralisation, transaction throughput and latency; has low or moderate scalability, resource consumption, deployment and maintenance complexity, and resistance to cyber-attacks; has high fault tolerance (CFT); is capable of ensuring data confidentiality. Desirable features include support for common programming languages for blockchain development, the use of DB structures capable of handling complex queries, and the flexibility to replace the consensus algorithm with a BFT-based one if required. Non-essential features include interoperability, integrated cryptocurrency support, and mechanisms for ensuring user anonymity. When selecting blockchain requirements, the current network infrastructure of USUST (Figure 1), which operates in a relatively trustworthy environment, the expected peak transaction intensity of approximately 100 TPS, and risk identification and analysis (Table 1) were considered.

Hyperledger Fabric (HLF) has met these requirements since version 3.0; therefore, we recommend adopting HLF for USUST. In HLF, every component, whether internal or external to the network – such as peers, orderers, client applications, and administrators – possesses a unique digital identity. Consequently, all network entities are explicitly identifiable.

The HLF ledger comprises two components: the current state of the ledger (the world state), which is maintained in a DB, and the transaction log (the history of updates to the current state), which forms the blockchain and is implemented as a file. HLF enables ledger operations – such as executing transactions, performing queries, and managing controlled access to the ledger – through smart contracts, which can be defined within the same chaincode or across separate chaincodes (Hyperledger Fabric, n.d.) HLF supports Go, Node.js, and Java for chaincode implementation. When a client application invokes a chaincode function to interact with the ledger (for read or write operations), a transaction is generated. These

transactions trigger state transitions, and their records collectively constitute the ledger in HLF.

In HLF, ledger hosting is carried out by a set of peer nodes (peers), which utilise smart contracts to enable read and write operations on the ledgers (Hyperledger Fabric, n.d.) Each peer maintains copies of the ledgers and smart contracts and exposes a set of APIs that allow client applications to interact with it. Beginning with HLF version 2.4, the Fabric Gateway Service (FGS) is deployed on peers, enabling them to manage transaction proposals and endorsements. The number of peers responsible for endorsing transactions (referred to as 'endorsers') is determined by the endorsement policy of a given network. In addition to peer nodes, there are orderer nodes (also known as orderers) and the corresponding ordering service, which ensure ledger consistency. Orderers aggregate transactions into blocks and subsequently distribute these blocks to peers for validation and commitment. At the same time, orderers do not maintain a state database and do not support unilateral switching from CFT (Replication and Fault Tolerant – Raft protocol, which is an example of CFT) to BFT (using the SmartBFT protocol).

Performance evaluation of the selected blockchain

Evaluating the performance of blockchain systems in general, and HLF in particular, is a non-trivial task due to the influence of numerous factors, including transaction size, block size, network scale, and available hardware resources such as CPU, memory, disk space, and disk or network I/O. As a rule, such an assessment is performed for a network of a specific configuration by changing the blockchain parameters using computer modelling.

With an increase in the number of peers, HLF performance is expected to improve due to the ability to load-balance endorsement requests across them (Hyperledger Fabric, n.d.) The results of HLF modelling reported by (Guggenberger et al., 2022) indicate that the highest sustainable throughput, approximately 400 TPS, was achieved with a peer count ranging from 8 to 32, a fixed number of endorsers (2) under an endorsement policy of '2 of 8 endorsers,' the Raft consensus mechanism, and CouchDB as the state DB. It is noteworthy that the mentioned modelling was performed for HLF version 2.0 and validated for version 2.2. Although the FGS has been available since version 2.4, this remains

the most comprehensive and detailed open-access study identified on the topic. We further assume that the introduction of FGS does not significantly influence the outcomes of the referenced modelling.

In general, an increase in the number of orderers is expected to reduce HLF performance. However, modelling results reported by Guggenberger et al. (2022) using the Raft consensus mechanism and CouchDB indicate that, for request rates below 1500 TPS, increasing the number of orderers to 64 has almost no impact on the maximum sustainable throughput. A slight improvement in throughput was observed when using four orderers.

To perform the max sustainable throughput assessment, depending on the number of peers and endorsers, we implemented and investigated the following machine learning models:

- Regressions: decision tree regression (DTR), ordinary least squares linear regression (LR), polynomial linear regression (PLR), random forest regression (RFR), support vector regression (SVR).
- Artificial neural networks (ANN): generalised regression ANN (GRANN), radial basis ANN (RBANN), exact radial basis ANN (ERBANN).

Regression models were implemented in Python using NumPy, pandas, and scikit-learn. ANN models were implemented in MATLAB (Deep Learning Toolbox). The raw dataset for training and validating these models was obtained from Guggenberger et al. (2022) and consists of 21 records. The dataset comprises three model inputs – total number of peers, endorsements, and orderers – and one output, the maximum sustainable throughput (TPS). Prior to model training, the dataset was analysed to identify correlations among the input variables. A strong correlation ($r = 0.89$) was observed between the total number of peers and the number of orderers. Consequently, orderers were excluded from the model inputs. The raw dataset was then randomly permuted and divided into 16 records for training and 5 for validation. Each subset was normalised to the range (0, 1).

Initially, we conducted the training and simulation of regression models. Their performance was evaluated using the coefficient of determination (R^2). For the model achieving the highest R^2 value, we computed the mean squared error (MSE) between the model's predicted values (obtained using the validation dataset) and the corresponding true values. This MSE value, denoted ε_{reg}, served as a reference for evaluating the ANN models' effectiveness.

Subsequently, we trained and simulated ANN models. The performance of the implemented models (GRANN, RBANN, and ERBANN) is strongly influenced by a key parameter – the spread of radial basis functions (hereinafter referred to as "spread"). The smoothness of function approximation increases proportionally with the spread value. To determine the optimal spread for each ANN model, we performed training, simulation, and MSE calculation following the same procedure used for regression models. The spread parameter for each model was sequentially varied within the range (1 × 10^-5, 2 × 10^-5… 10), encompassing 1,000 values. To identify the best-performing model, we compared the obtained MSE values with ε_{reg} and selected the ANN model with the smallest MSE below ε_{reg}, along with its corresponding spread value.

The choice of DB type is a key factor influencing HLF performance. When LevelDB is used, throughput can be two to three times higher than that achieved with CouchDB. Increasing the number of endorsers enhances system robustness but reduces throughput. Furthermore, the performance degradation observed with LevelDB is more pronounced than that with CouchDB (Guggenberger et al., 2022).

Optimising the deployment of HLF elements on different servers affects overall performance. According to the simulation results (Guggenberger et al., 2022), the highest throughput was observed when orderers, peers, and CouchDB instances were deployed on separate servers. When two or more of these entities were deployed on a single server, throughput decreased by approximately 35% with three entities per server.

Several other factors can influence HLF performance. A modest improvement of approximately 12% can be achieved by switching from a single-channel to a dual-channel configuration (Guggenberger et al., 2022). However, further increasing the number of channels does not yield a noticeable increase in maximum throughput. As the number of channels grows, so do the network's hardware requirements. Server hardware resources have a substantial impact on HLF performance. For instance, upgrading from two threads per physical CPU core (with 8 GiB of memory) to four threads (with 16 GiB of memory) improves CouchDB performance by approximately 90% (Guggenberger et al., 2022). It is worth noting that additional hardware upgrades yield diminishing returns

in performance. Therefore, given the associated cost increase, this approach to enhancing HLF performance requires economic justification.

Finally – the block parameters. Analysis of the dependencies of maximum sustainable throughput on block timeout, maximum block size, and size of data per transaction reveals certain values (approximately) at which a significant change in this performance occurs (Guggenberger et al., 2022): block timeout 0.5s, maximum block size about 100 transactions, size of data per transaction 1 KB.

Results

Hyperledger Fabric performance modelling results

Figure 2 shows the coefficient of determination values for five regression models (DTR, LR, PLR, RFR, and SVR). The MSE value for the RFR model is $\varepsilon_{reg} \approx 0.014$.

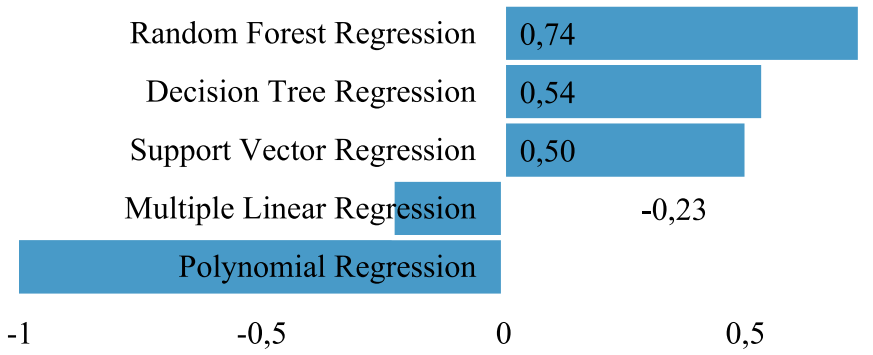


Figure 2. Coefficient of determination for the regression models

Source: compiled by the author.

Figure 3 shows the dependence of MSE on the “spread” value for each of the three considered ANN models. In this case, MSE was calculated for the validation dataset.

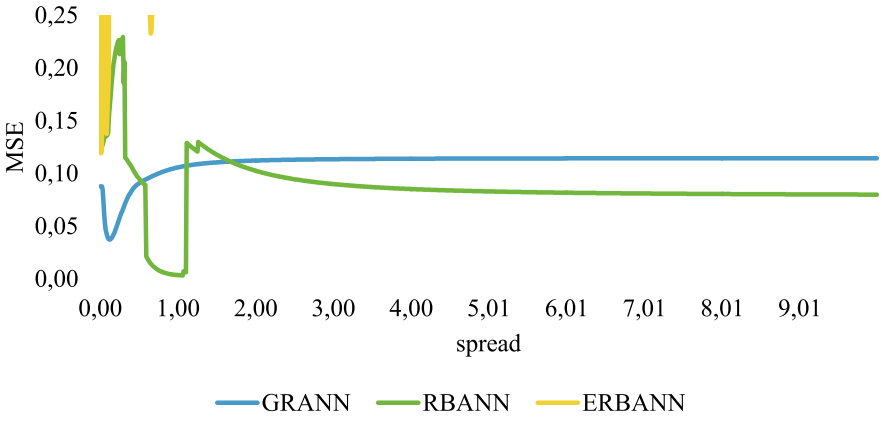


Figure 3. Mean-squared error for the ANN models

Source: compiled by the author.

Figure 4 presents the results of predicting the maximum sustainable throughput (TPS) for the RBANN model at a "spread" value of 1.06, with a corresponding MSE of 3.23×10^{-3}.

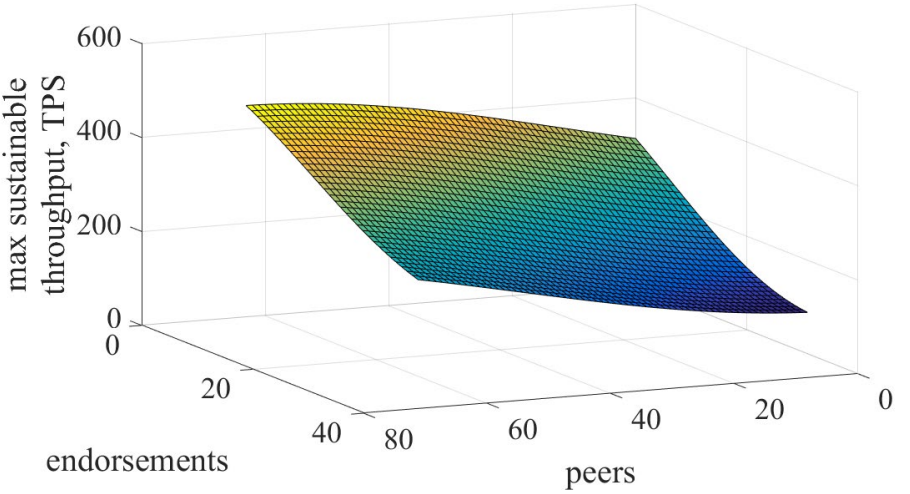


Figure 4. Max sustainable throughput predicted by RBANN

Source: compiled by the author.

Proposed network architecture

The proposed network infrastructure of USUST at the initial stage of HLF implementation is shown in Figure 5. In each institute (Organisation 1 … Organisation 5), one peer (P1) and one orderer (O1) are deployed on separate servers (Server 1 and Server 2). Thus, the USUST network comprises 5 orderer nodes and 5 peer nodes. On the same servers where the peers (P1) are deployed, a CouchDB is also deployed.

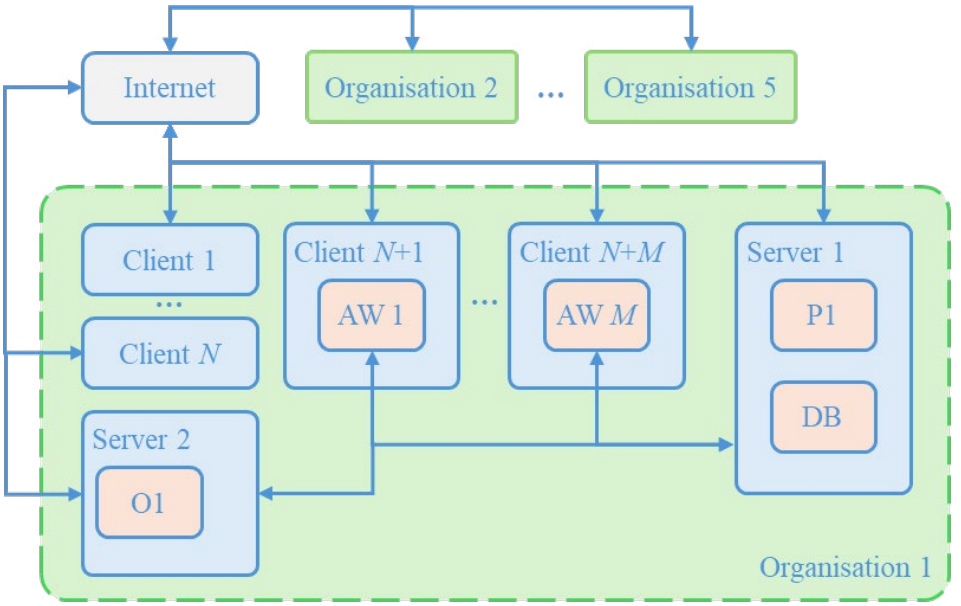


Figure 5. Proposed network infrastructure

Source: compiled by the author.

The configuration of HLF deployed on the proposed network infrastructure (Figure 5) is as follows: a single channel is available to all peers; the consensus mechanism is CFT (Raft protocol); no private collections are provided; the FGS enables an application connected to the institute's LAN to connect to one of the available trusted peers in this network; the number of endorsers is 2; the block timeout is 0.5 s; the maximum block size is about 100 transactions; the data size per transaction is 1 KB. According to the RBANN model evaluation shown

in Figure 4, for this number of peers and endorsers, the maximum sustainable throughput is expected to reach approximately 300 TPS.

Discussion

According to Table 1, risk R1 is identified as the highest-priority risk. To mitigate risk R1, it is sufficient to configure HLF to use the CFT (Raft protocol) consensus mechanism. Using five orderer nodes with Raft consensus ensures HLF operation in the presence of up to two non-Byzantine faults. Increasing the number of orderer nodes may reduce network performance. To tolerate up to three orderer node failures, their number should be increased to seven. The use of the Raft protocol also helps to mitigate the impact of risks R2 and R3 (Table 1), since they are similar to R1. Still, their effect is weaker, at least because communication is maintained within the institute's LAN, and at most because of possible access to a subset of servers in other institutes (Figure 3). Risk R6 (Table 1) in HLF is mitigated by verifying the digital identity of each network entity and by the ability of both available consensus mechanisms (CFT and BFT) to tolerate server access delays, which are equivalent to risks R1, R2, and R3. The characteristics of risk R5 (Table 1) remain unchanged and are more dependent on the organisation's access policy for authorised personnel than on HLF's built-in capabilities. To mitigate risk R4 (Table 1), we propose switching from CFT (Raft protocol) to BFT (SmartBFT protocol) by utilising HLF's built-in functionality. At the same time, we acknowledge that in a permissioned network such as HLF, resilience against specific cybersecurity threats is lower than in a permissionless network, for example, one that uses PoW.

Among the regression models evaluated by the coefficient of determination (Figure 2), the best-performing model was RFR, yielding a coefficient of determination of 0.74. The poorest performance was observed for PLR (less than -5.7 million). Therefore, the reference value ε_{reg} was computed specifically for RFR. As shown in Figure 3, starting from a "spread" value of approximately 0.64, the MSE for ERBANN becomes quite large and exceeds the plot range. The MSE curves for GRANN and RBANN tend to approach asymptotic values as the "spread" increases: 0.12 and 0.08, respectively. The most noticeable changes in the curves occur within the "spread" parameter range up to approximately 2.5. Among the three dependencies, the MSE for RBANN reaches the smallest value:

3.23×10^{-3} at a "spread" of 1.06, which is lower than ε_{reg}. Therefore, the RBANN model exhibits lower error than RFR and, consequently, than the other regression models. Hence, RBANN was selected for further analysis.

As shown in Figure 4, throughput increases with the number of peers when the number of endorsers remains constant. In this case, the throughput growth rate up to approximately 30 peers follows an O(n) order. For example, increasing the number of peers from 5 to 10 results in an approximate 4.5 % increase in throughput. Beyond 30 peers, throughput growth scales at O(log *n*), and increasing the number of peers from 30 to 60 results in approximately a 12% increase in throughput. This relationship can be used to assess the cost-effectiveness of adding new peers to the network. It is also worth noting that as the number of peers increases, the network's robustness improves. When the number of endorsers increases (while the number of peers remains fixed), the throughput declines more rapidly than when the number of peers decreases (with a fixed number of endorsers). This observation can guide rational adjustments to the network topology.

Additional performance gains can be achieved by deploying CouchDB on a separate server. Provided that the servers have sufficient processing capacity, further performance improvements can be achieved by using two channels instead of one. An interesting area for future research is integrating LevelDB with an off-chain database to enhance HLF (Hyperledger Fabric Docs Main Documentation, n.d.) performance. This approach allows, for example, the deployment of an SQL DB created from HLF while operating outside of it. In this way, significantly better DB management capabilities can be achieved without affecting HLF performance.

Conclusions

This study aimed to develop and justify a blockchain-based proposal to enhance the resilience of USUST's IT infrastructure against hardware-, software-, or network-related issues, as well as cyber threats. The findings of this study demonstrate that Hyperledger Fabric (HLF) can be deployed within the USUST network without significant modifications to its topology, thereby enabling the achievement of the defined objective. We have substantiated the parameters under which HLF delivers sufficient performance to support communication processes associated with the USUST network's educational environment. A key contribution

of this work is the development of eight predictive models: five regression-based and three artificial neural networks (ANNs) that capture the relationship between maximum sustainable throughput and the total number of peers and endorsements. Each model was evaluated, and the optimal one (RBANN) was selected. This model enables the prediction of throughput values as a function of the number of peers and endorsements, thereby facilitating the assessment of resource efficiency in configuring and maintaining the HLF network.

This study is limited to a relatively small number of available datasets, particularly those with peers that exhibit large step sizes at higher numbers (above 16); therefore, the findings cannot be generalised well. Despite this limitation, this study makes a significant contribution by applying it to the current topology of the USUST network, which is well-represented in the mentioned dataset (with fewer than 16 peers). As a result, it is suggested that future research conduct a study with a larger dataset.

References

1. Areshonkov, V. J. (2020). Cyfrovizacija vyshhoji osvity: Vyklyky ta vidpovidi. *Herald of the National Academy of Educational Sciences of Ukraine*, *2*(2). https://doi.org/10.37472/2707-305X-2020-2-2-13-2
2. Capocasale, V., Gotta, D., Perboli, G. (2023). Comparative analysis of permissioned blockchain frameworks for industrial applications. *Blockchain: Research and Applications*, *4*(1), 100113. https://doi.org/10.1016/j.bcra.2022.100113
3. Fernández, A., Gómez, B., Binjaku, K., Meçe, E. K. (2023). Digital transformation initiatives in higher education institutions: A multivocal literature review. *Education and Information Technologies*, *28*(10), 12351-12382. https://doi.org/10.1007/s10639-022-11544-0
4. Guggenberger, T., Sedlmeir, J., Fridgen, G., Luckow, A. (2022). An in-depth investigation of the performance characteristics of Hyperledger Fabric. *Computers & Industrial Engineering*, *173*, 108716. https://doi.org/10.1016/j.cie.2022.108716
5. Guru, A., Mohanta, B. K., Mohapatra, H., Al-Turjman, F., Altrjman, C., Yadav, A. (2023). A Survey on Consensus Protocols and Attacks on Blockchain Technology. *Applied Sciences*, *13*(4), 2604. https://doi.org/10.3390/app13042604
6. Kravtsov, O. V., Kravtsova, T. V., Lashchenko, O. V. (2021). Derzhavna polityka v sferi vyshchoi osvity Ukrainy v umovakh tsyfrovizatsii. *Bulletin of the National*

University of Civil Protection of Ukraine. Series: Public Administration, 1(14), 73–85. https://doi.org/10.52363/2414-5866-2021-1-10

7. Kyrylenko, S. A., Jakimova, M. E. (2025). Dejaki pytannja formuvannja cyfrovoji kompetentnosti zdobuvachiv vyshhoji osvity Ukrajiny ta Poljshhi. *Osvita Ta Rozvytok Obdarovanoji Osobystosti*, *1*, 129-133. https://doi.org/10.63437/2309-3935-2025-1(96)-17
8. Merlec, M. M., In, H. P. (2024). Blockchain-Based Decentralised Storage Systems for Sustainable Data Self-Sovereignty: A Comparative Study. *Sustainability*, *16*(17), 7671. https://doi.org/10.3390/su16177671
9. Nasir, N. M., Hassan, S., Mohd Zaini, K. (2024). Securing Permissioned Blockchain-Based Systems: An Analysis on the Significance of Consensus Mechanisms. *IEEE Access, 12*, 138211-138238. https://doi.org/10.1109/ACCESS.2024.3465869
10. Pavlysh, T., Basarab, V., Tereshhenko, O., Roghiv, M. (2023). Cyfrovizacija osvitnjogho procesu v zakladakh vyshhoji osvity v umovakh vojennogho stanu. *Osvitni Obriji*, *56*(1), 106-109. https://doi.org/10.15330/obrii.56.1.106-109
11. Tomić, N. Z. (2021). A Review of Consensus Protocols in Permissioned Blockchains. *Journal of Computer Science Research*, *3*(2), 19-26. https://doi.org/10.30564/jcsr.v3i2.2921
12. Yuan, F., Huang, X., Zheng, L., Wang, L., Wang, Y., Yan, X., Gu, S., Peng, Y. (2025). The Evolution and Optimization Strategies of a PBFT Consensus Algorithm for Consortium Blockchains. *Information*, *16*(4), 268. https://doi.org/10.3390/info16040268
13. Zhou, S., Li, K., Xiao, L., Cai, J., Liang, W., Castiglione, A. (2023). A Systematic Review of Consensus Mechanisms in Blockchain. *Mathematics*, *11*(10), 2248. https://doi.org/10.3390/math11102248

Netography

1. Ghalushhenko, G. (2024). *Za dvaroky rosija atakuvala ukrajinsjku energhetyku ponad 1000 raziv*. Ministerstvo energhetyky Ukrajiny. Retrieved from: https://mev.gov.ua/novyna/za-dva-roky-rosiya-atakuvala-ukrayinskuenerhetyku-ponad-1000-raziv-herman-halushchenko (dostęp: 02.09.2025).
2. Hyperledger Fabric. (n.d.). *A Blockchain Platform for the Enterprise*. Retrieved from: https://hyperledger-fabric.readthedocs.io/en/latest/ (access: 02.09.2025).

3. UDUNT. (n.d.). *Istorija uniwersytetu*. Retrieved from: https://ust.edu.ua/universytet/istoriya-universytetu/ (access: 02.09.2025).
4. Чайка, О. (2023). *Kiberataky na energhosystemu Ukrajiny: Jak kompaniji protystojatj khakeram?* Forbes. Retrieved from: https://forbes.ua/company/rosiyski-khakerikoordinuyut-dii-z-viyskovimi-ta-posilyuyut-ataki-naperedodni-zimi-yakukraina-protistoit-kiberatakam-na-energosistemu-08112023-17242 (access: 02.09.2025).

Barbara Drapikowska[1], Robert Zakrzewski[2]

Akademia Sztuki Wojennej w Warszawie

https://doi.org/10.29316/9788368103304_9

ZAUFANIE I LĘK WOBEC SZTUCZNEJ INTELIGENCJI – PSYCHOLOGICZNE MECHANIZMY A ODPOWIEDZIALNE WDRAŻANIE TECHNOLOGII

TRUST AND FEAR OF ARTIFICIAL INTELLIGENCE – PSYCHOLOGICAL MECHANISMS AND RESPONSIBLE IMPLEMENTATION OF TECHNOLOGY

Streszczenie

Celem rozdziału jest przedstawienie psychologicznych mechanizmów determinujących postawy wobec sztucznej inteligencji oraz określenie ich konsekwencji dla odpowiedzialnego i bezpiecznego wdrażania technologii w życie społeczne. Problem badawczy koncentruje się wokół pytania: jakie mechanizmy psychologiczne determinują postawy wobec AI oraz wpływają na odpowiedzialne i bezpieczne wdrażanie tej technologii w kontekście społecznym? Opracowanie ma charakter teoretyczny i opiera się na analizie aktualnej literatury z zakresu psychologii społecznej, poznawczej oraz interakcji człowiek – technologia.

[1] https://orcid.org/0000-0002-6852-2791, e-mail: b.drapikowska@akademia.mil.pl
[2] https://orcid.org/0000-0002-4019-9048, e-mail: r.zakrzewski@akademia.mil.pl

Przeprowadzone rozważania wskazują, że zaufanie i lęk wobec AI mają wspólne źródła psychologiczne, lecz mogą prowadzić do przeciwstawnych postaw, tj. od gotowości do współpracy po unikanie technologii. Zrozumienie tych procesów jest kluczowe dla wspierania odpowiedzialnego wdrażania sztucznej inteligencji w społeczeństwie.

Słowa kluczowe: sztuczna inteligencja (AI), antropomorfizacja, heurystyki poznawcze, psychologia technologii

Abstract

The aim of this chapter is to present the psychological mechanisms that determine attitudes toward artificial intelligence and to define their consequences for the responsible and safe implementation of technology in social life. The research problem is: What psychological mechanisms determine attitudes toward AI and influence the responsible, safe implementation of this technology in a social context? The study is theoretical and based on an analysis of current literature in social and cognitive psychology and human-technology interaction. The considerations presented indicate that trust and fear of AI have common psychological sources, but can lead to opposing attitudes, i.e., from a willingness to cooperate to avoidance of technology. Understanding these processes is key to supporting the responsible implementation of artificial intelligence in society.

Keywords: artificial intelligence (AI), anthropomorphism, cognitive heuristics, technology psychology

Wstęp

Dynamiczny rozwój technologii sztucznej inteligencji (AI) oraz jej integracja z codziennymi praktykami społecznymi, obejmująca tak zróżnicowane obszary jak zdrowie, bezpieczeństwo, edukacja i dostęp do usług, rodzi szereg pytań dotyczących sposobu, w jaki ludzie postrzegają i interpretują działanie systemów autonomicznych (West, Whittaker, Crawford, 2019). Użytkownicy często przypisują systemom AI cechy psychiczne, tj. intencje czy emocje, w procesie określanym jako antropomorfizacja. Zjawisko to wpływa zarówno na emocjonalne reakcje wobec maszyn, jak i na ocenę ich kompetencji oraz społecznego znaczenia (Złotowski, Proudfoot, Yogeeswaran, Bartneck, 2015).

Dynamiczny rozwój technologii sztucznej inteligencji oraz jej integracja z codziennymi praktykami społecznymi rodzi szereg pytań dotyczących sposobu, w jaki ludzie postrzegają i interpretują działanie systemów autonomicznych (West i in., 2019). W konsekwencji relacja człowiek – AI staje się obszarem silnego napięcia poznawczego i normatywnego. AI zaczyna być postrzegana nie tylko jako narzędzie, ale także jako potencjalny „aktor społeczny", wpływający na decyzje dotyczące obywateli i instytucji (Waytz, Cacioppo, Epley , 2010).

Antropomorfizacja może prowadzić do nadmiernego zaufania wobec systemów sztucznej inteligencji, nawet wówczas, gdy brak jest podstaw do zakładania ich intencjonalności. Przypisywanie AI podmiotowości może tworzyć fałszywe przekonanie o odpowiedzialności moralnej maszyn, co przesuwa ten ciężar z projektantów i operatorów na technologie (Bryson, 2018). Jednocześnie ludzie mają tendencję do postrzegania algorytmów jako bardziej obiektywnych i kompetentnych niż ludzie (Logg, Minson, Moore, 2019), nawet wówczas, kiedy nie potrafią ocenić poprawności ich działania.

Obecnie, kiedy systemy AI są integrowane z medycyną, edukacją, bezpieczeństwem publicznym czy rynkiem pracy, ich społeczna rola ulega gruntownej transformacji. Automatyczne podejmowanie decyzji w obszarach wcześniej zarezerwowanych dla człowieka wpływa na to, jak ludzie rozumieją sprawstwo, odpowiedzialność i kompetencje technologii. Coraz częściej podkreśla się, że stosunek ludzi do technologii zależy nie tylko od jej użyteczności, lecz także od potrzeb psychologicznych, takich jak poczucie bezpieczeństwa, przewidywalności czy kontroli. Postawy wobec AI mogą wynikać zarówno z realnych doświadczeń, jak i z przekazów medialnych oraz narracji kulturowych wzmacniających obawy przed utratą pracy, autonomii lub prywatności. Ambiwalencja ta sprawia, że ludzie jednocześnie oczekują od AI perfekcyjnego działania i obawiają się jej dominacji.

Systemy AI projektowane z cechami humanoidalnymi mogą wywoływać jeszcze silniejsze reakcje społeczne, tj. od sympatii i opiekuńczości po lęk przed utratą kontroli. Elementy formy i ruchu robotów automatycznie nasilają przypisywanie im cech mentalnych, co może prowadzić również do nadawania im osobowości czy wartości moralnej (Złotowski i in., 2015). W efekcie antropomorfizacja zaczyna wpływać nie tylko na ocenę użyteczności AI, lecz także na jej status nie jako technologii, lecz jako potencjalnego uczestnika życia społecznego.

W konsekwencji relacja człowiek – AI staje się obszarem silnego napięcia poznawczego i normatywnego. AI zaczyna być postrzegana nie tylko jako narzędzie, ale także jako potencjalny „aktor społeczny", wpływający na decyzje dotyczące obywateli i instytucji (West i in., 2019). Dlatego właśnie psychologia staje przed wyzwaniem opracowania ram teoretycznych pozwalających wyjaśnić, jakie procesy poznawcze i emocjonalne prowadzą do akceptacji lub odrzucenia technologii o wysokiej autonomii.

Zrozumienie tych procesów jest niezbędne nie tylko dla poszerzenia wiedzy teoretycznej, lecz także dla projektowania systemów, które będą budziły adekwatny poziom zaufania społecznego, jednocześnie nie wzmacniając nierealistycznych oczekiwań czy obaw.

Celem niniejszego rozdziału jest przedstawienie psychologicznych mechanizmów determinujących postawy wobec sztucznej inteligencji oraz określenie ich konsekwencji dla odpowiedzialnego i bezpiecznego wdrażania technologii w życie społeczne. Główny problem badawczy został skonstruowany w postaci pytania: jakie mechanizmy psychologiczne determinują postawy wobec AI oraz wpływają na odpowiedzialne i bezpieczne wdrażanie tej technologii w kontekście społecznym?

Na tej podstawie skonstruowano problematykę szczegółową obejmującą następujące kwestie:

1. Jakie są psychologiczne mechanizmy antropomorfizacji i zaufania do AI?
2. Jakie są psychologiczne mechanizmy lęku przed AI?
3. Jakie są psychologiczne konsekwencje oraz praktyczne implikacje dla projektowania zaufanej AI?

W kolejnych częściach rozdziału podjęto próbę rozwiązania przyjętych problemów badawczych poprzez analizę teoretyczną oraz prezentację wyników badań własnych.

Psychologiczne mechanizmy antropomorfizacji i zaufania do AI

W interakcjach człowieka z systemami sztucznej inteligencji istotną rolę odgrywają uproszczone mechanizmy poznawcze, które z jednej strony ułatwiają szybkie podejmowanie decyzji, lecz jednocześnie mogą prowadzić do błędnych ocen działania technologii. Tversky i Kahneman (1974) wykazali, że ludzie,

podejmując decyzje w warunkach niepewności, opierają się na heurystykach[3], tj. automatycznych skrótach myślowych, takich jak heurystyka reprezentatywności, dostępności czy zakotwiczenia. Mechanizmy te znajdują swoje odzwierciedlenie również w ocenach kompetencji AI. Innymi słowy, użytkownicy mogą ufać systemom informatycznym, ponieważ „wydają się" dokładne i obiektywne, nawet jeśli nie rozumieją zasad ich działania.

Lee i See (2004) podkreślają, że zaufanie do automatyzacji jest funkcją trzech czynników, tj. postrzeganej kompetencji, przewidywalności oraz intencji systemu. W praktyce oznacza to, że jeśli technologia szybko i poprawnie dostarcza odpowiedzi, użytkownik może ulegać iluzji niezawodności, nawet przy braku transparentności działania czynności algorytmu. W sytuacji błędu ludzie nierzadko reagują skrajnie, np. nagła utrata zaufania prowadzi do całkowitego odrzucenia systemu, mimo że wcześniejsze wyniki były korzystne.

Badania Dietvorsta i innych (2015) pokazują, że wystarczy jedna widoczna pomyłka algorytmu, aby zaufanie użytkowników spadło dramatycznie, nawet jeśli system statystycznie pozostaje dokładniejszy niż człowiek. Autorzy wskazują, że ludzie surowiej oceniają błędy popełnione przez algorytmy niż przez ludzi. Efekt ten ilustruje szersze zjawisko – mianowicie, decyzje dotyczące akceptacji technologii częściej opieramy na ocenie emocjonalnej, a nie na racjonalnym rachunku skuteczności.

Równocześnie istnieją również tendencje odwrotne. W niektórych kontekstach obserwuje się tendencję do bezkrytycznego polegania na systemach automatycznych (tzw. *automation bias*) nawet wtedy, gdy użytkownik ma możliwość zakwestionowania wyników (Logg, 2019). Oznacza to, że heurystyki mogą

[3] Heurystyka – uproszczona reguła poznawcza lub strategia myślenia, którą ludzie stosują, by podejmować szybkie decyzje lub wyciągać wnioski w sytuacjach niepewności lub ograniczonej informacji. Heurystyki zmniejszają obciążenie poznawcze, ale mogą prowadzić do systematycznych błędów (np. przewidywania, szacowania prawdopodobieństwa). Heurystyka reprezentatywności polega na tym, że oceniamy prawdopodobieństwo zdarzenia na podstawie tego, jak bardzo przypomina ono nasz prototyp w danej kategorii (zamiast na podstawie realnych danych statystycznych). Heurystyka dostępności polega na tym, że oceniamy częstotliwość lub ryzyko zdarzeń na podstawie tego, jak łatwo przychodzą nam na myśl przykłady, a nie na podstawie rzeczywistych statystyk. Heurystyka zakotwiczenia polega na tym, że przy podejmowaniu decyzji opieramy się na pierwszej dostępnej informacji (kotwicy), nawet jeśli jest przypadkowa lub niewiarygodna (Wojciszke, 2006).

prowadzić zarówno do nadmiernego polegania na AI, jak i do nagłego odrzucenia technologii, zależnie od doświadczeń użytkownika i charakteru zadania.

Należy zatem zaznaczyć, że decyzje dotyczące korzystania z AI nie są wynikiem chłodnego namysłu, lecz są warunkowane przez poznawcze skróty i uproszczenia. To sprawia, że zaufanie społeczne wobec sztucznej inteligencji jest zmienne, podatne na błędy oceny, a także podatne na manipulację, np. przez design interfejsu lub sposób prezentacji wyników działania systemu.

Jednym z kluczowych czynników wpływających na zaufanie do systemów sztucznej inteligencji jest antropomorfizacja, rozumiana jako przypisywanie cech ludzkich podmiotom nieożywionym. Ludzie wykazują naturalną tendencję do interpretowania zachowań technologii poprzez pryzmat kategorii psychologicznych, tj. intencji, emocji czy motywacji. Mechanizm ten ułatwia przewidywanie działania AI, a zarazem wzmacnia poczucie bezpieczeństwa w interakcji z nią (Waytz i in., 2010). W tym sensie antropomorfizacja pełni funkcję poznawczą i społeczną, czyli pozwala traktować AI jak partnera, a nie skomplikowany system informatyczny.

Badania pokazują, że elementy wyglądu i zachowania maszyn, takie jak humanoidalna forma, płynne ruchy czy naturalnie brzmiący głos, zwiększają skłonność użytkowników do przypisywania robotom stanów mentalnych. Złotowski i inni (2015) podkreślają, że już minimalne sygnały społeczne mogą automatycznie uruchamiać procesy mentalizacji, w wyniku których AI postrzegana jest jako bardziej kompetentna i godna zaufania. Co istotne, antropomorfizacja może skutkować również nadawaniem technologii elementów osobowości oraz wartości moralnej, co stanowi podstawę do traktowania jej jako „aktora społecznego".

Zaufanie budowane na antropomorficznych cechach AI niesie jednak ryzyko nadmiernego polegania na systemach, które nie posiadają realnych zdolności intencjonalnych. Bryson (2018) ostrzega, że traktowanie AI jako bytu moralnego jest nie tylko błędem poznawczym, lecz także może prowadzić do niebezpiecznego „przesuwania" odpowiedzialności z ludzi na technologie. Ludzie przypisują maszynom autonomię tam, gdzie w rzeczywistości decyzje pozostają wynikiem implementacji określonych algorytmów.

Co więcej, wysoki poziom antropomorfizacji nie zawsze sprzyja zaufaniu. Nadmiernie „ludzka" forma może wywoływać tzw. efekt niesamowitości (ang. *uncanny valley*), prowadząc do dyskomfortu i dystansu społecznego (Złotowski

i in., 2015). W takich przypadkach użytkownicy mogą reagować niechęcią lub wręcz nieufnością wobec AI, obawiając się naruszenia granic między tym, co ludzkie, a tym, co technologiczne. Efekt *uncanny valley* opisuje gwałtowny spadek akceptacji oraz wzrost niepokoju wobec robotów lub awatarów, które wyglądają niemal jak ludzie, lecz nadal są odbierane jako nienaturalne. W takich sytuacjach silne podobieństwo przyciąga uwagę, ale jednocześnie wywołuje niezgodność percepcyjną, prowadząc do lęku i odrzucenia technologii (Mori, 2012).

W badaniach C. Christoforakos i innych (2021) wykazano, że zaufanie wzrasta, gdy roboty oceniane są jako jednocześnie kompetentne i „ciepłe", czyli budzące pozytywne skojarzenia społeczne. Innymi słowy, skuteczność technologii nie wystarcza, gdyż użytkownik musi jeszcze postrzegać ją jako przyjazną i ukierunkowaną na dobro człowieka.

Warto zaakcentować, że antropomorfizacja stanowi jeden z fundamentów zaufania do AI, ale jej skutki są ambiwalentne. Może ułatwiać interakcję i akceptację technologii, jednocześnie prowadząc do nieadekwatnych oczekiwań i zaburzonej oceny odpowiedzialności. Kluczowe jest więc projektowanie systemów AI tak, aby wsparcie społeczne i emocjonalne nie odbywało się kosztem przejrzystości i bezpieczeństwa.

Nie sposób zaprzeczyć, że psychologiczne reakcje na sztuczną inteligencję i systemy automatyczne cechuje wysoka zmienność. W niektórych sytuacjach użytkownicy z dużą gotowością powierzają im zadania decyzyjne, uznając algorytmy za bardziej precyzyjne, bezstronne i przewidywalne niż oceny ludzkie. Taką tendencję określa się mianem *algorithm appreciation*, czyli preferencji wobec algorytmów (Logg, 2019). W innych przypadkach jednak nawet pojedynczy błąd systemu prowadzi do gwałtownego obniżenia zaufania i odrzucenia technologii – zjawiska znanego jako *algorithm aversion*, czyli awersja wobec algorytmów. Oba mechanizmy stanowią dwa przeciwstawne bieguny społecznego zaufania do technologii i odzwierciedlają sposób, w jaki ludzie interpretują kompetencje, intencje oraz ograniczenia systemów sztucznej inteligencji (Dietvorst i in., 2015).

Z jednej strony ludzie często zakładają, że algorytmy są bardziej obiektywne i wolne od emocji, co prowadzi do nadmiernego zaufania do ich rekomendacji. W kontekstach wymagających analizy danych i logiki, takich jak medycyna, systemy rekomendacyjne czy ocena ryzyka, automatyzacja bywa odbierana jako bardziej kompetentna niż człowiek (Dietvorst i in., 2015). Mechanizm ten może

skutkować bezkrytycznym poleganiem na AI, nawet gdy użytkownik dostrzega ograniczoną przejrzystość działania systemu.

Lee i See (2004) podkreślają, że zaufanie do automatyzacji opiera się głównie na postrzeganej kompetencji i przewidywalności. Gdy AI zachowuje się zgodnie z oczekiwaniami, pojawia się entuzjazm i gotowość do współpracy. Gdy jednak system zaskakuje, zwłaszcza negatywnie, reakcja emocjonalna prowadzi do gwałtownej utraty zaufania. Paradoksalnie więc to, co początkowo buduje zaufanie do AI, może stać się źródłem jego załamania.

Jednocześnie pojawia się fascynacja samą możliwością przekraczania ludzkich ograniczeń. W wielu kontekstach użytkownicy przypisują AI rolę sprzymierzeńca, zdolnego uzupełniać niedoskonałości ludzkiego poznania, zwiększać bezpieczeństwo i poprawiać komfort życia. Ta pozytywna perspektywa może wzmacniać gotowość do integracji technologii z codziennością, zwłaszcza jeśli AI zostaje zaprojektowana tak, aby budzić poczucie przewidywalności, kompetencji oraz „życzliwych intencji".

Przytoczone wcześniej zjawiska wskazują, że relacja człowiek – AI ma charakter niestabilny, czyli zależny nie tylko od faktycznych możliwości systemu, lecz także od oczekiwań, wcześniejszych doświadczeń oraz sposobu prezentacji wyników. Ta zmienność ma kluczowe znaczenie dla projektowania technologii, ponieważ nadmierne lub niewystarczające zaufanie może prowadzić do poważnych konsekwencji społecznych oraz etycznych.

Psychologiczne mechanizmy lęku przed AI

Postawy wobec sztucznej inteligencji rzadko są jednoznacznie pozytywne lub negatywne. Bez wątpienia w tej kwestii dominują emocje ambiwalentne, łączące zarówno entuzjazm wobec możliwości technologicznych, jak i obawy dotyczące ich konsekwencji dla człowieka oraz społeczeństwa. Z jednej strony AI bywa postrzegana jako narzędzie wspierające rozwój medycyny, edukacji czy bezpieczeństwa publicznego. Z drugiej jednak, budzi lęk przed utratą kontroli, dehumanizacją relacji społecznych oraz zagrożeniem dla pozycji człowieka w hierarchii sprawczości, choćby przez utratę pozycji zawodowej.

Jednym z kluczowych czynników generujących emocjonalną nieufność wobec zaawansowanych technologii jest lęk egzystencjalny. Zgodnie z Teorią

Opanowywania Trwogi[4] (ang. *Terror Management Theory*; TMT), świadomość własnej śmiertelności skłania ludzi do obrony symbolicznych struktur dających poczucie sensu i bezpieczeństwa (Greenberg, Solomon, Pyszczynski, 1997). AI, zdolna zastępować ludzkie zdolności poznawcze, może zostać odebrana jako zagrożenie dla wyjątkowości człowieka i jego statusu jako jedynego twórcy oraz interpretatora rzeczywistości (Greenberg i in., 1997). W rezultacie negatywne emocje mogą pojawiać się nie tylko w związku z możliwą utratą pracy, lecz także z bardziej fundamentalnym poczuciem, że rozwój AI „umniejsza" pozycję człowieka w świecie.

Wang i Siau (2019) zauważają, że emocje związane z AI wynikają z napięcia między nadzieją na poprawę jakości życia a poczuciem zagrożenia dla ludzkiej autonomii. Przekazy medialne i narracje kulturowe dodatkowo wzmacniają te reakcje. Obrazy przyjaznych, humanoidalnych robotów mogą wywoływać sympatię i ciekawość, natomiast wizje „buntu maszyn" – niepokój i opór. W konsekwencji każda nowa technologia AI zostaje szybko „osadzona" w istniejących schematach emocjonalnych.

Ambiwalencja ta ma znaczące konsekwencje dla procesów decyzyjnych. Silne emocje, zarówno pozytywne, jak i negatywne, mogą zniekształcać ocenę ryzyka i korzyści, prowadząc do tendencji skrajnych, czyli bezrefleksyjnej akceptacji lub nieuzasadnionego odrzucenia technologii. Zrozumienie psychologicznego podłoża tych reakcji jest zatem kluczowe dla projektowania systemów AI, które będą wzbudzały zaufanie adekwatne do ich rzeczywistych możliwości.

Jednak, jak wskazali Dietvorst i inni (2015), wystarczy jeden widoczny błąd algorytmu, aby zaufanie użytkownika spadło gwałtownie, czyli znacznie silniej, niż w przypadku błędów ludzkich. Ludzie surowiej oceniają technologię niż człowieka za tę samą pomyłkę, co autorzy tłumaczą oczekiwaniem perfekcyjności działania systemów sztucznej inteligencji (Greenberg i in., 1997). W efekcie

[4] Teoria opanowywania trwogi zakłada, że ludzie są świadomi własnej śmiertelności i w związku z tym doświadczają potencjalnie silnego lęku egzystencjalnego. Aby ten lęk zredukować i utrzymać psychologiczną stabilność, tworzą systemy kulturowych przekonań, które nadają życiu sens, porządek i trwałość. Kluczową rolę pełni również poczucie własnej wartości (samoocena), będące psychologiczną ochroną przed doświadczeniem trwogi. Gdy systemy światopoglądowe lub poczucie wartości zostają zagrożone, pojawiają się nasilone reakcje obronne, takie jak podtrzymywanie własnych przekonań, uprzedzenia czy niechęć wobec grup i obiektów postrzeganych jako zagrażające (Wojciszke, 2006).

użytkownicy mogą całkowicie odrzucić rozwiązania algorytmiczne, mimo że ich przeciętna skuteczność przewyższa skuteczność ludzką.

Postawy wobec sztucznej inteligencji nie kształtują się wyłącznie na poziomie jednostkowym. Istotną rolę odgrywa szerszy kontekst społeczny, czyli normy kulturowe, przekazy medialne oraz dyskurs publiczny, które wspólnie wpływają na sposób postrzegania AI. Wang i Siau (2019) zwracają uwagę, że nowe technologie są zawsze osadzone w realiach społecznych, dlatego emocje i oczekiwania wobec AI wynikają również z narracji dotyczących przyszłości pracy, utraty kontroli czy transformacji relacji międzyludzkich. W tym ujęciu sztuczna inteligencja staje się symbolem szans i zagrożeń społecznych, odzwierciedlając dominujące przekonania na temat rozwoju cywilizacyjnego.

Kulturowa interpretacja AI wiąże się również z mechanizmami obrony tożsamości. Zgodnie z wspomnianą wcześniej TMT, ludzie poszukują stabilnych ram znaczeniowych, które chronią poczucie sensu i wartości. Technologie, które mogą zakwestionować wyjątkowość człowieka, np. poprzez wykazywanie wyższej skuteczności w zadaniach intelektualnych, wywołują reakcje obronne, szczególnie tam, gdzie centralną wartością jest autonomia i sprawczość jednostki (Greenberg, 1997). W społeczeństwach silnie zorientowanych na indywidualizm obawy te mogą przybierać bardziej intensywną formę.

Znaczący wpływ na oceny AI mają także media i popkultura, które często wykorzystują skrajne narracje, tj. od idealistycznych wizji wszechmocnych technologii rozwiązujących globalne problemy, po katastroficzne scenariusze dominacji maszyn. Tego rodzaju przekazy wzmacniają emocjonalne bieguny postaw, tj. od entuzjastycznej akceptacji po silny opór i lęk. W rezultacie część społeczeństwa traktuje AI jako źródło nadziei, inni z kolei, jako zagrożenie dla porządku społecznego.

Istotne są również różnice wynikające z poziomu kompetencji cyfrowych. Osoby lepiej rozumiejące zasady działania systemów informatycznych wykazują większą zdolność do adekwatnej oceny ryzyka i korzyści, podczas gdy brak wiedzy może prowadzić do uogólnionego strachu lub przeciwnie – bezrefleksyjnej fascynacji technologią. To sugeruje, że edukacja w obszarze AI może pełnić kluczową rolę w kształtowaniu dojrzałego, zrównoważonego zaufania społecznego.

Podsumowując, ocena sztucznej inteligencji jest wynikiem złożonej gry wpływów jednostkowych emocji, przekonań kulturowych oraz narracji

społecznych. AI nie jest zatem postrzegana jedynie jako narzędzie, lecz jako aktor symboliczny, angażujący wartości, tożsamość i wyobrażenia o przyszłości człowieka w świecie.

Konsekwencje psychologiczne oraz implikacje praktyczne dla projektowania zaufanej AI

Właściwie ukształtowane zaufanie do sztucznej inteligencji może przynosić istotne korzyści zarówno jednostkom, jak i całym społeczeństwom. Gdy użytkownicy postrzegają system jako wystarczająco kompetentny i przewidywalny, są bardziej skłonni angażować się w współpracę człowiek – technologia, co przekłada się na poprawę skuteczności działań i redukcję obciążenia poznawczego (Lee, See, 2004). W obszarach wymagających przetwarzania dużych zbiorów danych AI może wspierać podejmowanie decyzji, ograniczając ryzyko błędów wynikających z ograniczeń ludzkiej pamięci i uwagi.

Zaufanie oparte na racjonalnych przesłankach zwiększa również bezpieczeństwo, szczególnie w zadaniach o podwyższonym ryzyku. Od medycyny po transport – automatyzacja może pełnić funkcję wsparcia w sytuacjach krytycznych, oferując szybszą analizę danych oraz natychmiastowe ostrzeganie o zagrożeniach. W takich przypadkach AI nie zastępuje człowieka, lecz pomaga mu podejmować decyzje bardziej świadomie i precyzyjnie (Lee, See, 2004).

Istotną korzyścią jest także zmniejszenie obciążenia emocjonalnego użytkowników. W systemach, w których algorytmy przyjmują część odpowiedzialności za analizę i rekomendacje, jednostka może doświadczać większego poczucia kontroli i bezpieczeństwa (Dietvorst i in., 2015). Jednocześnie pozytywne emocje, takie jak ciekawość i satysfakcja z kontaktu z technologią, mogą wzmacniać motywację do nauki i adaptacji nowych narzędzi.

W kontekście społecznym właściwe zaufanie do AI może przyczyniać się do zwiększenia dostępności usług, szczególnie w obszarze edukacji i opieki zdrowotnej. Systemy wspomagające komunikację, personalizację treści czy monitorowanie stanu zdrowia mogą poprawiać jakość życia osób z ograniczeniami funkcjonalnymi oraz zmniejszać nierówności wynikające z dostępu do specjalistów.

Warunkiem uzyskania tych korzyści jest adekwatna kalibracja zaufania, tj. zgodna z realnymi możliwościami i ograniczeniami systemu. Gdy użytkownik

ufa technologii na poziomie właściwym, proces współpracy człowiek – AI przebiega harmonijnie, a emocje i heurystyki poznawcze działają wspierająco, a nie zakłócająco. W rezultacie AI może stać się nie tylko narzędziem, ale partnerem, który wzmacnia ludzkie zdolności poznawcze, emocjonalne i społeczne.

Zaufanie do sztucznej inteligencji, choć pożądane z perspektywy efektywności i bezpieczeństwa, może również prowadzić do szeregu zagrożeń, jeśli nie zostanie odpowiednio skalibrowane. Niewłaściwe zaufanie można rozpatrywać jako nadmierne lub niedostateczne, przy czym każda z tych postaw wiąże się z odmiennymi konsekwencjami dla użytkownika i środowiska społecznego.

Nadmierne zaufanie (ang. *overtrust*) sprzyja bezkrytycznemu poleganiu na technologii, szczególnie w sytuacjach, w których użytkownik uznaje system za domyślnie bardziej kompetentny od człowieka (Lee, See, 2004). W rezultacie decyzje o dużym znaczeniu podejmowane są automatycznie, bez monitorowania i weryfikacji wyników działania algorytmu. Zjawisko to określa się mianem wspomnianego wcześniej *automation bias*. W obszarach takich jak medycyna czy transport może to prowadzić do poważnych błędów operacyjnych oraz zagrożeń dla zdrowia i życia.

Drugą skrajnością jest niedostateczne zaufanie, czyli unikanie lub odrzucanie technologii pomimo obiektywnie wyższej skuteczności systemów algorytmicznych w porównaniu z ludźmi. Przywołane badania dają podstawę, by sądzić, że nawet pojedynczy błąd prowadzi do gwałtownego spadku zaufania do AI, znacznie większego niż w przypadku błędów ludzkich. Postawa ta, określana jako *algorithm aversion*, może skutkować rezygnacją z innowacyjnych rozwiązań, które mogłyby poprawić jakość podejmowanych decyzji, np. w diagnozie medycznej czy analizie danych krytycznych.

Obie skrajne reakcje wynikają często z błędnych oczekiwań wobec technologii. Nadmierne zaufanie opiera się na przekonaniu, że AI jest nieomylna, z kolei niedostateczne, że jej błędy są bardziej znaczące niż błędy ludzkie. Tymczasem w literaturze podkreśla się, że zaufanie powinno być funkcją rzeczywistych możliwości systemu, uwzględniając zarówno jego mocne strony, jak i ograniczenia (Lee, See, 2004). Zachwianie tej równowagi prowadzi do dezorganizacji relacji człowiek – technologia.

Dodatkowym zagrożeniem jest przesunięcie odpowiedzialności. Gdy systemy prezentowane są jako inteligentne i autonomiczne, ludzie mogą przypisywać

im winę za błędne decyzje, unikając odpowiedzialności osobistej lub instytucjonalnej (Bryson, 2018). Diagnostyczna aura obiektywności przypisana algorytmom utrudnia krytyczną ewaluację wyników i stawia przed projektantami etyczne zobowiązanie do zapewnienia transparentności i nadzoru.

Aby zaufanie do sztucznej inteligencji było adekwatne, tj. ani zbyt wysokie, ani zbyt niskie, konieczne jest projektowanie systemów w sposób, który uwzględnia ograniczenia ludzkiej percepcji, emocji i procesów poznawczych. Lee i See (2004) podkreślają, że kluczowymi czynnikami kształtującymi zaufanie są kompetencja, przewidywalność oraz postrzegane intencje systemu. Oznacza to, że użytkownicy muszą rozumieć zarówno to, co technologia potrafi, jak również to, czego nie jest w stanie zrobić.

Jednym z najważniejszych elementów projektowania jest transparentność. Systemy AI powinny przedstawiać użytkownikowi czytelne informacje o zastosowanych danych, zakresie działania oraz poziomie pewności modelu. W procesie tym ogromną rolę odgrywa nurt *Explainable AI* (XAI), rozwijający narzędzia umożliwiające zrozumienie logiki decyzji algorytmu nawet osobom bez zaawansowanej wiedzy technicznej (Lee, See, 2004). Przejrzystość zmniejsza niepewność i obawy, co może zapobiegać zarówno nadmiernemu, jak i niedostatecznemu zaufaniu.

Istotnym obszarem jest również projektowanie komunikatów o błędach. Skoro ludzie tracą zaufanie do AI po jednym błędzie, odpowiednie projektowanie komunikacji nabiera dużego znaczenia. W tym kontekście można postulować wyjaśnianie przyczyn błędów i sposobów ich korekcji, co może łagodzić ten efekt i wzmacniać poczucie kontroli użytkownika.

Z kolei elementy antropomorficzne należy stosować w sposób funkcjonalny, a nie wyłącznie estetyczny. R. Złotowski i inni (2015) podkreślają, że sygnały społeczne AI mogą poprawiać jakość współpracy i budować zaufanie. Nadmierna antropomorfizacja może jednak prowadzić do błędnego przypisywania autonomii i intencji, co skutkuje niebezpiecznym przesunięciem odpowiedzialności (Bryson, 2018). Dlatego antropomorfizacja powinna być skalowalna i sytuacyjna, np. w robotach społecznych – uzasadniona, a w aplikacjach medycznych – ograniczona.

W projektowaniu systemów AI nie można pomijać norm kulturowych i społecznych. Jak wskazują Wang wraz z Siau (2019), akceptacja technologii różni się w zależności od kontekstu, narracji publicznych i poziomu kompetencji

cyfrowych. Z tego względu wdrożenie tego samego rozwiązania może przynieść odmienne efekty w różnych populacjach, zatem projekt musi być kontekstowy, oparty na badaniach użytkowników.

Można zatem przyjąć, że adekwatne zaufanie do AI wymaga zintegrowania wiedzy z zakresu psychologii, interakcji człowiek – technologia i etyki projektowania. Systemy powinny być przewidywalne, zrozumiałe i odpowiedzialne, a użytkownicy wyposażeni w narzędzia do świadomego nadzoru nad technologią. Tylko wówczas współpraca człowieka i AI będzie wspierała bezpieczeństwo społeczne i rozwój, zamiast generować nowe zagrożenia.

Warto również podkreślić, że rozwój systemów sztucznej inteligencji rodzi pytania nie tylko o zaufanie użytkowników, lecz także o odpowiedzialność za skutki decyzji podejmowanych z udziałem tych technologii. W wielu zastosowaniach AI jest postrzegana jako podmiot autonomiczny, zdolny działać „samodzielnie", co może prowadzić do błędnego przypisywania jej intencji oraz moralnej sprawczości. Bryson (2018) ostrzega, że traktowanie AI jako bytu etycznego jest nie tylko nieuzasadnione, ale również może skutkować niebezpiecznym odsunięciem odpowiedzialności od ludzi. Projektanci, operatorzy i instytucje wdrażające systemy muszą więc pozostać centralnymi aktorami odpowiedzialności etycznej i prawnej.

W wielu kontekstach AI pełni rolę doradczą, wpływając na decyzje człowieka poprzez rekomendacje, klasyfikacje czy przewidywania. Taka konfiguracja relacji rodzi ryzyko, że użytkownicy, zwłaszcza w sytuacjach nadmiernego zaufania, będą stawać się zależni od algorytmu, ograniczając własny nadzór i refleksyjność. Jednocześnie niewystarczające zaufanie może powodować odrzucenie technologii o wysokiej przydatności społecznej. Dlatego kluczowe znaczenie ma jasne określenie granic roli AI tak, aby jej udział wspierał decydenta, a nie zastępował jego odpowiedzialności (Lee, See, 2004).

Z punktu widzenia projektowania etycznego ważne jest zapewnienie zrozumiałych reguł funkcjonowania systemu, obejmujących m.in. informacje o jakości danych, poziomie błędu oraz o tym, kto ponosi odpowiedzialność za skutki działania algorytmu (Lee, See, 2004). W przypadku błędów AI, czy to w ocenie medycznej, czy w decyzjach dotyczących bezpieczeństwa, nie można dopuścić do sytuacji, w której odpowiedzialność staje się „niczyja". AI nie odczuwa bowiem

konsekwencji własnych decyzji i nie może zostać pociągnięta do odpowiedzialności moralnej ani prawnej (Bryson, 2018).

Granice sprawczości AI muszą być także komunikowane użytkownikom, w przeciwnym razie ryzykujemy zjawisko „moralnego rozmycia odpowiedzialności". Odpowiedzialne projektowanie zakłada stałą obecność człowieka „na pętli decyzyjnej", szczególnie w obszarach związanych z etyką, zdrowiem publicznym oraz bezpieczeństwem. Współpraca człowiek – AI powinna być zorganizowana tak, aby technologia wzmacniała kompetencje ludzi, a nie przyczyniała się do ich zastępowania lub degradacji.

Zaufanie do sztucznej inteligencji jest złożonym konstruktem psychologicznym, którego konsekwencje mogą być zarówno korzystne, jak i potencjalnie niebezpieczne dla jednostki oraz społeczeństwa. Odpowiednio skalibrowane zaufanie wspiera efektywność działań, poprawia bezpieczeństwo i pozwala wykorzystać mocne strony technologii w różnych dziedzinach życia, od medycyny po edukację. Jednak niewłaściwe zaufanie, zarówno nadmierne, jak i niedostateczne, może prowadzić do poważnych błędów decyzyjnych, utraty nadzoru nad systemem lub odrzucania rozwiązań, które mogłyby przynieść istotne korzyści. Dlatego projektowanie AI musi uwzględniać ludzkie ograniczenia poznawcze, emocjonalne i społeczne, zapewniając jednocześnie przejrzystość działania technologii, informowanie o jej możliwościach oraz wyraźne określenie granic odpowiedzialności człowieka i maszyny. Tylko wówczas współpraca człowiek – AI będzie sprzyjać rozwojowi i dobrostanowi, a nie generować nowych zagrożeń.

Podsumowanie

Celem rozdziału było ukazanie psychologicznych mechanizmów determinujących postawy wobec sztucznej inteligencji oraz określenie ich konsekwencji dla odpowiedzialnego i bezpiecznego wdrażania technologii w życie społeczne. Przeprowadzona analiza pozwala stwierdzić, że postawy wobec AI mają charakter głęboko ambiwalentny i są wynikiem współdziałania mechanizmów poznawczych, emocjonalnych oraz społeczno-kulturowych. W odniesieniu do problematyki szczegółowej rozdziału można wskazać, że zarówno zaufanie, jak i lęk wobec systemów autonomicznych mają swoje korzenie w tych samych procesach

psychologicznych – jednak prowadzą do odmiennych konsekwencji dla użytkownika oraz społeczeństwa.

Nawiązując do pierwszego problemu badawczego, dotyczącego psychologicznych mechanizmów antropomorfizacji i zaufania do AI, wyniki analizy wskazują, że uproszczone strategie poznawcze pełnią kluczową rolę w ocenie działania systemów autonomicznych. Heurystyki poznawcze sprawiają, że użytkownicy wyciągają wnioski na podstawie pozornych wskazówek, które sugerują kompetencję i obiektywność systemu, nawet gdy jego działanie nie jest w pełni zrozumiałe. Zaufanie do AI opiera się więc często na wrażeniu przewidywalności, skuteczności lub obiektywności, a nie na świadomości faktycznych możliwości technologii. Jak pokazały badania nad *algorithm aversion* i *algorithm appreciation*, reakcje użytkowników są skrajnie zmienne – od bezkrytycznego polegania na rekomendacjach algorytmu po nagłe odrzucenie technologii już po pierwszym błędzie. Zjawiska te ujawniają, że zaufanie nie jest konstruktem stabilnym ani prostym, lecz dynamicznym i silnie zależnym od wcześniejszych doświadczeń, kontekstu zadania oraz formy prezentacji danych.

Antropomorfizacja, która stanowi kolejny istotny element w omawianym obszarze, działa jako mechanizm ułatwiający interpretację zachowania AI poprzez ludzkie kategorie mentalne. Nadawanie technologii cech ludzkich, tj. intencji, emocji, kompetencji społecznych, może zwiększać jej akceptowalność, poczucie bliskości oraz przewidywalność. Jednocześnie nadmierna antropomorfizacja może prowadzić do fałszywego przekonania o podmiotowości moralnej AI oraz do „przesuwania" odpowiedzialności z ludzi na systemy techniczne. W skrajnych przypadkach nadmiernie „ludzka" forma robotów wywołuje efekt *uncanny valley*, który z kolei prowadzi do nieufności i dystansu emocjonalnego. Oznacza to, że antropomorfizacja działa dwutorowo, czyli może zarówno wspierać budowę zaufania, jak i je podważać, w zależności od stopnia podobieństwa do człowieka, cech formy oraz kontekstu społecznego.

Drugi problem badawczy dotyczył psychologicznych mechanizmów lęku przed AI. Analiza wskazuje, że źródłem obaw nie są jedynie funkcjonalne ograniczenia technologii, lecz głęboko zakorzenione mechanizmy emocjonalne. Szczególne znaczenie ma tu lęk egzystencjalny, opisany w ramach Teorii Opanowywania Trwogi, zgodnie z którą zagrożenie dla symbolicznych struktur wartości prowadzi do reakcji obronnych i negatywnych postaw wobec obiektów

podważających poczucie sensu i wyjątkowości człowieka. Sztuczna inteligencja, która jest zdolna wykonywać zadania intelektualne, przewidywać zachowania i podejmować decyzje, może być postrzegana jako zagrożenie dla ludzkiej autonomii oraz unikalności poznawczej. Dodatkowo lęk przed AI wzmacniają czynniki społeczne i kulturowe, tj. narracje medialne, wizje przyszłości pracy oraz wyobrażenia dotyczące utraty kontroli. W rezultacie lęk wobec AI nie jest wyłącznie reakcją na technologię, lecz na jej symboliczne znaczenie dla pozycji człowieka w świecie społecznym.

Odpowiadając na trzeci problem badawczy, dotyczący psychologicznych konsekwencji i implikacji praktycznych, można stwierdzić, że kluczowym wyzwaniem dla projektowania zaufanej AI jest właściwa kalibracja zaufania. Nadmierne zaufanie prowadzi do niebezpiecznych sytuacji, takich jak *automation bias* i nieumiejętność krytycznej oceny działań systemu, natomiast niedostateczne zaufanie może powodować odrzucanie technologii o wysokim potencjale społecznym. Oba te zjawiska mają realne konsekwencje dla bezpieczeństwa, efektywności i odpowiedzialności. Dlatego projektowanie systemów AI musi być oparte na transparentności, przewidywalności i jasnym komunikowaniu ograniczeń technologii. Szczególne znaczenie ma tu rozwój *Explainable AI* (XAI), który zwiększa zrozumienie działania algorytmów, a tym samym redukuje niepewność i obawy użytkowników.

W kontekście odpowiedzialnego wdrażania AI istotne jest również jednoznaczne określenie granic sprawczości technologii oraz przypisanie odpowiedzialności moralnej i prawnej ludziom, a nie maszynom. Psychologiczne procesy antropomorfizacji mogą prowadzić do błędnego przekonania, że technologie autonomiczne ponoszą odpowiedzialność za swoje decyzje, co wymaga przeciwdziałania poprzez odpowiednie projektowanie komunikatów i interfejsów. Równie istotne jest uwzględnienie różnic indywidualnych i kulturowych w poziomie umiejętności cyfrowych, ponieważ wpływają one na sposób, w jaki ludzie interpretują i oceniają działanie AI.

Podsumowując, analiza przeprowadzona w rozdziale wskazuje, że zaufanie i lęk wobec AI stanowią dwa komplementarne, lecz przeciwstawne rezultaty tych samych mechanizmów psychologicznych. Zrozumienie ich funkcjonowania jest kluczowe dla projektowania technologii, które wspierają autonomię człowieka, a jednocześnie nie prowadzą do nadmiernej zależności od systemów

algorytmicznych. Odpowiedzialne wdrażanie AI wymaga zatem połączenia wiedzy z zakresu psychologii, technologii oraz etyki, tak aby zapewnić nie tylko funkcjonalność technologii, lecz także jej akceptowalność społeczną. Tylko wówczas możliwe będzie budowanie dojrzałego, zrównoważonego zaufania do AI, które sprzyja bezpieczeństwu, dobrostanowi i rozwojowi społecznemu.

Literatura

1. Bryson, J. J. (2018). Patiency is not a virtue: The design of intelligent systems and systems of ethics. *Ethics and Information Technology, 20*(1), 15-26, https://doi.org/10.1007/s10676-018-9448-6
2. Christoforakos, L., Gallucci, A., Surmava-Große, T., Ullrich, D., Diefenbach, S. (2021). Can Robots Earn Our Trust the Same Way Humans Do? *A Systematic Exploration of Competence, Warmth, and Anthropomorphism as Determinants of Trust Development in HRI. Frontiers in Robotics and AI, 8*, 640444. https://doi.org/10.3389/frobt.2021.640444
3. Dietvorst, B. J., Simmons, J. P., Massey, C. (2015). Algorithm Aversion: People Erroneously Avoid Algorithms After Seeing Them Err. *Journal of Experimental Psychology: General, 144*(1), 114-131, https://doi.org/10.1037/xge0000033
4. Greenberg, J., Solomon, S., Pyszczynski, T. (1997). *Terror Management Theory of Self-Esteem and Cultural Worldviews: Empirical Assessments and Conceptual Refinements*. W: M. P. Zanna (red.), *Advances in Experimental Social Psychology*, vol. 29. Academic Press. https://doi.org/10.1016/S0065-2601(08)60016-7
5. Lee, J. D., See, K. A. (2004), Trust in automation: Designing for appropriate reliance. *Human Factors, 46*(1), 50-80, https://doi.org/10.1518/hfes.46.1.50_30392
6. Logg, J. M., Minson, J. A., Moore, D. A. (2019). Algorithm Appreciation: People Prefer Algorithmic to Human Judgment. *Organizational Behavior and Human Decision Processes, 151*, 90-103, https://doi.org/10.1016/j.obhdp.2018.12.005
7. Mori, M., MacDorman, K. F., Kageki, N. (2012). The Uncanny Valley. *IEEE Robotics & Automation Magazine, 19*(2), 98-100. https://doi.org/10.1109/MRA.2012.2192811
8. Tversky, A., Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science, 185*(4157), 1124-1131. https://doi.org/10.1126/science.185.4157.1124

9. Wang, W., Siau, K. (2019). Artificial Intelligence, Machine Learning, Automation, Robotics, Future of Work and Future of Humanity: A Review and Research Agenda. *Journal of Database Management, 30*(1), 61-79. https://doi.org/10.4018/JDM.2019010104
10. Waytz, A., Cacioppo, J., Epley, N. (2010). Who Sees Human? The Stability and Importance of Individual Differences in Anthropomorphism. *Perspectives on Psychological Science, 5*(3), 219-32, https://doi.org/10.1177/1745691610369336
11. West, S. M., Whittaker, M., Crawford, K. (2019). *Discriminating systems: Gender, race, and power in AI*. AI Now Institute Report. Pobrane z: https://ainowinstitute.org/publications/discriminating-systems-gender-race-and-power-in-ai-2 (dostęp: 02.09.2025).
12. Wojciszke, B. (2006), *Człowiek wśród ludzi. Zarys psychologii społecznej*. Scholar.
13. Złotowski, R., Proudfoot, D., Yogeeswaran, K., Bartneck, C. (2015). Anthropomorphism: Opportunities and Challenges in Human-Robot Interaction. *International Journal of Social Robotics*, *7*, 347-360. https://doi.org/10.1007/s12369-014-0267-6

Zofia Lubańska[1], Gabriel Tarasiuk[2]

Akademia Bialska im. Jana Pawła II, Biała Podlaska

https://doi.org/10.29316/9788368103304_10

AUTOMATYCZNE ROZPOZNAWANIE MOWY Z RUCHU UST PRZY UŻYCIU SZTUCZNEJ INTELIGENCJI – PRZEGLĄD METOD, ZASTOSOWAŃ I WYZWAŃ

AUTOMATIC SPEECH RECOGNITION FROM LIP MOVEMENTS USING ARTIFICIAL INTELLIGENCE – AN OVERVIEW OF METHODS, APPLICATIONS, AND CHALLENGES

Streszczenie

Rozpoznawanie mowy na podstawie ruchów warg (znane również jako czytanie z ruchu warg) stanowi kluczowy obszar na styku wizji komputerowej i przetwarzania mowy. Technologia ta umożliwia zrozumienie mowy poprzez analizę ruchów warg i języka mówiącego bez wykorzystania dźwięku. Jest ona szczególnie ważna dla osób z ubytkiem słuchu oraz w sytuacjach, w których sygnał audio jest niedostępny lub zakłócony przez hałas. W ostatnich latach poczyniono znaczne postępy w automatycznym odczytywaniu z ruchu warg dzięki rozwojowi sztucznej inteligencji, a zwłaszcza głębokich sieci neuronowych. Niniejszy rozdział zawiera przegląd aktualnych badań nad rozpoznawaniem

[1] e-mail: z.lubanska@dyd.akademiabialska.pl

[2] student, e-mail: tarasiuk22914@stud.akademiabialska.pl

mowy na podstawie ruchu warg z wykorzystaniem sztucznej inteligencji. Przedstawiono ewolucję podejść, od tradycyjnych metod opartych na cechach wizualnych i modelach statystycznych po nowoczesne architektury głębokiego uczenia (CNN, RNN, Transformer). Omówiono typowe metody i procesy rozpoznawania, a także popularne zbiory danych (takie jak GRID, LRW i LRS) oraz porównano wydajność wybranych modeli na tych zbiorach danych.

Słowa kluczowe: wizualne rozpoznawanie mowy, czytanie z ruchu warg, sieci neuronowe, głębokie uczenie, multimodalne rozpoznawanie mowy

Summary

Speech recognition based on lip movement (also known as lip reading) is a key area at the intersection of computer vision and speech processing. It enables understanding of speech by analysing the movements of the speaker's movements of the lips and other visible articulatory features without using sound. This technology is particularly important for individuals with hearing loss and in situations where the audio signal is unavailable or distorted by noise. In recent years, significant progress has been made in automatic lip reading, thanks to advances in artificial intelligence, particularly deep neural networks. This chapter provides an overview of current research on AI-based speech recognition from lip movements. The evolution of approaches is presented, from traditional methods based on visual features and statistical models to modern deep learning architectures (CNN, RNN, Transformer). Typical recognition methods and pipelines, as well as popular datasets (such as GRID, LRW, and LRS), are discussed, and the performance of selected models on these datasets is compared.

Keywords: visual speech recognition, lip reading, neural networks, deep learning, multimodal speech recognition

Wstęp

Komunikacja werbalna to złożony proces multimodalny, w którym poza dźwiękiem mowy istotną rolę odgrywają sygnały wizualne – ruchy ust, mimika twarzy, gesty itp. (Battista i in., 2025). Ludzie intuicyjnie wykorzystują informację wizualną podczas słuchania, zwłaszcza w warunkach hałasu. Zdolność czytania z ruchu ust (ang. *lip-reading*, nazywana też wizualnym rozpoznawaniem mowy) odgrywa ważną rolę w rozumieniu mowy przy zaburzeniach słuchu lub

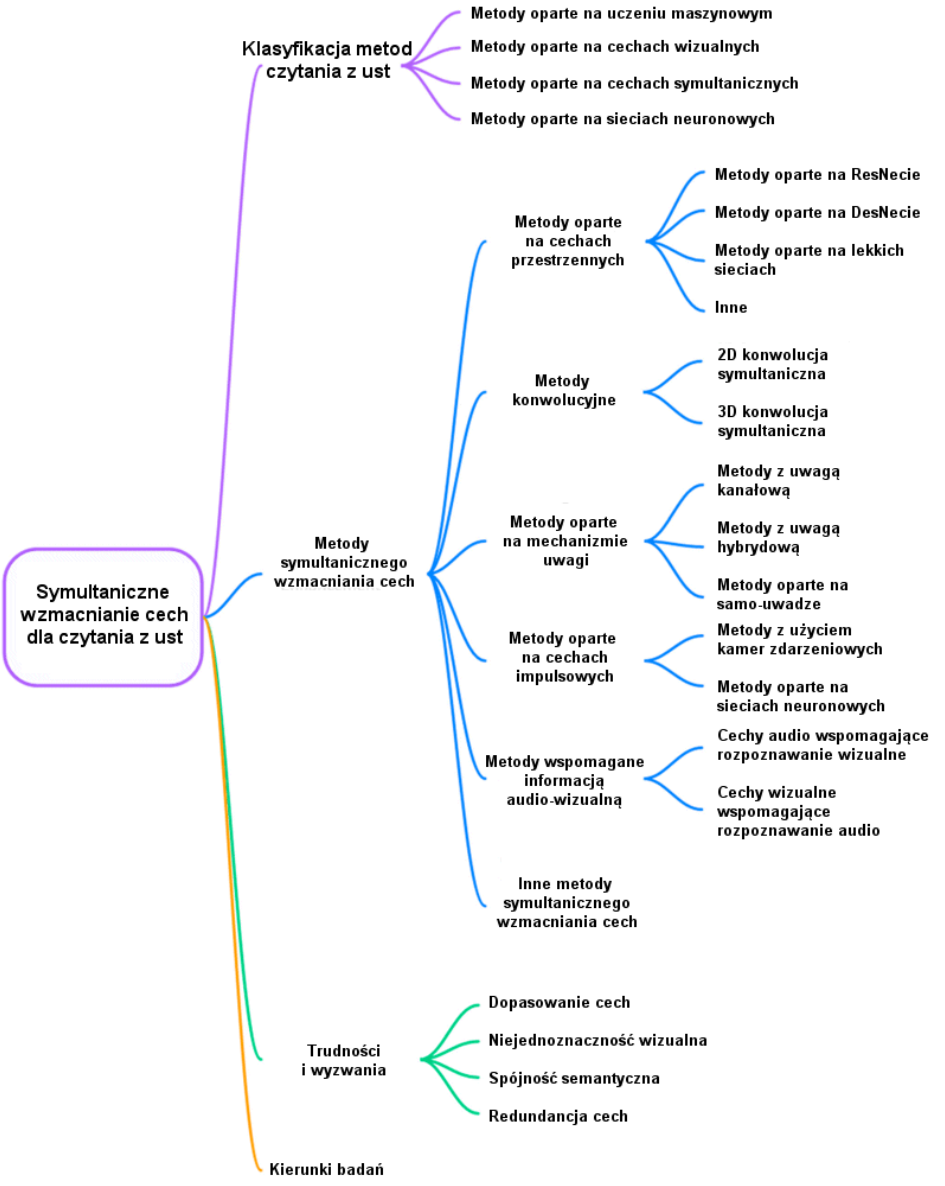
zakłóceniach akustycznych (Battista i in., 2025; Santana i in., 2025). U osób z ubytkiem słuchu umiejętność czytania z ust często staje się podstawowym kanałem komunikacji, znacząco poprawiając rozumienie przekazu (Santana i in., 2025). Również w codziennych sytuacjach, gdy np. rozmawiamy w hałaśliwym otoczeniu albo widzimy mówiącą osobę z daleka za szklaną przegrodą, analizujemy ruch ust, by wspomóc percepcję mowy (Battista i in., 2025). Nic więc dziwnego, że od dekad trwają prace nad zautomatyzowaniem tej umiejętności – tak, by komputery potrafiły "słyszeć oczami", czyli rozpoznawać mowę wyłącznie na podstawie obrazu twarzy mówiącego.

Automatyczne rozpoznawanie mowy z ruchu ust jest wyzwaniem interdyscyplinarnym, leżącym na pograniczu przetwarzania obrazów i rozpoznawania mowy (Potamianos i in., 2003; Rahmatullah i in., 2025). Tradycyjnie do tego zadania stosowano podejścia oparte na analizie cech wizualnych i klasycznych modelach uczenia maszynowego. Już na początku lat 90. badacze zademonstrowali pierwsze systemy rozpoznające proste wypowiedzi na podstawie nagrań wideo – przykładowo Mase i Pentland (1991) zastosowali analizę przepływu optycznego do śledzenia ruchu warg i rozpoznawania głosek. Pierwsze te systemy działały jednak w ograniczonym zakresie słów i wymagały ręcznej inżynierii cech. W kolejnych latach rozwijano metody ekstrakcji cech wizualnych z okolic ust, takie jak kształt konturu warg, sygnały kinematyczne czy transformacje obrazu (np. współczynniki DCT, DWT, modele kształtu) (Matthews i in., 2002; Lin i in., 2017; Miled i in., 2022). Cechy te zasilano klasyczne klasyfikatory sekwencyjne (ukryte modele Markowa HMM, modele k-NN, SVM itp.) do rozpoznawania ciągów głosek czy słów (Potamianos i in., 2003; Erbey, Barışçı, 2022). Mimo pewnych sukcesów (np. Matthews i wsp. (2002) osiągnęli ok. 63% trafności rozpoznawania cyfr na podstawie analizy kształtu ust), tradycyjne podejścia borykały się z wieloma ograniczeniami – małą odpornością na zmienność mówców, oświetlenia, kąt kamery oraz z fundamentalnym problemem niejednoznaczności wizualnej wielu fonemów (tzw. wizemy – różne dźwięki mowy mogą wyglądać tak samo na ustach) (Bear, Harvey, 2017; Rahmatullah i in., 2025). Przykładowo, głoski /p/, /b/, /m/ mają identyczny układ warg (jednostka wizualna), przez co system bazujący tylko na obrazie nie odróżni ich bez dodatkowego kontekstu. To tzw. zjawisko wizem, które utrudnia rozpoznawanie mowy wyłącznie z ruchu ust.

Dopiero rozwój głębokiego uczenia na przełomie ostatniej dekady przyniósł przełom w omawianej dziedzinie (Fenghour i in., 2021; Erbey, Barışçı, 2022; Rahmatullah i in., 2025). Pojawienie się dużych zbiorów danych wideo oraz technik głębokich sieci neuronowych (konwolucyjnych CNN i rekurencyjnych RNN/LSTM) pozwoliło budować modele automatycznie uczące się reprezentacji ruchu ust, znacznie przewyższające wcześniejsze rozwiązania (Fenghour i in., 2021; Wang i in., 2022). Na Rysunku 1 zebrano schematycznie technologie czytania z ust i ich rozwój (Po lewej wyróżniono tradycyjne podejścia *machine learning* (bazujące na ręcznie projektowanych cechach wizualnych i klasyfikatorach, np. HMM, SVM)). Po prawej pokazano podział metod głębokiego uczenia ze względu na typ przetwarzanych cech. W 2016 r. grupa z Oxford University zaprezentowała zestaw danych LRW obejmujący 500 słów wypowiadanych przez setki mówców w warunkach zbliżonych do rzeczywistych (materiały z programów BBC). Już dwa lata później najlepsze modele osiągnęły na nim ~83% skuteczności rozpoznawania słów, podczas gdy pierwsze próby dekady wcześniej dawały wyniki rzędu 50-60% (Fenghour i in., 2021). W kolejnych latach powstały jeszcze większe bazy, m.in. LRS2 i LRS3 z tysiącami pełnych zdań w języku angielskim oraz LRW-1000 – obejmujący 1000 słów w języku mandaryńskim (718 tysięcy próbek od >2000 mówców) (Wang i in., 2022; Rahmatullah i in., 2025). Dzięki temu możliwe stało się trenowanie end-to-end głębokich modeli, które automatycznie uczą się zarówno optymalnych cech wizualnych, jak i modelu języka (sekwencji). Rezultaty są imponujące – np. model LipNet (Assael i in., 2017) osiągnął 95% trafności rozpoznawania pełnych zdań w korpusie GRID, co przewyższyło wynik doświadczonych ludzkich "czytaczy z ust" (~52%). Również w zadaniach rozpoznawania pojedynczych słów w trudnych warunkach tzw. "in the wild" notuje się dziś dokładność przekraczającą 90% (Fenghour i in., 2021). Mimo to automatyczne czytanie z ust nadal pozostaje wyzwaniem. W warunkach rzeczywistych skuteczność spada – na dużym zestawie LRW-1000 (język chiński) najlepsze obecnie modele osiągają tylko ~45% poprawnych rozpoznań (Wang i in., 2022). Ponadto modele wymagają olbrzymich zbiorów danych treningowych i miewają problemy z generalizacją do nowych mówców czy języków (Rahmatullah i in., 2025).

Celem niniejszego rozdziału jest przedstawienie przeglądu stanu badań nad rozpoznawaniem mowy z ruchu ust z wykorzystaniem sztucznej inteligencji. Omówiono główne podejścia techniczne, począwszy od klasycznych metod

machine learning, aż po współczesne głębokie sieci neuronowe, wraz z charakterystyką architektur sieci oraz sposobami reprezentacji ruchu ust.



Rysunek 1. Taksonomia metod rozpoznawania mowy z ruchu ust z punktu widzenia rodzaju wykorzystywanych cech i architektur

Źródło: opracowanie własne na podstawie: Erbey, Barışçı, 2022; Fenghour i in., 2021.

W dalszej części zestawiono i porównano wybrane rozwiązania, przedstawiając dwie tabelaryczne syntezy: (a) porównanie najważniejszych modeli AI do czytania z ust (architektury i dokładności) oraz (b) porównanie właściwości najczęściej używanych zbiorów danych. Następnie omówiono praktyczne zastosowania technologii, bieżące ograniczenia i otwarte problemy oraz wskazano perspektywiczne kierunki dalszych badań.

Metody

Automatyczne rozpoznawanie mowy z ruchu ust można ogólnie podzielić na dwa etapy: (1) ekstrakcja cech wizualnych z sekwencji obrazów przedstawiających usta/twarz mówiącego, oraz (2) modelowanie sekwencji cech w celu zdekodowania treści mowy (np. klasyfikacja do słów lub transkrypcja zdania) (Potamianos i in., 2003; Fenghour i in., 2021). Historycznie najpierw dominowały podejścia szkieletowe (ang. handcrafted) – ręcznie projektowane cechy i klasyfikatory. Obecnie standardem są metody *end-to-end* z głębokimi sieciami neuronowymi, które uczą się reprezentacji i dekodowania jednocześnie. W niniejszym rozdziale przedstawiono najważniejsze techniki z obu tych paradygmatów.

Metody klasyczne (machine learning)

W początkowym okresie badań nad automatycznym czytaniem z ust (lata 90. i pocz. 2000) stosowano głównie podejścia oparte na klasycznym uczeniu maszynowym bez głębokich sieci (Potamianos i in., 2003). Kluczowym zadaniem była inżynieria dobrych cech wizualnych opisujących układ ust w kolejnych klatkach wideo. Wykorzystano szereg technik: cechy geometryczne (np. wysokość i szerokość otwarcia ust, kształt konturu warg), cechy opisujące teksturę w obszarze ust (np. współczynniki transformat – dyskretnej kosinusowej DCT lub falkowej DWT – obliczone z obrazu ust), cechy ruchu (wektory przepływu optycznego w regionie ust) oraz modele parametryczne (np. aktywne modele kształtu – AAM, śledzące kształt warg) (Matthews i in., 2002; Lin i in., 2017; Miled i in., 2022). Tabela 1 (pierwszy wiersz) zestawia typowe metody ekstrakcji cech stosowane w tych podejściach.

Następnie sekwencje takich wektorów cech z kolejnych klatek były klasyfikowane za pomocą sekwencyjnych modeli statystycznych. Najpopularniejsze były ukryte modele Markowa (HMM) – analogicznie jak w klasycznym rozpoznawaniu mowy audio – a także modele k-NN, GMM czy sieci neuronowe o płytkiej architekturze (Potamianos i in., 2003; Erbey, Barışçı, 2022). Przykładowo, znana praca Matthews i wsp. (2002) użyła modelu AAM do śledzenia ruchu ust i ekstrakcji parametrów kształtu oraz intensywności, a następnie zastosowała klasyfikator HMM do rozpoznawania głosek i słów – osiągając ~62% dokładności dla słów i ~79% dla prostych fraz na zbiorze MIRACL-VC1. Inna ówczesna metoda (Zhou i in., 2011, opisana m.in. Erbey, Barışçı, 2022) wykorzystała transformację Random Forest Manifold Alignment do dopasowania cech wizualnych różnych mówców przed klasyfikacją. Ogólnie jednak skuteczność klasycznych metod była umiarkowana, a każdy krok (wykrywanie ust, ekstrakcja cech, rozpoznawanie) wymagał osobnego dostrajania.

Dużym wyzwaniem okazała się wspomniana wcześniej niejednoznaczność fonetyczno-wizualna – wiele fonemów mapuje się na ten sam wizem (np. głoski {p, b, m} wyglądają identycznie). Rozwiązania próbowano szukać w definiowaniu odpowiednich mapowań fonem-wizem, które grupują dźwięki w klasy wizualne. Bear i Harvey (2017) dokonali obszernej analizy takich mapowań, pokazując ich wpływ na skuteczność rozpoznawania – nawet najlepsze mapowania wprowadzają jednak pewne straty informacji (Bear, Harvey, 2017). Podsumowując, tradycyjne podejścia *machine learning* stworzyły podwaliny automatycznego czytania z ust, ale miały liczne ograniczenia. Dopiero pojawienie się głębokich sieci neuronowych pozwoliło znacząco podnieść dokładność dzięki automatycznemu odkrywaniu złożonych cech czasowo-przestrzennych z danych.

Metody oparte na głębokim uczeniu

Obecnie dominującą rolę w rozpoznawaniu mowy z ruchu ust odgrywają metody wykorzystujące głębokie sieci neuronowe (Fenghour i in., 2021; Erbey, Barışçı, 2022; Deshpande i in., 2025). Cechuje je podejście *end-to-end*: surowe dane wizyjne (sekwencje obrazów) są przekazywane przez sieć, która uczy się zarówno reprezentacji cech, jak i modelowania ich zmian w czasie, a na wyjściu bezpośrednio generuje decyzję – rozpoznane słowo, zdanie albo transkrypcję

fonetyczną. W tej sekcji omawiamy główne typy architektur sieci stosowanych w literaturze, zgodnie z podziałem zaznaczonym na Rysunku 1.

Architektury CNN+RNN (cechy wizualne + cechy temporalne)

Pierwsze udane modele *deep learning* dla *lip-reading* łączyły dwie kluczowe składowe: konwolucyjną sieć neuronową (CNN) jako ekstraktor cech z obrazów oraz sieć rekurencyjną (RNN), najczęściej z komórkami długiej pamięci LSTM, do modelowania dynamiki sekwencji (Fenghour i in., 2021; Erbey, Barışçı, 2022; Lin i in., 2017). Taka architektura pozwalała uczyć się zarówno aspektów przestrzennych (układ ust), jak i czasowych (zmiany w czasie). Przykładem jest słynny model LipNet (Assael i in., 2017) – pierwsza w pełni *end-to-end* sieć do rozpoznawania całych zdań z ruchu ust. W LipNet zastosowano 3 warstwy CNN 2D (wyodrębniające cechy wizualne z kolejnych klatek) połączone z 2-warstwowym Bi-LSTM, a na wyjściu użyto mechanizmu CTC (*Connectionist Temporal Classification*) do generowania sekwencji znaków. Model ten, trenując się na dużym korpusie GRID (34 tys. zdań, prosty ograniczony słownik), osiągnął ~95% dokładności zdań, deklasując wcześniejsze metody (dla porównania – wynik doświadczonych ludzkich ekspertów to ~52%). Innym przykładem jest podejście Stafylakisa i Tzimiropoulosa (2017), które połączyło wydajną sieć ResNet-34 (do ekstrakcji cech z każdej klatki) z warstwami Bi-LSTM – osiągając wtedy rekordowe wyniki zarówno dla pojedynczych słów (97% na zbiorze GRID), jak i ciągłych ujęć *in the wild* (ponad 76% rozpoznawalności słów na 500-słownym zbiorze LRW). Architektury CNN+RNN stały się bazą wielu kolejnych prac; stosowano różne warianty: głębsze sieci konwolucyjne (np. VGG, Inception) jako *front-end* oraz sieci LSTM lub GRU (jednokierunkowe i dwukierunkowe) jako *back-end* (Fenghour i in., 2021; Erbey, Barışçı, 2022). Niektóre systemy zamiast przewidywać ciągi znaków dokonywały klasyfikacji do ustalonego słownika słów (przydatne np. do rozpoznawania komend). Ogólnie podejście CNN+RNN zdominowało dekadę 2015-2020 w dziedzinie *visual speech recognition*, przynosząc ogromny skok jakości.

Architektury 3D-CNN (cechy spatiotemporalne)

Alternatywą dla kombinacji 2D CNN + RNN stały się modele wykorzystujące konwolucje trójwymiarowe (3D CNN) lub ogólniej *spatiotemporal convolution*. Są to sieci, które uczą się filtrów konwolucyjnych działających jednocześnie w dwóch wymiarach przestrzennych (obrazu) i w wymiarze czasu – co pozwala od razu wydobywać lokalne cechy dynamiczne (ruch) bez osobnej warstwy rekurencyjnej (Ma, Sun, 2025; Wang i in., 2022). Przykładem jest model zaproponowany przez Zhanga i współautorów (2019, omówiony w Wang et al., 2022), w którym wykorzystano splot 3D do analizowania sekwencji obrazów ust, a następnie wyjście 3D-CNN sekwencyjnie przetwarzano już klasyczną siecią 2D-CNN (tzw. podwójna konwolucja: najpierw na rozdzielczości czasowej, potem przestrzennej). Modele z konwolucją czasową często cechują się większą równoległością przetwarzania niż RNN, przez co są łatwiejsze do trenowania na GPU. W literaturze opisano wiele wariantów – od prostych *Temporal Convolutional Networks (TCN)* po bardziej złożone architektury. Często stosuje się też połączenie: najpierw warstwy 3D-CNN, a potem klasyfikator sekwencyjny (np. TCN + CTC). Spatiotemporal CNN okazały się skuteczne zwłaszcza dla rozpoznawania pojedynczych słów i krótkich fraz. Przykładowo, Wang i in. (2022) zaproponowali model *Vision Transformer* z wbudowanym 3D-CNN, który na zbiorze LRW osiągnął ~89,5% dokładności – najlepszy wynik bez użycia dodatkowych danych. Zaletą 3D-CNN jest zdolność do jednoczesnego uchwycenia kształtu ust i jego zmian między sąsiadującymi klatkami (np. ruchu szczęki, pojawiania się zębów/przełyku itp.). Wyzwanie stanowi jednak większa złożoność obliczeniowa – filtry 3D mają więcej parametrów, co utrudnia trenowanie bardzo głębokich modeli na ograniczonych danych (tzw. ryzyko nadmiernego dopasowania). Dlatego często stosuje się wieloskładowe podejścia hybrydowe – np. łączenie bloków ResNet (uczących cech przestrzennych) z wstawkami 3D (dla dynamiki) lub stosowanie konwolucji separowalnych w czasie (np. model *Time-Depth Separable CNN*).

Mechanizmy uwagi i transformatory

Najnowsze lata przyniosły adaptację do zadania czytania z ust modeli opartych na mechanizmie *attention* oraz Transformerów, które zrewolucjonizowały

przetwarzanie języka naturalnego. Modele te rezygnują z recyrencji na rzecz mechanizmu *self-attention*, który potrafi globalnie modelować zależności w sekwencji. Już około 2019 roku pojawiły się pierwsze prace, w których moduły uwagi poprawiały dokładność rozpoznawania mowy wizualnej – np. poprzez selektywne skupianie się na kluczowych ramach (momentach artykulacji) lub ważnych obszarach obrazu (np. okolice ust kontra inne części twarzy) (Fenghour i in., 2021; Ma i Sun, 2025). W 2020-2021 zaczęto stosować pełne architektury Transformer, często łączone z CNN. Przykładowo, model *Vision Transformer + CTC* opisany przez Martineza i in. (omówiony w Wang i in., 2022) osiągnął bardzo dobre wyniki w rozpoznawaniu ciągłych zdań, zastępując LSTM blokiem Transformer. Zaletą uwagi jest umiejętność modelowania długoterminowych zależności – ważne np. dla uchwycenia kontekstu zdaniowego i uwzględnienia informacji językowej (składni, częstości słów). Nowoczesne systemy często wykorzystują tzw. architekturę *encoder-decoder Transformer*, analogiczną jak w tłumaczeniu maszynowym: enkoder (np. typu Conformer czy Swin Transformer) przetwarza sekwencję obrazów ust na wewnętrzną reprezentację, a dekoder Transformer generuje sekwencję wyjściową (np. liter lub słów) uwzględniając kontekst językowy. Renomowane grupy (np. z Baidu, DeepMind) zaprezentowały modele sekwencyjne z uwagą, które dla długich zdań potrafią zbliżyć się do skuteczności rozpoznawania audio. Należy wspomnieć, że mechanizm uwagi bywa też używany w połączeniu z CNN/RNN – np. moduły CBAM (*Convolutional Block Attention Module*) integrujące mapy uwagi kanałowej i przestrzennej z wyjściami CNN poprawiły wyniki w chińskim systemie *lipreadingu* Fu i Ni (2023). Ogółem, architektury z uwagą są obecnie jednym z głównych kierunków badań – umożliwiają bowiem bardziej efektywne wykorzystanie informacji o kontekście wypowiedzi niż klasyczne LSTM, choć wymagają bardzo dużych zbiorów danych do treningu ze względu na znaczną liczbę parametrów (Ma, Sun, 2025; Deshpande i in., 2025).

Inne podejścia głębokie

W literaturze znaleźć można także mniej konwencjonalne techniki. Jedna z nich to wykorzystanie sieci neuronowych (SNN) i przetwarzania impulsowego. Ponieważ ruch ust ma charakter sygnału czasowego, próbowano modelować go za pomocą sieci neuronowych przetwarzających dyskretne zdarzenia

(tzw. *spiking networks*). Zhang i wsp. (2021, przywoływani przez Ma, Sun, 2025) zaproponowali sieć kolczystą do rozpoznawania sylab, jednak jej skuteczność była niższa niż klasycznych podejść CNN/RNN. Ciekawą ścieżką jest natomiast użycie kamer zdarzeniowych (*event cameras*) do rejestrowania ruchu ust. Kamery te zapisują zmiany jasności w pikselach z wysoką częstotliwością (mikrosekundy) jako asynchroniczny strumień zdarzeń. Wstępne prace (Ma, Sun, 2025) sugerują, że takie sensory mogą uchwycić szybkie ruchy artykulacyjne niewidoczne w klasycznych nagraniach wideo, potencjalnie ułatwiając rozpoznawanie. Wymaga to jednak nowych algorytmów – często łączy się to z sieciami kolczystymi lub innymi architekturami dostosowanymi do danych zdarzeniowych.

Podsumowując, pod wpływem technik głębokiego uczenia nastąpiła znacząca zmiana paradygmatu w automatycznym czytaniu z ust. Tabela 1 zestawia cztery główne klasy podejść, omawiane powyżej, wraz z charakterystyką wykorzystywanych przez nie cech i modeli. W kolejnej części rozdziału przyjrzymy się, jakie konkretne wyniki udało się osiągnąć dzięki tym metodom oraz na jakich danych były one testowane.

Wyniki przeglądu badań

W celu syntetycznego porównania dotychczasowych osiągnięć badawczych zebrano informacje o wybranych modelach AI do rozpoznawania mowy z ruchu ust oraz o najczęściej wykorzystywanych zbiorach danych. Tabela 1 przedstawia porównanie reprezentatywnych podejść technicznych i ich skuteczności, zaś Tabela 2 podsumowuje właściwości kluczowych korpusów danych używanych w tej dziedzinie (Fenghour i in., 2021; Rahmatullah i in., 2025; Deshpande i in., 2025).

Jak wynika z Tabeli 1, na ograniczonych zadaniach (np. pojedyncze słowa z ustalonego słownika, zdania o prostej składni) współczesne modele AI osiągają bardzo wysokie wyniki – często powyżej 90%, a w niektórych przypadkach dorównują lub przewyższają umiejętności ludzkie (Fenghour i in., 2021). Dla przykładu, LipNet czy późniejsze modele ResNet+LSTM prawie bezbłędnie odczytują zdania z korpusu GRID, w którym składnia i słownictwo są mocno ograniczone. Jednak w miarę przechodzenia do trudniejszych, realnych danych (dłuższe wypowiedzi, dowolne słowa, różni mówcy i warunki) dokładność spada – na dużych korpusach typu LRS czy LRW-1000 najlepsze dotąd modele osiągają 40-50%

rozpoznawania (Rahmatullah i in., 2025; Wang i in., 2022), co pozostawia jeszcze duży margines do poprawy.

Tabela 1. Przegląd wybranych modeli do automatycznego czytania z ust z wykorzystaniem głębokiego uczenia. Zestawiono charakterystyki architektury oraz uzyskane wyniki na standardowych zbiorach danych

Model (rok)	**Główne cechy architektury**	**Zbiór danych (język)**	**Dokładność / Wynik**
LipNet – Assael i in. (2017)	end-to-end CNN + BiLSTM, CTC (rozpoznawanie zdań)	GRID (ang.; proste zdania)	95,2% dokładność zdań (WER ~5%) – przewyższa wynik ludzki
ResNet+LSTM – Stafylakis, Tzimiropoulos (2017)	Głęboki ResNet34 (ekstrakcja cech) + BiLSTM	GRID / LRW (ang.; słowa)	97,0% słowa na GRID; 76,2% słowa na LRW (500 słów)
Deep AVSR – Chung i in. (2017)	CNN + LSTM z połączeniem audiowizualnym (*Watch, Listen, Attend & Spell*)	LRS2 (ang.; zdania)	~50% wskaźnik BLEU (ok. 50% zdań poprawnie) – system audiowizualny (audio+video)
TMseq2seq – Afouras i in. (2018)	Transformer Encoder + Decoder seq2seq (uwaga wielogłowa)	LRS3 (ang.; zdania TED)	~41% WER (błąd słowa) w trybie wizualnym; ok. 2× lepszy niż wcześniejsze LSTM
Shillingford, i in (2019)	Masowy model *seq2seq* (DFCNN + LSTM, 2 tys. godz. danych)	własny zbiór (ang.; zdania)	40,9% WER dla zdań (setki godzin YouTube) – pierwszy model skalujący się na duże dane
Vision Transformer – Ma i in. (2022)	3DCNN + 2D SwinTransformer, klasyfikacja słów	LRW (ang.; 500 słów)	89,5% dokładność (Top1) – stan na 2022 r., bez danych dodatkowych (Wang i in., 2022)
EfficientGhostNet + GRU – Fu, Ni (2023)	Lekka sieć CNN (GhostNet zmodyfikowany) + BiGRU	własny (chiń.; 50 słów)	~85% dokładność słów – model zoptymalizowany pod urządzenia mobilne (Fu i in., 2023)

Model (rok)	Główne cechy architektury	Zbiór danych (język)	Dokładność / Wynik
SwinTransformer 3D + PTM – Zhang i in. (2023)	Model hybrydowy: SwinTransformer 3D + augmentacja *PartitionTime Masking*	LRW (ang.) / LRW 1000 (mandar.)	94,1% słowa na LRW; 45% słowa na LRW 1000 – najlepsze wyniki ok. 2023 r. (Wang i in., 2022)

Źródło: opracowanie własne na podstawie: Fenghour i in., 2021; Wang i in., 2022; Rahmatullah i in., 2025.

W Tabeli 2 zestawiono natomiast właściwości najpopularniejszych zbiorów danych wykorzystywanych w badaniach nad automatycznym czytaniem z ust. Są to zarówno korpusy kontrolowane (nagrania studyjne z ograniczonym słownikiem), jak i zbiory typu "*in the wild*" (klipy z telewizji, Internetu, o dużej różnorodności). Wielkość i złożoność tych zestawów danych w ogromnym stopniu determinują postęp – jak podkreślają Rahmatullah i in. (2025), brak dużych i zróżnicowanych danych przez lata hamował rozwój skutecznych modeli *lip-readingu*. Dopiero relatywnie niedawno – około 2016 r. – zaczęły powstawać naprawdę duże korpusy, co natychmiast przełożyło się na skok wyników (Fenghour i in., 2021; Rahmatullah i in., 2025).

Tabela 2. Przegląd wybranych publicznych zbiorów danych do wizualnego rozpoznawania mowy (lip-reading). Zestawiono charakterystykę korpusów: język, rodzaj i zakres wypowiedzi, liczebność i przykładowe osiągi najlepszych modeli (stan ok. 2023 r.)

Zbiór danych (rok)	Język / zakres mowy	Zawartość korpusu	Przykładowa skuteczność
GRID (2006)	angielski (proste zdania o ustalonej formie)	34 mówców, 1000 zdań każdy (ok. 33,5 tys. nagrań). Słownik ~51 słów, zdania 6-wyrazowe o stałej strukturze. Nagrania studyjne, frontalne wideo 25 Hz.	95% dokładność zdań (LipNet 2017). Ludzki poziom: ok. 50% (Assael i in., 2017).

Zbiór danych (rok)	Język / zakres mowy	Zawartość korpusu	Przykładowa skuteczność
CUAVE (2002)	angielski (ciągi cyfr)	36 mówców (19 par). Nagrania cyfr 0-9 (ciągi, izolowane). Ponadto nagrania w ruchu i z zakłóceniami. Łącznie ok. 7 tys. klipów.	~80-85% dla cyfr (modele z lat 2000-2010); rzadziej używany we współczesnych pracach.
AVLetters (1998) / AVLetters2 (2015)	angielski (alfabet)	AVLetters: 10 mówców, każda litera A–Z powtórzona 7 razy (łącznie 1820 klipów). AVLetters2: 5 widoków kątowych × 20 mówców × 26 liter (2600 klipów).	>90% dokładność liter w najnowszych modelach (Fenghour i in., 2021).
OuluVS (2009) / OuluVS2 (2015)	angielski (krótkie frazy)	OuluVS: 20 mówców, 10 fraz (np. “Excuse me”) – łącznie 200 nagrań. OuluVS2: 53 mówców, nagrania 5 słów z 5 kątów (0°, 30°, 45°, 60°, 90°) – ~2650 klipów.	>90% dokładność słów (stan ok. 2018); małe zbiory, głównie do testów *multi-view*.
LRW – Lip Reading in the Wild (2016)	angielski (słowa, *in the wild*)	~500 klas słów (brytyjski angielski, różne tematy). Nagrania z BBC. Ponad 1000 wystąpień na klasę, łącznie >500k klipów, długość ~1,16 s każdy. Duża zmienność mówców, ujęć, oświetlenia.	83% dokładność Top 1 (2018); ok. 94% (SOTA, 2023) (Wang et al.., 2022).
LRS2 – Lip Reading Sentences (2017)	angielski (zdania z TV)	Część *pre-train*: ~96 tys. zdań. Część *train-val*: 45 839 zdań (do 5 s) z 144 godzin programów BBC. Test: 1243 zdań. Słownik >40k słów.	~50% WER (DeepMind, 2017), spadek do ~32% WER (2021, Transformer; Rahmatullah i in., 2025).

Zbiór danych (rok)	Język / zakres mowy	Zawartość korpusu	Przykładowa skuteczność
LRS3 (TED) (2018)	angielski (zdania z TED)	Nagrania z TED/TEDx; ~118 516 segmentów zdań (do 30 s, średnio 4 s) z 409 wystąpień. ~475 godzin wideo.	~59% WER (2018, *seq2seq*) → ~33% WER (2022, ConformerTransformer; Rahmatullah i in., 2025).
LRW1000 (2019)	mandaryński (słowa, *in the wild*)	1000 klas słów (różne sylaby; każde słowo 1-2 znaki). 718 018 próbek wideo od >2000 osób (programy telewizyjne).	Tylko 43-45% dokładność Top 1 najlepszych modeli (2021-2023) – znacznie trudniejsza niż LRW (Wang i in., 2022).

Źródło: opracowanie własne.

Tabela 2 ilustruje rosnącą skalę i złożoność nowszych zbiorów danych. Wczesne korpusy (AVLetters, OuluVS) liczyły kilkaset nagrań i ograniczały się do pojedynczych słów czy liter, podczas gdy współczesne (LRW, LRS, LRW-1000) to setki tysięcy próbek reprezentujących realne warunki. Ten skok ilościowy był krytyczny – jak zauważają Rahmatullah i wsp. (2025), dopiero po 2014 r. pojawiło się >100 prac rocznie, gdyż badacze zyskali materiał do trenowania głębokich modeli. Mimo to nadal brakuje pełnej różnorodności – większość dużych zbiorów dotyczy języka angielskiego, co rodzi pytanie o przenoszalność metod na inne języki (np. tonalne, alfabet sylabiczny itp.) (Rahmatullah i in., 2025). Z tabeli widać też, że wyniki dla różnych korpusów bardzo się różnią. Najłatwiejsze (GRID – zamknięty kontekst) osiągnęły poziom niemal perfekcyjny, podczas gdy najbardziej wymagające (LRW-1000, spontaniczne zdania) pozostają dalekie od doskonałości. Ta rozbieżność jest istotna przy rozważaniu zastosowań praktycznych, co omówiono w następnym rozdziale.

Dyskusja

Zastosowania i znaczenie technologii

Automatyczne rozpoznawanie mowy z ruchu ust ma szereg potencjalnych zastosowań o dużym znaczeniu społecznym. Jednym z najważniejszych jest wsparcie komunikacyjne dla osób niesłyszących i niedosłyszących. Choć

implanty ślimakowe i aparaty słuchowe poprawiają percepcję dźwięku, wiele osób z ubytkiem słuchu i tak polega na czytaniu z ust, zwłaszcza w hałasie lub gdy rozmówca nie mówi wyraźnie (Battista i in., 2025; Santana i in., 2025). Systemy automatycznego czytania z ust mogłyby w przyszłości działać jako "tłumacz", zamieniając ruch ust rozmówcy na napisy w czasie rzeczywistym – analogicznie do napisów w filmach, ale generowanych na żywo. Badania Santany i in. (2025) pokazały, że czytanie z ust jest efektywną strategią poprawy rozumienia mowy przez osoby z ubytkiem słuchu, zwłaszcza w połączeniu z innymi metodami (np. aparatami słuchowymi). Automatyzacja tej strategii za pomocą AI może zatem znacząco zwiększyć dostępność informacji i jakość życia takich osób. Już teraz powstają prototypy aplikacji mobilnych, które potrafią rozpoznawać krótkie wypowiedzi z kamery smartfona i wyświetlać tekst – choć na razie ich skuteczność bywa ograniczona do prostych zdań i pojedynczych mówców (Erbey, Barışçı, 2022).

Drugim istotnym obszarem jest rozpoznawanie mowy w warunkach zakłóconego lub niedostępnego dźwięku. W wielu sytuacjach mikrofony zawodzą – np. z powodu głośnego szumu (przemysł, koncerty) albo gdy źródło dźwięku jest daleko. Człowiek do pewnego stopnia radzi sobie, patrząc na twarz mówiącego. Podobnie systemy audio-wizualnego rozpoznawania mowy (AV-ASR) integrujące sygnał dźwiękowy i obraz warg potrafią osiągnąć lepsze wyniki niż klasyczne ASR oparte wyłącznie na audio (Potamianos i in., 2003; Rahmatullah i in., 2025). Na przykład, gdy w nagraniu ktoś mówi w hałasie ulicznym – model audio-wizualny może wciąż poprawnie zidentyfikować słowa dzięki sygnałom wizualnym, podczas gdy model audio sam może się mylić. Ephrat i in. (2017, cyt. za Rahmatullah i in., 2025) pokazali system, który z niemego wideo potrafił częściowo odtworzyć przybliżony sygnał mowy (*Vid2Speech*), a Afouras i in. (2018) – demonstrację generowania mówiącej twarzy z samego audio (problem odwrotny, ale blisko spokrewniony). To wszystko wskazuje na komplementarność modalności audio i wideo. W praktyce przyszłe systemy rozpoznawania mowy w smartfonach czy robotach mogą korzystać z kamery do "wspomagania" rozpoznawania w trudnych warunkach akustycznych.

Innym polem jest bezpieczeństwo i nadzór. Czytanie z ust na nagraniach z monitoringu mogłoby pomóc zrozumieć, co mówią osoby obserwowane, nawet jeśli nie słychać dźwięku. W filmach sensacyjnych często pojawia się scena, gdzie

służby specjalne "czytają z ust" podejrzanych na odległość – AI mogłaby kiedyś to umożliwić na większą skalę. Rzecz jasna, wiąże się to z wątpliwościami natury prywatności i etyki (potencjalna masowa inwigilacja poprzez kamery CCTV), które należałoby rozważyć (Battista i in., 2025).

Kolejna kategoria to interfejsy człowiek-komputer w warunkach ciszy. Wyobraźmy sobie asystenta głosowego, z którego chcemy skorzystać w cichym biurze czy bibliotece – nie możemy głośno mówić, ale możemy szeptać lub poruszać ustami bezgłośnie. System *lip-reading* mógłby rozpoznać naszą wypowiedź i wykonać polecenie, nie emitując hałasu. Już teraz rozwijane są algorytmy rozpoznające tzw. *silent speech*, czyli mowę bezgłośną – czasem z pomocą specjalnych czujników (EMG mięśni twarzy, kamera termowizyjna itp.), a czasem tylko kamery. Przykładowo, praca Kumar i in. (2020, cyt. za Deshpande i in., 2025) wprowadziła system "Hush-Hush Speak" rekonstruujący mowę szeptaną z wideo za pomocą modelu generatywnego. Takie technologie mogą znaleźć zastosowanie w dyskretnym sterowaniu urządzeniami (np. wydawaniu poleceń smartfonowi bez mówienia na głos).

Warto też wspomnieć o wykorzystaniu czytania z ust w połączeniu z innymi modalnościami w systemach wielomodalnej komunikacji. Przykładowo, w rzeczywistości wirtualnej (VR) animowanie awatara postaci ludzkiej wymaga synchronizacji ruchu ust z mową, co można wspomóc algorytmami *lip-syncu*. W medycynie analiza wzorców ruchu ust może wesprzeć diagnozę zaburzeń mowy lub zaburzeń neurologicznych. Wreszcie technologia ta ma walor naukowy: pomaga badać proces integracji słuchowo-wizualnej (efekt McGurka – iluzja słuchowa pod wpływem bodźca wzrokowego) i lepiej rozumieć, jak ludzki mózg "łączy" obraz i dźwięk w percepcji mowy (Battista i in., 2025; Potamianos i in., 2003).

Ograniczenia obecnych metod

Pomimo imponujących postępów opisanych w rozdziałach poprzednich, istnieje wiele wyzwań i ograniczeń, z którymi mierzą się obecne systemy rozpoznawania mowy z ruchu ust. Wskazują je zarówno rezultaty licznych eksperymentów, jak i kompleksowe przeglądy (Rahmatullah i in., 2025; Santana i in., 2025; Deshpande i in., 2025). Poniżej omówiono kluczowe problemy.

Niedobór danych i generalizacja.

Trening głębokich sieci wymaga ogromnych zbiorów danych, a te – choć coraz większe – wciąż mogą być niewystarczające dla pełnej generalizacji. Większość obecnych modeli jest trenowana i testowana na względnie ograniczonych domenach (np. prezenterzy telewizyjni, wyraźna artykulacja). Gdy system zostanie użyty poza takim domenowym zbiorem – np. do rozpoznawania mowy przypadkowej osoby na ulicy – skuteczność może drastycznie spaść. Jest to problem niezależności od mówcy i warunków. Ludzie uczą się czytać z ust latami, oglądając wiele osób – analogicznie sieci potrzebują bardzo zróżnicowanych danych, by nie "przyuczyć się" do konkretnych twarzy, kształtów warg czy stylów mówienia. Jak zauważają Rahmatullah i in. (2025), nadal istnieją poważne wyzwania, takie jak niewystarczająca ilość danych, nieadekwatne warunki środowiskowe i różnorodność językowa, które ograniczają uniwersalność modeli. Przykładowo, model wytrenowany na brytyjskim angielskim może słabo radzić sobie z odczytywaniem amerykańskiego angielskiego (inne akcenty, ruchy ust), a tym bardziej z innym językiem – bo różnice fonetyczne i artykulacyjne są znaczne. Zatem brak wielojęzycznych, wielodialektalnych zbiorów to poważne ograniczenie.

Wizemy i kontekst językowy.

Jak już omawiano, wiele fonemów jest wizualnie nierozróżnialnych. Oznacza to, że system odczytujący mowę z ust ma fundamentalną niejednoznaczność, którą musi rozwiązać inaczej – poprzez kontekst językowy. Człowiek radzi sobie z wizemami dzięki temu, że wie, jakie słowa w danym języku są możliwe, więc może zgadnąć na podstawie części informacji. System AI musi czynić podobnie, integrując model języka (np. poprzez warstwę dekodera albo użycie statystyk n-gramów). Niemniej, w obecnych modelach – zwłaszcza pracujących na poziomie słów – czasem brakuje głębszego rozumienia kontekstu. Model może popełniać błędy semantycznie niedorzeczne, np. odczytać zdanie jako pozbawione sensu, bo wybrał wizualnie pasujące słowo, które jednak nie pasuje do reszty zdania. Problem wizemów szczególnie utrudnia rozpoznawanie krótkich wypowiedzi i pojedynczych słów bez kontekstu – np. rozróżnienie "mat" vs "bat" tylko z obrazu ust jest prawie niemożliwe (Bear, Harvey, 2017; Rahmatullah i in., 2025).

Warunki oświetleniowe, pozycja kamery i jakość wideo.

Większość badań wykorzystuje dość dobrej jakościowo dane – frontalne ujęcia twarzy, przyzwoita rozdzielczość (≥ 100 px szerokość ust), dobre oświetlenie. W praktyce kamery monitoringu czy smartfony nie zawsze dają taki obraz. Silne cienie na twarzy, słabe światło, duży kąt (profil zamiast *en face*), niska rozdzielczość – to wszystko obniża skuteczność czytania z ust. Badania wskazują, że przy rozdzielczości ust < 20 px systemy praktycznie tracą zdolność rozpoznawania (Deshpande i in., 2025). Również opóźnienia w klatkażu (np. kamera 5 fps) sprawiają, że ruch wydaje się skokowy, co utrudnia analizę. Niektóre prace starają się radzić z tym poprzez *superresolution* lub poprawę jakości obrazu jako etap wstępny, ale to wciąż aktywna dziedzina (Ma, Sun, 2025).

Złożoność obliczeniowa i opóźnienia.

Głębokie modele *lip-reading* bywają bardzo duże (np. transformatory z setkami milionów parametrów) i wymagają potężnych GPU do działania w czasie rzeczywistym. To ogranicza ich zastosowanie np. w urządzeniach mobilnych. Powoli pojawiają się próby optymalizacji – np. Fu i Ni (2023) zaproponowali lekką sieć Efficient GhostNet, redukując liczbę parametrów bez dużej utraty dokładności. Mimo to osiągnięcie płynnego działania, np. na smartfonie z samą kamerą (bez wysyłania danych do chmury), nadal jest wyzwaniem. Trzeba też pamiętać o opóźnieniu predykcji: przy rozpoznawaniu zdań modele często potrzebują "obejrzeć" całą sekwencję, zanim wydadzą wynik (zwłaszcza te z dekoderem *seq2seq*). To oznacza, że przy długiej wypowiedzi system zadziała dopiero po jej zakończeniu. Dla interaktywnych zastosowań lepsze są modele *streamingowe* lub działające w oknach czasowych. CTC oferuje tu przewagę, bo pozwala częściowo *online* dekodować, ale większość badań nadal koncentruje się na trybie *offline*.

Bezpieczeństwo i etyka.

Zastosowania nadzorcze tej technologii niosą pytania etyczne: czy wolno maszynowo odczytywać z ust ludzi bez ich wiedzy? To budzi obawy o prywatność – ktoś może być de facto "podsłuchiwany", nawet jeśli nie rozmawia przez telefon, a jedynie rozmawia twarzą w twarz, ale w zasięgu kamery. Już teraz

technologie typu *lip-reading CCTV* budzą kontrowersje (Battista i in., 2025). Z technicznego punktu widzenia ograniczeniem (a z etycznego może zaletą) jest tu to, że systemy te wciąż są zawodne w warunkach rzeczywistych – więc raczej nie są masowo stosowane. Jednak przy dalszym postępie prawdopodobnie pojawi się konieczność regulacji prawnych dotyczących wykorzystania tej technologii.

Kierunki dalszych badań

Biorąc pod uwagę wymienione ograniczenia, społeczność naukowa wyznaczyła szereg kierunków badań, które mają na celu udoskonalenie automatycznego rozpoznawania mowy z ruchu ust (Fenghour i in., 2021; Rahmatullah i in., 2025; Deshpande i in., 2025).

Zwiększanie i urozmaicanie danych treningowych

Należy tworzyć większe i bardziej zróżnicowane zestawy danych – obejmujące różne języki, dialekty, grupy wiekowe, warunki nagrania. Szczególnie paląca jest potrzeba korpusów dla języków innych niż angielski (np. chiński LRW-1000 jest krokiem w tę stronę, ale przydałby się np. duży zbiór dla języków z innym systemem fonetycznym – arabskiego, hindi, hiszpańskiego, polskiego). Ponadto dane w warunkach naturalnych – nagrania z różnych kątów, w ruchu – pomogą modelom lepiej generalizować. Jedną z dróg jest generowanie danych syntetycznych – np. użycie modeli generatywnych do tworzenia nowych nagrań ust mówiących zadane frazy (speech-to-video), co może służyć augmentacji danych dla rzadkich słów czy nietypowych warunków (Deshpande i in., 2025).

Uczenie niewspierane i transfer wiedzy

Aby zmniejszyć zależność od etykietowanych danych (transkrypcji), intensywnie bada się metody *self-supervised learning*. Modele mogłyby uczyć się reprezentacji ruchu ust bez transkrypcji, np. predykując przyszłe klatki, dopasowując zdania do klipów (*contrastive learning*) itp. (Ma, Sun, 2025). Taki wstępnie wytrenowany model można potem dostroić na mniejszej liczbie danych oznaczonych. Innym podejściem jest *transfer learning* z pokrewnych zadań: np. model do

rozpoznawania mowy audio lub do rozpoznawania mimiki twarzy może dostarczyć część reprezentacji dla modelu *lip-reading* (Fenghour i in., 2021).

Integracja z modelem językowym

W miarę postępów w NLP, rozsądnym kierunkiem jest ściślejsze powiązanie systemów czytania z ust z nowoczesnymi modelami językowymi. Obecne architektury *seq2seq* już to robią w pewnym stopniu, ale można pójść dalej: np. użycie dużego modelu językowego do autokorekty wyjścia systemu *lip-reading*. Model mógłby wygenerować listę najlepszych hipotez, a LLM ocenić, która jest najbardziej sensowna w kontekście (Deshpande i in., 2025). Integracja na poziomie semantyki mogłaby częściowo rozwiązać problemy związane z wizemami i błędami gramatycznymi.

Multimodalność – systemy audiowizualne

Badania pokazują, że modele uczone jednocześnie na audio i wideo często uczą się lepszych reprezentacji – nawet jeśli potem audio nie jest dostępne (Potamianos i in., 2003; Rahmatullah i in., 2025). Dlatego warto kontynuować rozwój modeli multimodalnych. W szczególności technika *teacher-student*: model audio (łatwiejszy do trenowania, bo ASR audio jest dojrzały) naucza model wizualny, przekazując mu wskazówki (*knowledge distillation*). Integracja audio-wideo jest celem samym w sobie – przyszłe systemy rozpoznawania mowy w realnych zastosowaniach powinny wykorzystywać wszystko, co dostępne (mikrofony + kamery), by zminimalizować błędy. Można też rozszerzyć modalność: np. łączyć czytanie z ust z rozpoznawaniem mowy migowej (ruch rąk) dla komunikacji z osobami głuchymi.

Nowe architektury sieciowe

Choć transformatory dominują, eksplorowane są nowe architektury mogące lepiej pasować do sekwencji wizyjnych – np. wspomniane sieci kolczyste (*spiking neural nets*) potencjalnie bardzo wydajne energetycznie, działające w czasie rzeczywistym na sprzęcie neuromorficznym (Ma, Sun, 2025). Jeśli uda się opracować SNN skutecznie rozpoznającą wzorce ruchu ust, mogłoby to otworzyć drogę

do miniaturowych układów (np. implantów rozumiejących mowę z mikrokamer). Innym trendem są modele z atencją czasowo-kanoniczną, potrafiące radzić sobie z różnymi tempami mowy.

Lepsze zrozumienie i interpretowalność

Większość modeli to tzw. czarne skrzynki – wiemy, że działają (lub nie), ale trudno powiedzieć, dlaczego np. pomyliły "mat" z "bat". Przyszłe badania mogą skupić się na uczynieniu modeli bardziej interpretowalnymi. Np. wizualizacje map uwagi pozwalają zobaczyć, na które klatki czy obszary model patrzył, podejmując decyzję (Deshpande i in., 2025). W kontekście prawnym (np. wykorzystania w sądzie odczytu maszynowego z monitoringu) potrzebna będzie ocena wiarygodności – system musiałby sygnalizować, na ile pewne jest rozpoznanie. Kalibracja takich miar zaufania dla mowy wizualnej jest osobnym wyzwaniem.

Podsumowując, perspektywy rozwoju automatycznego czytania z ust są bardzo obiecujące. Jak zauważają Rahmatullah i in. (2025), postęp w ostatniej dekadzie był ogromny – nowe techniki głębokiego uczenia pozwoliły osiągnąć przekonujące wyniki *state-of-the-art*, choć wciąż pozostaje wiele do zrobienia, aby zrównać wydajność systemów wizualnych z audio i pokonać takie przeszkody, jak niedobór danych czy niejednoznaczność wizualna. Trend wskazuje, że integracja modalności, większe modele i lepsze algorytmy uczenia będą sukcesywnie zmniejszać tę lukę.

Wnioski

Rozpoznawanie mowy na podstawie ruchu ust i warg jest zagadnieniem o długoletniej historii badawczej, które dzięki połączeniu z nowoczesnymi technikami sztucznej inteligencji wkroczyło w nową fazę rozwoju. W niniejszym przeglądzie dokonano podsumowania stanu badań w tym zakresie. Wprowadzenie nakreśliło znaczenie problemu – zarówno dla osób z ubytkiem słuchu, jak i w kontekście wzbogacania klasycznego rozpoznawania mowy w hałasie czy budowy multimodalnych interfejsów. Przypomniano też wyzwania, takie jak niejednoznaczność wizualna głosek (problem wizemów), które utrudniają to zadanie.

W części metod prześledzono ewolucję technik: od wczesnych podejść opartych na cechach geometrycznych i HMM, poprzez pojawienie się sieci neuronowych CNN+RNN zdolnych do automatycznej ekstrakcji cech z obrazów ust, aż po współczesne architektury Transformer i hybrydowe modele *spatiotemporalne*. Na Rysunku 1 zobrazowano taksonomię tych metod, uwzględniając różne kategorie cech (wizualne vs. czasowo-przestrzenne), mechanizmy uwagi oraz nowatorskie podejścia (sieci kolczyste, kamery zdarzeniowe). Wyniki przeglądu zestawione w tabelach pokazują, że na ograniczonych zadaniach (np. zamknięty słownik słów) systemy AI osiągają już znakomite rezultaty (rzędu 90-95% poprawności), natomiast w warunkach zbliżonych do rzeczywistych wciąż istnieje spora luka do zapełnienia – dokładności rzędu 40-50% wskazują, że jest miejsce na poprawę (Rahmatullah i in., 2025; Wang i in., 2022). Analiza najczęściej używanych korpusów danych (Tabela 2) ujawnia dramatyczny wzrost skali dostępnych danych w ostatnich latach, co było kluczowe dla sukcesu metod głębokiego uczenia.

W dyskusji podkreślono praktyczne zastosowania technologii czytania z ust (wspieranie niesłyszących, rozpoznawanie mowy w hałasie, nadzór i bezpieczeństwo, interfejsy cichej mowy), ale także bariery, które utrudniają ich pełną realizację. Obecne ograniczenia – takie jak niedostatek danych spoza języka angielskiego, wrażliwość na warunki nagrania, duża złożoność obliczeniowa modeli czy kwestie prywatności – wymagają dalszych badań. Zaproponowano szereg kierunków przyszłego rozwoju: od gromadzenia nowych wielojęzycznych zbiorów danych, przez uczenie samonadzorowane i transfer wiedzy z modeli audio, po udoskonalenia architektur (np. integracja z modelami językowymi, zastosowanie sieci Transformer również po stronie ekstrakcji cech, rozwój modeli multimodalnych audio-wizualnych). Zwrócono również uwagę na aspekt etyczny i potrzebę interpretowalności – tak, by rosnące możliwości technologiczne szły w parze z zaufaniem do systemów AI.

Reasumując, rozpoznawanie mowy z ruchu ust przeszło drogę od ciekawostki naukowej do poważnej dziedziny badań nad sztuczną inteligencją. Wykorzystanie głębokich sieci neuronowych przyniosło ogromny skok jakościowy – do poziomu, na którym systemy te mogą realnie konkurować z ludzkimi umiejętnościami w ograniczonych warunkach (Fenghour i in., 2021). Jednak dopiero rozwiązanie otwartych problemów (wizemy, generalizacja, dane) pozwoli na

powszechne wdrożenia. Wysiłki badawcze trwają: jak podaje najnowszy systematyczny przegląd literatury, w latach 2014-2024 opublikowano ponad 100 oryginalnych prac w tej tematyce, a tempo badań rośnie (Rahmatullah i in., 2025). Można oczekiwać, że w kolejnych latach będziemy świadkami dalszego zbliżania wydajności systemów czytania z ust do rozpoznawania mowy z dźwięku – być może aż do punktu, gdy obie te modalności połączą się, dając w pełni multimodalne systemy rozumienia mowy działające niezawodnie w różnych warunkach.

Literatura

1. Battista, M., Collesei, F., Orzan, E., Fantoni, M., Bottari, D. (2025). LipReading: Advances and Unresolved Questions in a Key Communication Skill. *Audiology Research, 15*(4), 89. https://doi.org/10.3390/audiolres15040089
2. Bear, H. L., Harvey, R. (2017). Phoneme-to-viseme mappings: the good, the bad, and the ugly. *Speech Communication*, 95, 40-67. https://doi.org/10.1016/j.specom.2017.07.001
3. Deshpande, S., Shirsath, K., Pashte, A., Loya, P., Shingade, S., Sambhe, V. (2025). A Comprehensive Survey of Advancement in Lip Reading Models: Techniques and Future Directions. *IET Image Processing, 19*(1). https://doi.org/10.1049/ipr2.70095
4. Erbey, A., Barışçı, N. (2022). A Survey on LipReading with Deep Learning. *Uluslararası Mühendislik Araştırma ve Geliştirme Dergisi (International Journal of Engineering Research and Development), 14*(2), 844-860. https://doi.org/10.29137/umagd.1038899
5. Fenghour, S., Chen, D., Guo, K., Li, B., Xiao, P. (2021). Deep LearningBased Automated LipReading: A Survey. *IEEE Access, 9*, 121184-121205. https://doi.org/10.1109/ACCESS.2021.3107946
6. Fu, Y., Lu, Y., Ni, R. (2023). Chinese LipReading Research Based on ShuffleNet and CBAM. *Applied Sciences, 13*(2), 1106. https://doi.org/ 10.3390/app13021106
7. Lin, B. S., Yao, Y. H., Liu, C. F., Lien, C. F., Lin, B. S. (2017). Development of Novel LipReading Recognition Algorithm. *IEEE Access, 5*, 794-801. https://doi.org/10.1109/ACCESS.2017.2649838
8. Ma, Y., Sun, X. (2025). Spatiotemporal Feature Enhancement for LipReading: A Survey. *Applied Sciences, 15*(8), 4142. https://doi.org/10.3390/app15084142

9. Mase, K. and Pentland, A. (1991), Automatic lipreading by optical-flow analysis. Syst. Comp. Jpn., 22: 67-76. https://doi.org/10.1002/scj.4690220607
10. Matthews, I., Cootes, T. F., Bangham, J. A., Cox, S., Harvey, R. (2002). Extraction of visual features for lipreading. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 24*(2), 198-213. https://doi.org/10.1109/34.982900
11. Miled, M., Messaoud, M. A. B., Bouzid, A. (2023). Lip reading of words with lip segmentation and deep learning. Multimedia Tools and Applications, *82*, 551-571. https://doi.org/10.1007/s11042-022-13321-0
12. Potamianos, G., Neti, C., Gravier, G., Garg, A., Senior, A. (2003). Recent advances in the automatic recognition of audiovisual speech. *Proceedings of the IEEE, 91*(9), 1306-1326. https://doi.org/10.1109/JPROC.2003.817150
13. Rahmatullah, G. M., Ruan, S.J., Wisesa, I. W. W., Li, L. P.H. (2025). Enhancing visual speech perception through deep automatic lipreading: A systematic review. *Computers in Biology and Medicine, 190*, 110019. https://doi.org/10.1016/j.compbiomed.2025.110019
14. Santana, R. S., Menezes, E. C., Ralin, V., Givigi, S. (2025). Lipreading as a communication strategy to enhance speech recognition in individuals with hearing impairment: A scoping review. *Disability and Rehabilitation: Assistive Technology, 20*(5), 1235-1246. https://doi.org/10.1080/17483107.2025.2449984
15. Wang, H., Pu, G., Chen, T. (2022). A lip-reading method based on a 3D convolutional vision transformer. *IEEE Access, 10*, 77205-77212. https://doi.org/10.1109/ACCESS.2022.3193231

Piotr Lichograj[1], Marta Chodyka[2], Michał Szałkowski[3], Vladimir Golovko[4]

Akademia Bialska im. Jana Pawła II, Biała Podlaska

https://doi.org/10.29316/9788368103304_11

ANALIZA METOD YOLO W DETEKCJI OBIEKTÓW Z OBRAZÓW REJESTROWANYCH PRZEZ UAV ORAZ ICH ZASTOSOWANIE W POMIARZE PRĘDKOŚCI PIESZYCH

ANALYSIS OF YOLO METHODS IN DETECTING OBJECTS FROM IMAGES RECORDED BY UAVS AND THEIR APPLICATION IN MEASURING PEDESTRIAN SPEED

Streszczenie

W ostatnich latach bezzałogowe statki powietrzne (UAV) stały się kluczowym narzędziem pozyskiwania danych wizyjnych do analiz ruchu w środowisku miejskim. Perspektywa lotnicza umożliwia szerokie zastosowania w monitoringu pieszych, jednak jednocześnie generuje liczne wyzwania związane z niewielkim rozmiarem obiektów, zmienną

[1] https://orcid.org/0000-0002-1351-4874, e-mail: p.lichograj@dyd.akademiabialska.pl
[2] https://orcid.org/0000-0002-8819-2451, e-mail: m.chodyka@dyd.akademiabialska.pl
[3] student, e-mail: szalkowski22913@stud.akademiabialska.pl
[4] https://orcid.org/0000-0003-2615-289X, e-mail: v.golovko@dyd.akademiabialska.pl

wysokością lotu oraz niestabilnością detekcji. W rozdziale przedstawiono kompleksowy przegląd architektur YOLO w kontekście ich wykorzystania w obrazach z UAV oraz omówiono metody estymacji prędkości obiektów na materiale wideo. W części badawczej zaprezentowano prototypowy system detekcji i estymacji prędkości pieszych z wykorzystaniem modelu YOLOv8, wytrenowanego na specjalnie przygotowanym zbiorze danych UAV. Oceniono skuteczność detekcji, stabilność śledzenia oraz wiarygodność pomiaru prędkości. Wyniki potwierdzają, że nowsze generacje modeli YOLO, po odpowiednim dostrojeniu do danych lotniczych, umożliwiają dokładną analizę ruchu pieszych, a temat ten pozostaje obszarem o dużym potencjale badawczym.

Słowa kluczowe: UAV, detekcja obiektów, YOLOv8, analiza ruchu pieszych, estymacja prędkości, głębokie uczenie, śledzenie obiektów

Abstract

In recent years, unmanned aerial vehicles (UAVs) have become essential tools for acquiring visual data to support motion analysis in urban environments. The aerial perspective enables effective pedestrian monitoring but poses significant challenges, including small object size, variable flight altitude, and detection instability. This chapter provides a comprehensive review of YOLO architectures and their applicability to UAV imagery, along with an overview of velocity estimation methods in video sequences. The experimental section introduces a prototype system for pedestrian detection and speed estimation using a YOLOv8 model trained on a custom UAV dataset. The system's detection accuracy, tracking stability, and speed estimation reliability were evaluated. The findings demonstrate that modern YOLO models, when properly fine-tuned on UAV-specific data, enable accurate pedestrian motion analysis, while the topic itself remains an underexplored research area with substantial potential for future studies.

Keywords: UAV, object detection, YOLOv8, pedestrian motion analysis, speed estimation, deep learning, object tracking

Wstęp

W ostatniej dekadzie bezzałogowe statki powietrzne (UAV, ang. Unmanned Aerial Vehicles) stały się jednym z kluczowych narzędzi pozyskiwania danych wizyjnych w zastosowaniach naukowych, komercyjnych oraz inżynieryjnych. W przeciwieństwie do klasycznych, stacjonarnych systemów monitoringu, UAV

oferują możliwość dynamicznego rejestrowania obrazów z różnych wysokości i perspektyw, przy jednoczesnym zachowaniu stosunkowo niskich kosztów eksploatacji. Dzięki swojej mobilności oraz elastyczności platformy te znajdują zastosowanie w licznych scenariuszach, takich jak monitoring ruchu drogowego, analiza zachowań pieszych, badania dynamiki tłumu w przestrzeni miejskiej, kontrola infrastruktury krytycznej czy nadzór bezpieczeństwa podczas wydarzeń masowych. W każdej z tych dziedzin kluczową rolę odgrywa szybkie i wiarygodne przetwarzanie wizji komputerowej, które umożliwia automatyzację procesów decyzyjnych i zwiększa efektywność analiz.

Rozwój metod głębokiego uczenia, a w szczególności konwolucyjnych sieci neuronowych (CNN), wywarł fundamentalny wpływ na postęp w obszarze detekcji obiektów. Nowoczesne architektury, takie jak R-CNN, SSD czy rodzina modeli YOLO (You Only Look Once), umożliwiły znaczące podniesienie jakości i szybkości detekcji, otwierając drogę do zastosowań wymagających działania w czasie rzeczywistym. Spośród nich to właśnie YOLO stało się jednym z najważniejszych rozwiązań, oferujących kompromis pomiędzy wysoką dokładnością a bardzo dużą szybkością inferencji. Cecha ta jest szczególnie istotna w systemach UAV, w których przetwarzanie obrazu odbywa się często w warunkach ograniczonej mocy obliczeniowej, a jednocześnie wymaga natychmiastowej reakcji – na przykład podczas monitoringu bezpieczeństwa, sterowania autonomicznego czy wspomagania decyzji operatora.

Pomimo licznych zalet, detekcja obiektów na nagraniach UAV pozostaje zadaniem trudnym. Wynika to m.in. z niewielkiej liczby pikseli reprezentujących obiekty, zwłaszcza pieszych i małych pojazdów, z częstych zmian wysokości i kąta rejestrowania obrazu oraz z występowania dynamicznych zakłóceń wynikających z warunków środowiskowych. W literaturze wielokrotnie zwracano uwagę, że modele trenowane na standardowych zbiorach danych – takich jak MS COCO – charakteryzujących się głównie perspektywą naziemną, nie osiągają satysfakcjonujących wyników w warunkach UAV bez dodatkowego dostrajania. W odpowiedzi na te ograniczenia powstają liczne rozwiązania dedykowane obrazom lotniczym, obejmujące zarówno modyfikacje architektury, jak i adaptację procesu trenowania, rozszerzanie zbiorów danych oraz integrację informacji kontekstowych.

Detekcja obiektów stanowi jednak dopiero pierwszy element bardziej złożonego procesu, jakim jest analiza ruchu. W sytuacjach takich jak monitorowanie pojazdów czy pieszych konieczne jest nie tylko zidentyfikowanie obiektów, ale również oszacowanie ich prędkości oraz kierunku ruchu. Tradycyjne podejścia wykorzystują algorytmy śledzenia – takie jak KLT, SORT czy Deep SORT – działające w kombinacji z detektorami obiektów. Obliczanie prędkości opiera się najczęściej na translacji obiektu między kolejnymi klatkami, przeliczonej na jednostki metryczne w oparciu o parametry kalibracyjne kamery lub znane odcinki referencyjne. W przypadku UAV dochodzą do tego dodatkowe trudności, takie jak niestabilność kadru, zmieniająca się wysokość lotu czy brak jednoznacznych punktów odniesienia na powierzchni terenu.

W literaturze istnieje wiele prac dotyczących pomiaru prędkości pojazdów, jednak badania dotyczące estymacji prędkości pieszych w rejestrach UAV pozostają stosunkowo nieliczne. Wynika to głównie z trudności percepcyjnych: piesi widziani z wysokości kilkudziesięciu metrów zajmują bardzo mało pikseli, poruszają się w sposób nieregularny, a ich detekcja wymaga modeli odpornych na zmiany warunków wizualnych oraz częste zasłonięcia obiektów. Tym samym stanowi to obszar o dużym potencjale badawczym, szczególnie w kontekście rozwoju inteligentnych systemów monitoringu miejskiego oraz zastosowania UAV w badaniach przestrzennych.

W niniejszym rozdziale przedstawiono szeroki przegląd rozwiązań z rodziny YOLO w kontekście detekcji obiektów z UAV oraz przeprowadzono analizę istniejących podejść do estymacji prędkości obiektów na podstawie materiału wideo. W dalszej części zaprezentowano również badania własne autora, obejmujące opracowanie systemu detekcji I estymacji prędkości pieszych z wykorzystaniem modelu YOLOv8 wytrenowanego na dedykowanym zbiorze danych UAV. Szczególny nacisk położono na stabilność detekcji, proces śledzenia obiektów oraz wpływ filtracji i wygładzania sygnału na uzyskiwane wyniki, a także na praktyczne problemy związane z kalibracją przestrzenną. Całość pracy uzupełnia analiza porównawcza YOLOv4, YOLOv5 oraz YOLOv8 w kontekście danych UAV oraz omówienie ograniczeń i możliwości dalszego rozwoju metod.

Przegląd literatury

Rozwój detektorów obiektów i znaczenie architektur YOLO

Postęp w dziedzinie detekcji obiektów w ostatnich latach jest nierozerwalnie związany z rozwojem architektur głębokiego uczenia, w szczególności konwolucyjnych sieci neuronowych (CNN). Wprowadzenie modeli takich jak R-CNN, Fast R-CNN, Faster R-CNN czy SSD znacząco podniosło poziom dokładności detekcji, jednak ich wydajność obliczeniowa często nie pozwalała na zastosowanie w warunkach wymagających pracy w czasie rzeczywistym. Przełom przyniosła rodzina modeli YOLO (You Only Look Once), która wprowadziła jednoprzejściową (single-shot) koncepcję detekcji, umożliwiając przetwarzanie obrazu w sposób bezpośredni i szybki, bez konieczności wcześniejszego generowania propozycji regionów.

Klasyczne modele YOLO opierają się na pojedynczej sieci konwolucyjnej, która przetwarza obraz wejściowy i jednocześnie przewiduje współrzędne ramek ograniczających oraz prawdopodobieństwa klas. Architektura ta charakteryzuje się wyjątkową szybkością, co czyni ją szczególnie przydatną w zadaniach czasu rzeczywistego i zastosowaniach mobilnych, w tym na platformach UAV. Wprowadzenie YOLO stanowiło zatem istotny krok w kierunku praktycznego wykorzystania detekcji w systemach o ograniczonej mocy obliczeniowej.

YOLOv4: optymalizacja dokładności i wydajności

YOLOv4 (Bochkovskiy i in., 2020) odegrał kluczową rolę w rozwoju rodziny algorytmów YOLO (*You Only Look Once* – „widzisz tylko raz") poprzez wprowadzenie szeregu istotnych innowacji metodologicznych, które zostały podzielone na dwa zbiory rozwiązań. Wśród najważniejszych elementów architektury i procesu uczenia znalazł się ekstraktor cech CSPDarknet53 (*Cross Stage Partial Darknet* – częściowe łączenie etapów sieci Darknet), który poprawia przepływ gradientów oraz redukuje złożoność obliczeniową modelu. Istotną rolę odegrały również zaawansowane techniki augmentacji danych (*data augmentation* – sztuczne zwiększanie zbioru danych), w szczególności metoda mozaikowa, polegająca na łączeniu kilku obrazów w jeden, co zwiększa różnorodność danych treningowych i poprawia generalizację modelu. Kolejną innowacją było

zastosowanie samoadwersarialnego uczenia (SAT – ang. *Self-Adversarial Training*), które polega na celowym modyfikowaniu obrazów wejściowych w trakcie uczenia w celu zwiększenia odporności sieci na zakłócenia. Ponadto w YOLOv4 wprowadzono nowe funkcje straty (ang. loss functions) oparte na uogólnionych miarach współczynnika części wspólnej do sumy pól (IoU - ang. *Intersection over Union*), które umożliwiają dokładniejsze i stabilniejsze dopasowanie ramek ograniczających podczas procesu detekcji obiektów.

YOLOv4 osiągnął bardzo wysokie wyniki na zbiorze MS COCO, a jednocześnie utrzymał znaczną szybkość inferencji na popularnych GPU, co znacznie przyczyniło się do jego rozpowszechnienia w środowisku akademickim i inżynieryjnym. Model ten stał się w krótkim czasie standardem referencyjnym w wielu pracach badawczych dotyczących UAV, w których konieczne jest zachowanie kompromisu pomiędzy dokładnością a wydajnością.

YOLOv5: praktyczne wdrożenia i elastyczność architektoniczna

YOLOv5, rozwijany przez Ultralytics, mimo braku początkowej publikacji naukowej, bardzo szybko zyskał popularność dzięki implementacji w PyTorch, łatwości użytkowania oraz bogatym narzędziom do trenowania, wizualizacji i ewaluacji modeli. Rodzina YOLOv5 obejmuje szereg wariantów – od bardzo lekkich modeli „nano" po duże wersje „x-large" – co umożliwia ich dobór do konkretnych ograniczeń sprzętowych.

W literaturze znajdują się liczne prace, w których YOLOv5 zastosowano m.in. w detekcji obiektów naziemnych, obrazowaniu medycznym, rolnictwie precyzyjnym, monitoringu infrastruktury czy analizie zachowań ludzi. Jednakże w zastosowaniach UAV model ten często wymaga dostrajania, ponieważ dane lotnicze istotnie różnią się od konwencjonalnych zbiorów treningowych.

YOLOv8: najnowsza generacja i detekcja anchor-free

YOLOv8 (Ultralytics, 2020) stanowi obecnie najbardziej zaawansowaną, stabilną i powszechnie wykorzystywaną generację detektorów z rodziny YOLO (*You Only Look Once* – „widzisz tylko raz"), znajdując szerokie zastosowanie zarówno w badaniach naukowych, jak i w rozwiązaniach przemysłowych. Jedną

z kluczowych zmian wprowadzonych w tej wersji jest przejście z podejścia opartego na kotwicach (ang. *anchor-based*), w którym detekcja obiektów bazuje na wcześniej zdefiniowanych ramkach odniesienia, na podejście bez użycia kotwic (ang. *anchor-free*), co upraszcza proces uczenia oraz zwiększa elastyczność modelu w wykrywaniu obiektów o zróżnicowanych rozmiarach. Istotną innowacją jest również zastosowanie rozdzielonych gałęzi detekcyjnych, w których proces regresji ramek ograniczających oraz klasyfikacji obiektów realizowany jest niezależnie, co przekłada się na lepszą stabilność uczenia i wyższą dokładność predykcji. Ponadto YOLOv8 charakteryzuje się zredukowaną złożonością obliczeniową, co umożliwia jego efektywne wykorzystanie w systemach czasu rzeczywistego oraz na urządzeniach o ograniczonych zasobach sprzętowych. Dodatkowym atutem tej generacji jest integracja wielu zadań wizji komputerowej w ramach jednej platformy, obejmującej detekcję obiektów, segmentację oraz klasyfikację, co znacząco zwiększa uniwersalność i praktyczną użyteczność algorytmu.

YOLOv8 wyznacza obecnie standard w czasie rzeczywistym, zarówno w badaniach akademickich, jak i w rozwiązaniach wdrażanych w sektorze przemysłowym. Ultralytics podkreśla, że model ten jest zoptymalizowany pod kątem efektywności, dzięki czemu może działać nie tylko na wydajnych GPU, ale również – w mniejszym zakresie – na CPU.

Specyfika detekcji w obrazach UAV

Dane pozyskane z platform UAV stawiają przed systemami detekcji liczne wyzwania, wynikające m.in. z:

- niewielkich rozmiarów obiektów (piechota, małe UAV, drobne pojazdy),
- zmiennej wysokości lotu i kąta nachylenia kamery,
- dynamicznego tła,
- zakłóceń związanych z warunkami środowiskowymi,
- częstych częściowych zasłonięć (ang. occlusions).

Literatura wskazuje, że bezpośrednie zastosowanie standardowych modeli detekcyjnych do obrazów UAV prowadzi do spadku skuteczności, dlatego konieczne jest dostrajanie modeli do specyfiki danych lotniczych oraz wprowadzanie modyfikacji architektonicznych (Zeng i in., 2025; Fan i in., 2025).

Modyfikacje YOLO dedykowane UAV

W literaturze zaprezentowano szereg ulepszeń architektury YOLO mających na celu poprawę jakości detekcji obiektów w danych lotniczych:

- C4D-YOLOv8 (Zeng i in., 2025) – ulepszony ekstraktor cech i integracja wielorozdzielczych reprezentacji, poprawiające detekcję małych obiektów w zbiorach VisDrone.
- LUD-YOLO (Fan i in., 2025) – model zoptymalizowany pod kątem ograniczonych zasobów obliczeniowych UAV.
- Badania Alhassana i Yılmaza (2025) wskazują, że YOLOv5 może osiągać lepsze wyniki niż YOLOv4 w scenariuszach UAV, choć zależy to od danych i konfiguracji hiperparametrów.
- Jeszcze szerszą analizę stanowią prace Kaur i in. (2025), które porównują siedem wersji YOLO w zadaniu air-to-air, pokazując, że skuteczność modeli jest silnie zależna od zbioru danych i scenariusza, a nowsza wersja nie daje automatycznie przewagi.

Estymacja prędkości obiektów

Estymacja prędkości pojazdów i pieszych z materiału wideo stanowi ważny element systemów analizy ruchu. Klasyczne podejścia opierają się na:

- detekcji obiektu,
- śledzeniu jego pozycji w kolejnych klatkach,
- konwersji przesunięć pikselowych na jednostki metryczne,
- filtracji i stabilizacji wyników.

Najczęściej wykorzystywane algorytmy śledzenia to KLT, SORT i Deep SORT. W systemach opartych na YOLO procedura jest analogiczna: środek ramki ograniczającej stanowi reprezentację pozycji obiektu, a przemieszczenia pomiędzy klatkami – podstawę estymacji prędkości (Ultralytics, 2024a).

Znaczną trudność stanowi stabilność detekcji – nawet pojedyncze błędne wykrycie może prowadzić do artefaktów w obliczanych prędkościach. Dlatego literatura zaleca stosowanie metod wygładzania czasowego (np. średniej ruchomej), filtrów progowych oraz dokładnej kalibracji przestrzennej.

W odniesieniu do badań pieszych są nieliczne, co wynika z trudności percepcyjnych związanych z niewielkimi rozmiarami obiektów w ujęciach UAV.

Metodologia badań

Celem badań własnych było opracowanie oraz przetestowanie systemu detekcji i estymacji prędkości pieszych na nagraniu wideo zarejestrowanym z UAV z wykorzystaniem modelu YOLOv8. Najważniejszym założeniem było stworzenie kompletnego rozwiązania, które umożliwia nie tylko wykrywanie obiektów na nagraniu, lecz także obliczanie ich prędkości w sposób stabilny, powtarzalny i zbieżny z faktycznym ruchem pieszych. Kluczowe znaczenie miało zapewnienie odpowiedniej jakości detekcji obiektów niewielkich rozmiarów, obserwowanych z wysokości ok. 30 metrów, co – jak wskazuje literatura – stanowi jedno z największych wyzwań w analizie danych UAV.

Projekt obejmował przygotowanie dedykowanego zbioru danych, trenowanie modelu YOLOv8, opracowanie modułu do przetwarzania materiału wideo oraz wykonanie procedury śledzenia i estymacji prędkości obiektów. Dodatkowo przeprowadzono analizę stabilności wyników oraz porównanie YOLOv8 z modelami YOLOv4 i YOLOv5.

W pierwszym etapie badań przygotowano zbiór danych obejmujący 200 obrazów pochodzących bezpośrednio z nagrań UAV, przedstawiających pieszych obserwowanych z perspektywy pionowej. Obrazy zostały ręcznie oznaczone w środowisku Roboflow. Zdefiniowano jedną klasę detekcji – person. Jednokomórkowa klasyfikacja była decyzją celową, gdyż upraszczała zadanie oraz zwiększała dokładność rozpoznawania obiektu kluczowego dla późniejszych obliczeń prędkości.

Zbiór danych został następnie podzielony na część treningową i walidacyjną zgodnie z rekomendacjami Ultralytics. Model YOLOv8 poddano procesowi trenowania z typowymi hiperparametrami, a uzyskany najlepszy model zapisano jako plik best.pt. W trakcie procesu trenowania zastosowano standardowe augmentacje, wspomagające generalizację modelu na różnorodne warunki oświetleniowe i geometryczne.

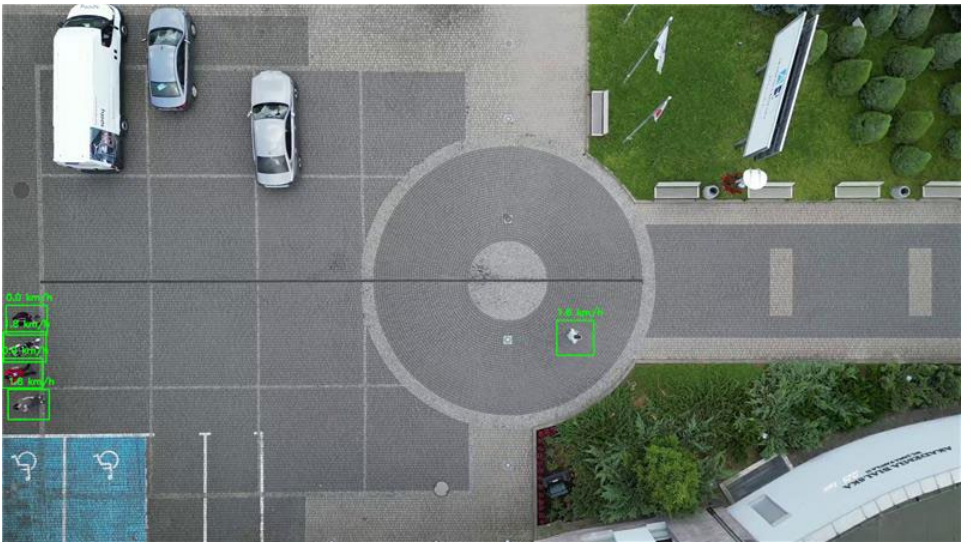
Nagranie zarejestrowane z wysokości 32 metrów było przetwarzane klatka po klatce przez model YOLOv8, który generował ramki ograniczające wykrytych osób, współrzędne ich środków oraz poziom ufności detekcji.

Wydajność modelu okazała się wystarczająca do działania w czasie rzeczywistym. W każdej klatce pozycja środka ramki była zapisywana jako punkt referencyjny służący do obliczania przemieszczeń. Wykorzystano jedynie jedną klasę (person), co pozwoliło uniknąć problemów z wieloklasową analizą sceny.

Detekcja obiektów w kolejnych klatkach wideo nie gwarantuje zachowania tej samej tożsamości obiektów, ponieważ YOLOv8 w każdej klatce wykonuje detekcję niezależnie. Aby zapewnić spójność ruchu poszczególnych osób, konieczne było opracowanie uproszczonego mechanizmu śledzenia obiektów.
Zastosowano procedurę:

1. Dla każdego wykrytego obiektu obliczano odległość pomiędzy jego aktualną pozycją (środek ramki) a pozycjami obiektów z poprzedniej klatki.
2. Jeśli odległość była mniejsza niż 50 pikseli, uznawano, że jest to ten sam obiekt i przypisywano mu ten sam identyfikator ID.
3. Jeśli przekroczono próg 50 px, tworzono nowy identyfikator, co odpowiadało pojawieniu się nowej osoby w scenie.

Metoda ta, choć prosta, okazała się wystarczająca. W analizowanym nagraniu liczba pieszych była niewielka, a trajektorie nie przecinały się w sposób istotnie złożony, dzięki czemu nie wystąpiły problemy charakterystyczne dla gęstych scen. Rezultat przedstawiono na Rysunku 1.



Rysunek 1. Klatka video z uruchomionym algorytmem YOLO-8 trenowanym do detekcji osób

Źródło: opracowanie własne.

Kalibracja przestrzenna i przeliczanie pikseli na metry

Aby oszacować prędkość ruchu pieszych, konieczne było określenie relacji pomiędzy odległością w pikselach a rzeczywistą odległością w metrach. W tym celu zastosowano współczynnik kalibracyjny:

$$pixel_to_meter_ratio = \frac{32 \text{ metry}}{375 \text{ pikseli}} \approx 0{,}085 \text{ m/px}.$$

Wartość ta została wyznaczona na podstawie znanego odcinka znajdującego się w analizowanej scenie, co stanowi standardową praktykę w analizie nagrań UAV, gdy dostęp do parametrów kamery jest ograniczony.
Przemieszczenie obiektu pomiędzy klatkami wyliczano jako:

$$\Delta s = pixel_distance \times pixel_to_meter_ratio.$$

Z kolei czas pomiędzy klatkami, przy znanej liczbie klatek na sekundę nagrania, pozwalał obliczyć prędkość:

$$v = \frac{\Delta s}{\Delta t}.$$

Prędkość wyrażano następnie w km/h.

Filtracja i stabilizacja wyników prędkości

W trakcie badań zaobserwowano, że nawet niewielkie niestabilności detekcji prowadzą do skoków w wartościach prędkości. Dotyczyło to:

- chwilowych przesunięć ramek,
- błędnych detekcji,
- zaniku obiektu na pojedynczą klatkę,
- lokalnych zasłonięć obiektów.

Aby temu przeciwdziałać, wprowadzono dwa mechanizmy stabilizujące. Zastosowano uśrednianie ostatnich pięciu wartości prędkości:

$$v_{smooth}(t) = \frac{1}{5}\sum_{i=0}^{4} v(t-i).$$

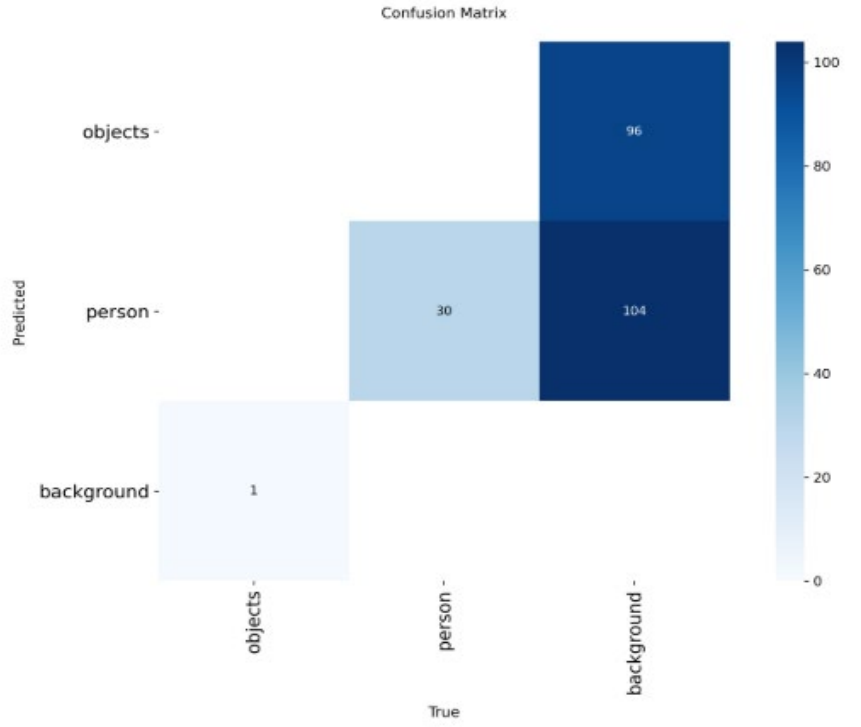
Mechanizm ten redukował fluktuacje wynikające z nieregularnych detekcji.

W przypadku gdy chwilowa prędkość wzrastała o ponad 30% w stosunku do wartości z poprzedniej iteracji, była ona zastępowana wartością wcześniejszą. Zastosowany filtr skutecznie ograniczał wpływ błędów detekcji, zapobiegał występowaniu nienaturalnych skoków prędkości, a jednocześnie umożliwiał swobodne spadki prędkości, na przykład w sytuacji zatrzymania się pieszego. Połączenie obu metod filtracji pozwoliło na uzyskanie stabilnych i realistycznych przebiegów prędkości w czasie. Końcowa wizualizacja wyników obejmowała wyświetlanie ramek ograniczających wokół wykrytych pieszych oraz wygładzone wartości ich prędkości wyrażone w kilometrach na godzinę, przy jednoczesnym pominięciu identyfikatorów obiektów w celu poprawy czytelności prezentowanych danych.

Celem była jak najprostsza prezentacja informacji ilościowej, tj. prędkości ruchu pieszych, bez przeładowania obrazu dodatkowymi elementami.

Ewaluacja działania modelu YOLOv8

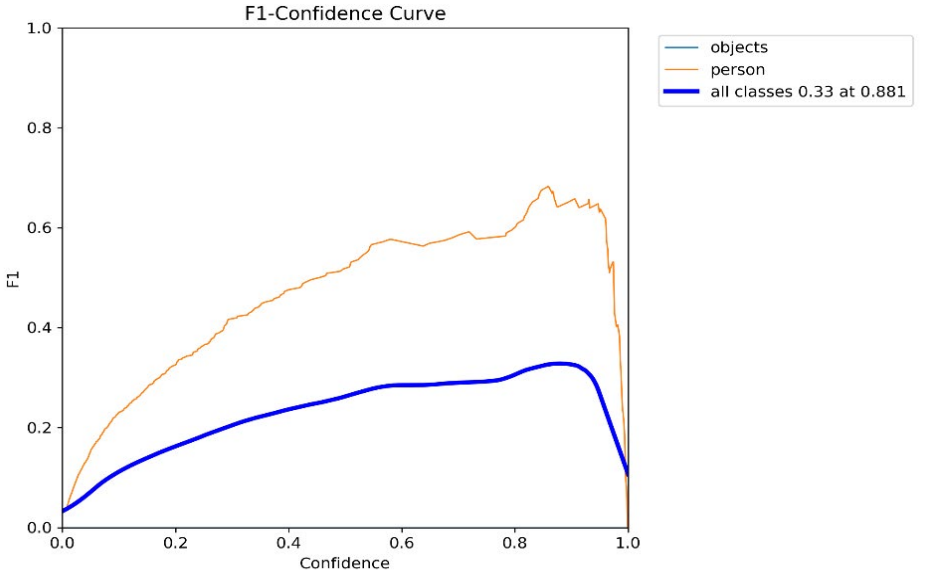
Po przejściu procesu treningu modelu YOLOv8 opracowano macierz pomyłek celem pokazania, jak często detekcje modelu pokrywają się z rzeczywistymi etykietami (Rysunek 2).



Rysunek 2. Macierz pomyłek modelu YOLOv8
Źródło: opracowanie własne.

Wyniki oceny modelu YOLOv8 wskazują, że przy progu ufności równym 0,88 osiągnął on wartość F1-score (Rysunek 3) na poziomie około 0,6. W trakcie analizy odnotowano 104 poprawne wykrycia osób oraz 96 przypadków poprawnego rozpoznania tła jako tła, czyli sytuacji, w których nie występowały obiekty. Jednocześnie zaobserwowano jedynie niewielką liczbę fałszywych wyników pozytywnych i negatywnych. Analiza macierzy pomyłek oraz krzywej F1 w funkcji progu ufności potwierdziła wysoką skuteczność modelu w warunkach

obrazowania z UAV, co podkreśla istotną rolę trenowania sieci na danych dostosowanych do specyfiki perspektywy lotniczej.



Rysunek 3. Zależność F1-score od progu ufności

Źródło: opracowanie własne.

W ramach badań wykonano porównanie detekcji trzech modeli YOLO: YOLOv4, YOLOv5 (wersje s oraz l) i YOLOv8. Analizę możliwości zastosowania w analizie sekwencji wideo z poziomu bezzałogowych statków powietrznych przedstawiono w Tabeli 1.

Wyniki

- YOLOv4 i YOLOv5 nie wykrywały pieszych z wysokości 32 metrów. Detektory te, trenowane na COCO, nie były przystosowane do tak małych obiektów.
- YOLOv8, trenowany na dedykowanym zbiorze UAV, zapewniał detekcje stabilne i użyteczne.

Tabela 1. Porównanie cech algorytmów YOLO v4, v5 i v8

Cecha	YOLOv4	YOLOv5	YOLOv8
Wersja	Darknet	PyTorch	Ultralytics (PyTorch)
Detekcja osób z góry	Bardzo niska	Niska	Wysoka (po dotrenowaniu)
Wymagana dodatkowa nauka	Tak	Tak	Tak, lecz nie w wysokim stopniu złożoności
Integracja z cv2	Wymaga konwersji	Prosta	Natychmiastowa
Czas detekcji (CPU)	31,1 ms	23,3 ms	16,1 ms
Obsługa klas własnych	Ręczna konfiguracja	Wbudowana	Automatyczna (Roboflow)
Stabilność predykcji	Niska przy małych osobach	Niestabilna	Dobra z wygładzaniem

Źródło: opracowanie własne.

Uzyskane wyniki są zgodne z obserwacjami prezentowanymi w literaturze przedmiotu, które wskazują na ograniczoną transferowalność modeli uczonych na danych pozyskiwanych z perspektywy naziemnej do zastosowań wykorzystujących obrazowanie z innych punktów widzenia. Jednocześnie potwierdzają one wyraźną przewagę nowszych wersji algorytmów z rodziny YOLO w zadaniach detekcji obejmujących małe obiekty, dla których kluczowe znaczenie mają zarówno architektura sieci, jak i zastosowane strategie uczenia.

Dyskusja

Wyniki przeprowadzonych badań własnych pozwalają na sformułowanie szeregu istotnych obserwacji dotyczących detekcji pieszych oraz estymacji ich prędkości na nagraniach z UAV przy wykorzystaniu modeli YOLO, w szczególności YOLOv8. Analiza eksperymentalna potwierdza zakładane wcześniej trudności związane z przetwarzaniem obrazów lotniczych, a jednocześnie pozwala na weryfikację skuteczności współczesnych rozwiązań detekcyjnych w tym specyficznym środowisku.

Jednym z kluczowych wniosków wynikających z badań jest potwierdzenie obserwacji powszechnie opisywanej w literaturze (m.in. Kaur i in., 2025; Zeng i in., 2025; Fan i in., 2025), że skuteczna detekcja obiektów w obrazach UAV wymaga dostrojenia modeli do specyfiki danych lotniczych oraz rozwiązań lepiej

radzących sobie z małymi obiektami. Dotyczy to zwłaszcza pieszych obserwowanych z wysokości ok. 30 metrów, których reprezentacja wizualna jest ograniczona do zaledwie kilkudziesięciu pikseli. Badania porównawcze przeprowadzone w niniejszej pracy wykazały, że zarówno YOLOv4, jak i YOLOv5 – pomimo swojej popularności w zastosowaniach naziemnych – nie były w stanie poprawnie wykrywać pieszych w analizowanej scenie.

Z kolei YOLOv8, po odpowiednim dostrojeniu do danych UAV, osiągnął znacząco lepsze wyniki, co potwierdza kluczową rolę dostrojenia algorytmu oraz doboru reprezentatywnego zbioru danych. Osiągnięty F1-score na poziomie około 0,6 przy wysokim progu ufności (0,88) należy uznać za zadowalający w analizowanych warunkach, zwłaszcza że dane pochodzą z perspektywy lotniczej, która charakteryzuje się wysokim stopniem trudności i stosunkowo małą liczbą pikseli reprezentujących sylwetkę pieszego.

Wyniki te są zgodne z podejściami zaprezentowanymi w literaturze – m.in. C4D-YOLOv8 (Zeng i in., 2025), LUD-YOLO (Fan i in., 2025) oraz Enhanced YOLOv8 (Luo, Yuan, 2025) – w których autorzy potwierdzają konieczność wzbogacania informacji kontekstowej i modyfikowania architektury detektora, tak aby lepiej radził sobie z małymi obiektami. W szczególności modyfikacje dotyczące przetwarzania cech wielorozdzielczych odgrywają kluczową rolę w poprawie dokładności detekcji.

Kolejnym istotnym aspektem analizowanych badań jest wpływ stabilności detekcji na estymację prędkości obiektów. Nawet pojedyncze błędne wykrycie, przesunięcie ramki lub chwilowy brak detekcji prowadzą do nienaturalnych skoków w wartościach prędkości, dlatego w literaturze poświęconej estymacji prędkości i ekstrakcji danych ruchowych z materiału UAV podkreśla się znaczenie stabilnej detekcji oraz śledzenia obiektów (Jiao, Fei, 2023; Manjusha i in., 2025). W analizowanym projekcie istotną poprawę stabilności uzyskano dzięki zastosowaniu dwóch metod filtracji: wygładzania średnią ruchomą oraz filtru zmian powyżej 30%. Mechanizmy te znacząco zredukowały liczbę anomalii i poprawiły spójność czasową wyników.

Podejście zastosowane w badaniach własnych jest zbieżne z zalecaniami Ultralytics (2024a), w których podkreśla się konieczność wykorzystania filtracji w zadaniach obejmujących przeliczanie trajektorii na jednostki metryczne. Literatura wskazuje, że przy braku takich mechanizmów nawet bardzo dobre detektory

mogą generować dane ruchowe niskiej jakości z punktu widzenia analizy dynamiki obiektów.

Jednocześnie należy zauważyć, że zastosowane rozwiązania stanowią kompromis pomiędzy złożonością obliczeniową a jakością wyników. Szczególnie interesujące wydaje się porównanie tych metod z bardziej zaawansowanymi algorytmami śledzenia – takimi jak Deep SORT czy ByteTrack – które mogą dodatkowo poprawić spójność trajektorii, choć kosztem większej złożoności implementacyjnej.

W badaniach zdecydowano się na zastosowanie uproszczonego algorytmu śledzenia opartego na minimalnej odległości pomiędzy środkami ramek detekcji. Jest to zgodne z praktyką stosowaną w wielu projektach badawczych, w których liczba obiektów jest niewielka, a trajektorie nie przecinają się w sposób złożony. W analizowanym scenariuszu mechanizm ten zapewnił pełną funkcjonalność systemu, pozwalając na jednoznaczne przypisywanie identyfikatorów ID.

W bardziej złożonych środowiskach – np. przy dużym zagęszczeniu pieszych, licznych zasłonięciach lub dynamicznych zmianach perspektywy UAV – konieczne może być zastosowanie bardziej złożonych metod śledzenia. Literatura sugeruje, że wykorzystanie algorytmów takich jak SORT, Deep SORT, a nawet rozwiązań opartych na transformatorach sekwencyjnych (np. TrackFormer) może znacznie poprawić ciągłość trajektorii, szczególnie w scenach wieloobiektowych.

W niniejszym projekcie zastosowano prostą metodę kalibracji przestrzennej opartą na znanej długości odcinka widocznego w scenie. Jest to podejście powszechnie stosowane w sytuacjach, gdy brak jest pełnych parametrów kamery lub gdy UAV jest platformą ruchomą o zmiennym położeniu. Niemniej jednak metoda ta ma swoje ograniczenia – w szczególności:

- współczynnik piksel-metr może zmieniać się dynamicznie wraz z wysokością lotu,
- nachylenie kamery powoduje zniekształcenia perspektywiczne,
- brak jednorodności rozdzielczości przestrzennej w obrębie kadru.

Dokumentacja Ultralytics oraz liczne prace naukowe wskazują, że dokładniejsze systemy estymacji prędkości powinny wykorzystywać informacje o geometrii kamery, dane telemetryczne UAV lub metody rekonstrukcji 3D. W niniejszym projekcie celem było jednak zbudowanie prototypowego systemu, który

działa w praktycznych warunkach minimalnych, co uzasadnia zastosowanie prostej kalibracji.

Zarówno wyniki własne, jak i obserwacje z literatury, wskazują na kilka kierunków rozwoju omawianych metod:

1. Zwiększenie rozmiaru i różnorodności zbioru treningowego – pozwoli to na poprawę generalizacji modelu YOLOv8 i jego odporności na zmienne warunki lotu UAV.
2. Integracja zaawansowanych algorytmów śledzenia – takich jak Deep SORT, ByteTrack, BoT-SORT, zwiększyłaby to stabilność trajektorii pieszych.
3. Wykorzystanie danych telemetrycznych UAV umożliwiłoby dynamiczną kalibrację przestrzenną, eliminując zniekształcenia perspektywy.
4. Porównanie różnych wariantów YOLO w warunkach UAV – szczególnie ekscytujące byłoby przetestowanie YOLOv9, YOLOv11 oraz wariantów zoptymalizowanych dla małych obiektów (np. SOD-YOLO, LRDS-YOLO).
5. Pełna walidacja na większej liczbie nagrań referencyjnych – najlepiej z rzeczywistymi pomiarami prędkości pieszych.
6. Możliwość integracji analizy ruchu pieszych z modelami predykcji trajektorii – w scenach z UAV mogłoby to znaleźć zastosowanie w inteligentnych systemach bezpieczeństwa.

Wnioski

Analiza literatury oraz przeprowadzone badania własne jasno wskazują, że nowoczesne detektory obiektów z rodziny YOLO – w szczególności YOLOv8 – stanowią obecnie jedno z najbardziej obiecujących narzędzi do analizy obrazów pochodzących z bezzałogowych statków powietrznych (UAV). Ich zdolność do działania w czasie rzeczywistym, połączenie wysokiej dokładności z wydajnością oraz możliwość dostosowania architektury do specyfiki danych lotniczych czynią je wyjątkowo przydatnymi w zadaniach obejmujących monitorowanie ruchu pieszych.

Przegląd literatury potwierdza, że skuteczna detekcja obiektów w scenach lotniczych wymaga dostrojenia modeli do specyfiki danych UAV oraz rozwiązań lepiej radzących sobie z małymi obiektami. Wybrane prace (Kaur i in., 2025; Zeng i in., 2025; Fan i in., 2025) podkreślają znaczenie adaptacji architektury oraz

wzbogacania zbiorów danych, co znajduje potwierdzenie także w wynikach badań własnych. Badania eksperymentalne wykazały, że YOLOv8 – wytrenowany na dostosowanym zbiorze danych UAV – umożliwia skuteczną detekcję pieszych oraz stabilną estymację ich prędkości. Zastosowanie metod filtracji sygnału (średnia ruchoma i filtr gwałtownych zmian) znacząco poprawiło jakość uzyskiwanych wyników i pozwoliło na eliminację artefaktów wynikających z chwilowych błędów detekcji. Prosty algorytm śledzenia, oparty na minimalnej odległości pomiędzy detekcjami, okazał się wystarczający w analizowanej scenie, choć w bardziej złożonych środowiskach zasadna byłaby integracja z bardziej zaawansowanymi metodami śledzenia.

Porównanie YOLOv4, YOLOv5 i YOLOv8 wykazało wyraźną przewagę nowszej architektury, co jest zgodne z trendami opisanymi w literaturze. W szczególności YOLOv4 oraz YOLOv5 – bez dodatkowego trenowania – nie były w stanie wykrywać pieszych na wysokości 30 m, podczas gdy YOLOv8, przystosowany do tej perspektywy, zapewniał detekcje wystarczająco stabilne, aby możliwe było estymowanie prędkości.

Wnioski te prowadzą do kilku kluczowych obserwacji:

1. Nowoczesne architektury YOLO (v8 i nowsze) stanowią fundament praktycznych rozwiązań detekcyjnych dla UAV, pod warunkiem odpowiedniego dostosowania do danych lotniczych.
2. Trenowanie na dedykowanych zbiorach danych jest konieczne, aby uzyskać wysoką skuteczność detekcji pieszych z perspektywy pionowej.
3. Realistyczna estymacja prędkości wymaga zastosowania filtracji oraz śledzenia obiektów, nawet w systemach opartych na wysokiej jakości detektorach.
4. Obszar estymacji prędkości pieszych w danych UAV pozostaje stosunkowo słabo zbadany, co otwiera szerokie możliwości dalszych badań naukowych.
5. Zwiększenie skali zbiorów danych, integracja danych telemetrycznych UAV oraz zastosowanie zaawansowanych metod śledzenia mogą znacząco podnieść jakość przyszłych systemów.

Przedstawiona praca stanowi istotny wkład w rozwój narzędzi i metod analizy ruchu pieszych w obrazach UAV, łącząc aktualny stan wiedzy z praktyczną implementacją działającego prototypu systemu.

Literatura

1. Alhassan, M. A. M., Yılmaz, E. (2025). Evaluating YOLOv4 and YOLOv5 for Enhanced Object Detection in UAV-Based Surveillance. *Processes, 13*(1), 254. https://doi.org/10.3390/pr13010254
2. Bochkovskiy, A., Wang, C.-Y., Liao, H.-Y. M. (2020). YOLOv4: Optimal Speed and Accuracy of Object Detection. *arXiv*. https://doi.org/10.48550/arXiv.2004.10934
3. Fan, Q., Li, Y., Deveci, M., Zhong, K., Kadry, S. (2025). LUD-YOLO: A novel lightweight object detection network for unmanned aerial vehicle. *Information Sciences, 686*, 121366. https://doi.org/10.1016/j.ins.2024.121366
4. Jiao, D., Fei, T. (2023). Pedestrian walking speed monitoring at street scale by an in-flight drone. *PeerJ Computer Science, 9*, e1226. https://doi.org/10.7717/peerj-cs.1226
5. Kaur, D., Battish, N., Bhavsar, A., Poddar, S. (2025). Analysis of vision-based Air-to-Air UAV detection using YOLO based architectures. *Robotics and Autonomous Systems, 193*, 105074. https://doi.org/10.1016/j.robot.2025.105074
6. Luo, W., Yuan, S. (2025). Enhanced YOLOv8 for small-object detection in multiscale UAV imagery: Innovations in detection accuracy and efficiency. *Digital Signal Processing, 158*, 104964. https://doi.org/10.1016/j.dsp.2024.104964
7. Manjusha, M., Niveditha, R., Sunitha, V. (2025). Estimation of YOLO algorithms for extracting traffic data using UAV and stationary camera. *Innovative Infrastructure Solutions, 10*, 151. https://doi.org/10.1007/s41062-025-01898-y
8. Zeng, W., Wu, P., Wang, J., Hu, G., Zhao, J. (2025). C4D-YOLOv8: Improved YOLOv8 for object detection on drone-captured images. *Signal, Image and Video Processing, 19*, 862. https://doi.org/10.1007/s11760-025-04457-1

Netografia

1. Ultralytics. (2024a). Speed Estimation using Ultralytics YOLO26. Pobrane z: https://docs.ultralytics.com/guides/speed-estimation/ (dostęp: 12.12.2025).
2. Ultralytics. (2024b). Comprehensive Guide to Ultralytics YOLOv5. Pobrane z: https://docs.ultralytics.com/yolov5/ (dostęp: 12.12.2025).
3. Ultralytics. (2020). YOLOv5 vs. YOLOv8: Evaluating the Evolution of Ultralytics Vision AI. Pobrane z: https://docs.ultralytics.com/compare/yolov5-vs-yolov8/ (dostęp: 12.12.2025).

ISBN 978-83-68103-30-4