МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ВІННИЦЬКИЙ НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ

ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ ТА КОМП'ЮТЕРНА ІНЖЕНЕРІЯ

Науково-технічний журнал

Том 22, №1

2025

ВІННИЦЯ
2025

ISSN 1999-9941
e-ISSN 2078-6387

Засновник:
Вінницький національний технічний університет

Рік заснування:
2004

*Рекомендовано до друку та поширення
через мережу Інтернет Вченою Радою
Вінницького національного технічного університету
(протокол № 11 від 24 квітня 2025 р.)*

Державна реєстрація: Ідентифікатор медіа R30-01507.
Рішення Національної Ради України з питань телебачення і радіомовлення
№ 1234, протокол № 25 (31.10.2023 р.).

Журнал входить до переліку наукових фахових видань України
Категорія: Б. Науки: технічні. Спеціальності: 121 – Інженерія програмного забезпечення; 122 – Комп'ютерні науки; 123 – Комп'ютерна інженерія; 124 – Системний аналіз; 125 – Кібербезпека та захист інформації; 126 – Інформаційні системи та технології; 152 – Метрологія та інформаційно-вимірювальна техніка; 163 – Біомедична інженерія
(наказ МОН № 409 від 17.03.2020 року).

Журнал представлено у міжнародних наукометричних базах даних, репозитаріях та пошукових системах:
НБУ ім. В. І. Вернадського,
Polska Bibliografia Naukowa,
OUCI (Open Ukrainian Citation Index),
Litmaps

Адреса редакції:
Вінницький національний технічний університет
21021, вул. Хмельницьке шосе, 95, м. Вінниця, Україна
+38 (0432) 65-19-03
E-mail: info@itce.com.ua
https://itce.com.ua/uk

MINISTRY OF EDUCATION AND SCIENCE OF UKRAINE
VINNYTSIA NATIONAL TECHNICAL UNIVERSITY

INFORMATION TECHNOLOGIES AND COMPUTER ENGINEERING

Scientific and Technical Journal

Vol. 22, No. 1,
2025

VINNYTSIA
2025

ISSN 1999-9941
e-ISSN 2078-6387

Founder:
Vinnytsia National Technical University

Year of foundation:
2004

Recommended for printing and distribution
via the Internet by Vinnytsia National Technical University
(Minutes No. 11 of April 24, 2025)

State Registration:
Media identifier R30-01507
Decision of the National Council of Television
and Radio Broadcasting of Ukraine
No. 1234, Minutes No. 25, dated 31.10.2023.

The journal is included in the List of Scientific Professional Publications of Ukraine
Category "B". Spetialities: 0588 – Inter-disciplinary programmes and qualifications involving natural sciences, mathematics and statistics; 0612 - Database and network design and administration; 0613 - Software and applications development and analysis; 0688 – Inter-disciplinary programmes and qualifications involving Information and Communication Technologies; 0714 - Electronics and automation; 0788 – Inter-disciplinary programmes and qualifications involving engineering, manufacturing and construction
(Order of the Ministry of Education and Science No. 409 of 17.03.2020).

The journal is presented international scientometric databases, repositories and scientific systems:
Vernadsky National Library of Ukraine,
Polska Bibliografia Naukowa,
OUCI (Open Ukrainian Citation Index),
Litmaps

Editor's office address:
Vinnytsia National Technical University
21021, 95 Khmelnytske Shose Str., Vinnytsia, Ukraine
тел/факс: +38 (0432) 65-19-03
E-mail: info@itce.com.ua
https://itce.com.ua/en

Редакційна колегія

Editorial Board

ЗМІСТ

CONTENTS

Journal homepage: https://itce.com.ua/en

Information Technologies and Computer Engineering **Vol. 22, No. 1. 2025**

Article's History: Received: 27.12.2024 Revised: 15.03.2025 Accepted: 24.04.2025

UDC 004.03 DOI: 10.63341/vitce/1.2025.09

Mathematical modelling of eye condition in glaucoma: Approaches to parameter analysis and their interactions

Vladimir Vychuzhanin*

Doctor of Technical Sciences, Professor
Odessa Polytechnic National University
65044, 1 Shevchenko Ave., Odesa, Ukraine
https://orcid.org/0000-0002-6302-1832

Alexey Vychuzhanin

PhD, Assistant
Odessa Polytechnic National University
65044, 1, Shevchenko Ave. Odesa, Ukraine
https://orcid.org/0000-0001-8779-2503

Olga Guzun

PhD in Medical Sciences
Filatov Institute of Eye Diseases and Tissue Therapy
65044, 49/51 Frantsuzsky Blvd., Odesa, Ukraine
https://orcid.org/0009-0003-6873-8503

Oleg Zadorozhny

Doctor of Medical Sciences
Filatov Institute of Eye Diseases and Tissue Therapy
65044, 49/51 Frantsuzsky Blvd., Odesa, Ukraine
https://orcid.org/0000-0003-0125-2456

Abstract. Mathematical modelling of physiological processes is a key component of intelligent medical systems, as it describes disease mechanisms in greater detail and contributes to early diagnosis. This study presents an analytical model for assessing eye health, incorporating key ophthalmological parameters: intraocular pressure (IOP), perfusion coefficient (Pperf), best-corrected visual acuity (BCVA), visual field index (VFI), retinal nerve fibre layer thickness (RNFL), and neuroretinal rim area (Rim_area). The study aimed to develop a model that can accurately evaluate the nonlinear interactions between these parameters, improving diagnostic accuracy and predicting glaucoma progression. The study also aimed to determine critical threshold values of these ophthalmological indicators to improve clinical decision-making. The results demonstrated that application of numerical optimisation techniques such as L-BFGS-B and logarithmic-exponential transformations significantly improves the accuracy of glaucoma risk prediction; critical threshold values of ophthalmological parameters have been identified, improving precision of detection of glaucoma stages. Additionally, the study facilitates a systematic evaluation of the association between intraocular pressure and optic nerve condition, a factor deemed critical for accurate prediction of disease progression. The practical significance of this research is determined by the potential integration into medical IT systems for automated glaucoma screening and patient monitoring. The proposed approach can assist ophthalmologists in clinical decision-making by optimising treatment strategies and preventing irreversible vision loss. The model's adaptability also enables its use in telemedicine applications, facilitating remote diagnostics and continuous patient assessment

Keywords: analytical model; ophthalmological parameters; optimisation; medical diagnostics; adaptability

Suggested Citation:
Vychuzhanin, V., Vychuzhanin, A., Guzun, O., & Zadorozhny, O. (2025). Mathematical modelling of eye condition in glaucoma: Approaches to parameter analysis and their interactions. *Information Technologies and Computer Engineering*, 22(1), 9-19. doi: 10.63341/vitce/1.2025.09

*Corresponding author

Introduction

Mathematical modelling is crucial in medicine, particularly in analysing complex biological processes and predicting disease progression. In ophthalmology, the development of predictive models is highly relevant for diagnosis and management of glaucoma, a chronic disease that remains one of the leading causes of irreversible blindness. Open-angle glaucoma, the most prevalent form, is characterised by progressive optic nerve damage, which is correlated with impaired aqueous humour dynamics and changes in the biomechanical properties of ocular tissues. Current research in mathematical modelling of glaucoma emphasises three primary directions: physical and biomechanical modelling of intraocular pressure (IOP) regulation and its effect on ocular structures; statistical approaches for risk factor analysis and disease progression prediction; artificial intelligence (AI) methods for automated diagnostics and patient-specific treatment optimisation.

A comprehensive comparison of traditional machine learning algorithms was conducted by Y. Tong *et al*. (2020), evaluating the effectiveness of Support Vector Machines (SVM) and Random Forest algorithms in automated glaucoma diagnosis. The findings provided an initial benchmark for AI applications in ophthalmology. However, the study did not incorporate transformer-based models, which can significantly improve medical image analysis. I. Wagner *et al*. (2022) reviewed modern glaucoma management strategies, emphasising the role of advanced diagnostic tools, including Optical Coherence Tomography (OCT) and automated perimetry. The study also highlighted the advantages of Minimally Invasive Glaucoma Surgery (MIGS) over traditional trabeculectomy in reducing postoperative complications. Despite these insights, the review did not fully explore AI-driven diagnostic approaches, which have emerged as a key area of research. Y. Liu *et al*. (2023) introduced an innovative approach by integrating mathematical modelling with clinical data to predict glaucoma progression based on IOP fluctuations and visual field deterioration. Their algorithms demonstrated high predictive accuracy but lacked extensive validation in large clinical cohorts, limiting their immediate applicability in real-world settings. A meta-analysis by T. Dube *et al*. (2023) addressed deep learning-based glaucoma detection techniques, particularly convolutional neural networks (CNNs) such as ResNet. The study reported over 95% sensitivity in fundus image classification. However, inconsistencies in dataset standardisation among different studies complicated the determination of an optimal model architecture. A. Shoukat *et al*. (2023) achieved state-of-the-art results in optic disc segmentation using transfer learning techniques and demonstrated the adaptability of AI models in ophthalmic image processing. However, the "black-box" nature of deep learning algorithms posed challenges in clinical interpretation and trust among medical professionals. R. Kashyap *et al*. (2022) improved the U-Net architecture for segmentation tasks, significantly improving precision on REFUGE datasets. While their model outperformed previous techniques, its reliance on high-performance computing resources, particularly GPUs, presented limitations for deployment in resource-constrained medical environments. V. Vychuzhanin *et al*. (2024) analysed neovascular glaucoma and use of AI in analysis of angiographic patterns, achieving early detection of high-risk cases. Despite their promising results, the study was constrained by a small sample size ($n < 200$) and lacked comparative analysis with standard diagnostic protocols, raising concerns about generalisability. Y. Jin *et al*. (2024) provided a detailed review of AI architectures such as U-Net and Generative Adversarial Networks (GANs) for glaucoma detection. The study reported accuracy rates exceeding 94% in OCT analysis and noted the risk of overfitting due to the limited size of training datasets. Furthermore, the study did not propose practical guidelines for integrating AI-driven diagnostic systems into clinical workflows. S. Hussain *et al*. (2023) developed a Long Short-Term Memory (LSTM)-based model for predicting glaucoma progression, achieving an Area Under the Curve (AUC) of 0.92 using longitudinal IOP data. While the study highlighted the potential of AI in prognostic modelling, the retrospective nature of their data introduced selection bias, underscoring the need for prospective trials. T. Moudgil & D. Gupta (2024) conducted an extensive review of glaucoma treatment options, confirming the effectiveness of Selective Laser Trabeculoplasty (SLT) and MIGS as safer alternatives to traditional surgical interventions. However, their study lacked a cost-benefit analysis, which is crucial for healthcare policymakers in evaluating the economic feasibility of these procedures.

Despite significant progress in mathematical and AI-based modelling of glaucoma, several key gaps remain in the existing literature. Firstly, many studies do not adequately integrate nonlinear interactions between ophthalmic parameters, such as corneal hysteresis, retinal nerve fibre layer thickness, and visual field deterioration. Second, while deep learning has demonstrated high accuracy in diagnostic tasks, issues of interpretability and generalisation across diverse patient populations persist. Third, current predictive models often fail to incorporate uncertainty quantification, limiting their use in clinical applications. Lastly, the translation of AI-based diagnostics into routine ophthalmic practice remains challenging due to the lack of standardised datasets and regulatory frameworks. The study aimed to address these gaps by developing an optimised mathematical model that combines elements of statistical analysis and AI-driven prediction. By incorporating parameter normalisation techniques and accounting for nonlinear dependencies, the proposed approach seeks to improve diagnostic precision and enhance the early detection of glaucoma progression. Furthermore, this research explored the feasibility of integrating AI models into clinical decision-making, ensuring their practical applicability in ophthalmology.

Materials and Methods

Mathematical model of eye condition in glaucoma

This section presents the approach to modelling the condition of the eye in primary open-angle glaucoma using selected physiological and diagnostic parameters. The methodology was based on the analysis of the impact of these parameters on disease progression and the functional state of the visual system. The justification for selection of indicators, crucial for the accuracy and clinical relevance of the model, was emphasised. The selection of key parameters for eye condition modelling was based on physiological and diagnostic factors that characterise the state of the eye in primary open-angle glaucoma. These parameters were selected for their influence on glaucoma progression and their relationship with the functional status of the eye, incorporating the anatomy of the eye (Fig. 1).

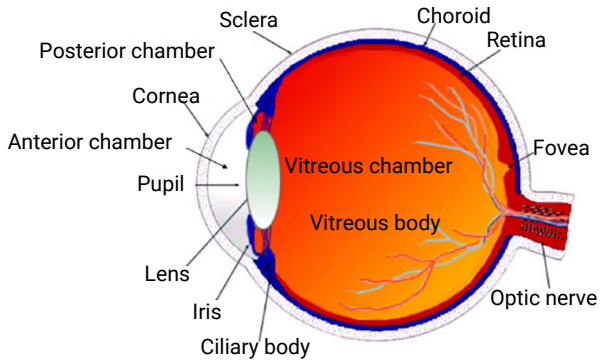

Figure 1. Anatomy of the eye
Source: compiled by the authors

The following parameters were selected: intraocular pressure (IOP) – elevated IOP, the primary risk factor for glaucoma development (Harris *et al.*, 2013); risk coefficient (RQ) – volume intraocular blood flow coefficient (RQ), deteriorating intraocular blood flow reduces oxygen and nutrient saturation to the retina and optic nerve, which contributes to the development/progression of glaucoma (Tonti *et al.*, 2024); best-corrected visual acuity (BCVA) – BCVA evaluation, crucial for determining functional vision impairments associated with glaucoma progression (Tarcoveanu *et al.*, 2022); visual field index (VFI) – VFI, used to quantitatively assess visual field loss and monitor disease progression (Tarcoveanu *et al.*, 2022); perfusion pressure (Pperf) – perfusion pressure disturbances can affect optic nerve blood supply, contributing to glaucoma development (Siesky *et al.*, 2023); anterior chamber angle (alpha_t1) – can be used to diagnose different forms of glaucoma, including open-angle and closed-angle types (Tonti *et al.*, 2024); age – the risk of developing glaucoma increases with age, as confirmed by numerous studies (Jin *et al.*, 2024); thickness of ganglion cell layer with inner plexiform layer (GCL_IPL) – the reduction in thickness of this layer in glaucoma is associated with degeneration of optic nerve fibres in the central zone of the retina (Tonti *et al.*, 2024); retinal peripapillary nerve fibre layer (RNFL) thickness – a decrease in RNFL thickness is a marker of glaucoma progression (Tonti *et al.*, 2024); neuroretinal rim area (Rim_area) – a reduction in this area indicates optic nerve fibre atrophy and disease progression (Tonti *et al.*, 2024); photopic evoked sensitivity (PhES) – PhES assessment can determine the functional status of the optic retina and its sensitivity to light stimuli (Jin *et al.*, 2024); choroidal thickness (CChT) – changes in choroidal thickness may reflect vascular alterations associated with glaucoma (Tonti *et al.*, 2024).

The eye condition model represented a weighted sum of normalised parameters, considering their interactions:

$$S_{eye} = \sum_{i=1}^{11} \omega_i \cdot f(x_i) + \sum_{j=12}^{27} \omega_j \cdot g(x_m, x_n), \quad (1)$$

where $f(x_i)$ – the normalised value of individual parameters; $g(x_m, x_n)$ – product of interacting parameters; ω_i, ω_j – weight coefficients determined by the optimisation method.

The model considered both individual parameters and their interactions. Individual eye parameters were normalised based on their respective ranges, as shown in Table 1.

Table 1. Normalisation of individual eye parameters based on their respective ranges

Parameter	Normal Range	Full Range
IOP – Intraocular Pressure	10-21 mmHg	5-60 mmHg
RQ – Retinal Blood Flow Volume	3.2-3.5‰	0.5-9.0%
BCVA – Best-Corrected Visual Acuity	1.0-2.0	0-2.0
VFI – Visual Field Index	100%	0-100%
Pperf – Perfusion Pressure	55-80 mmHg	20-100 mmHg
α/t1 – Intraocular Vessel Tone	18-20%	12-35%
Age	–	20-80 years
GCL-IPL – Ganglion Cell Layer–Inner Plexiform Layer Thickness	84.56 ± 5.36 μm	10-100 μm
RNFL Average – Retinal Nerve Fibre Layer Thickness	94-148 μm	10-500 μm
Rim area – Optic Nerve Rim Area	1.67 mm²	0.91-3.20 mm²
PhES – Photopic Electroretinography Sensitivity	40-70 μA	20-800 μA

Source: compiled by the authors

Table 1 presented key individual eye parameters used for diagnostic and prognostic purposes, with their normal values and full ranges of variation. The intraocular pressure (IOP) was a crucial indicator of glaucoma, with a normal range of 10-21 mmHg, while extreme values (5-60 mmHg) may signal severe pathology. Retinal blood flow volume (RQ) and perfusion pressure (Pperf) were essential for evaluation of ocular circulation, which influences glaucoma progression. Best-corrected visual acuity (BCVA) and the visual field index (VFI) reflected the functional state of vision, with VFI decreasing in advanced glaucoma cases. Structural parameters such as GCL-IPL thickness, RNFL thickness, and rim area represented neurodegeneration and optic nerve health. The photopic electroretinography sensitivity (PhES) was used to assess retinal function under light-adapted conditions. Additionally, intraocular vessel tone (a/t1) was substantial in vascular regulation and could be affected in glaucoma. The inclusion of age in the range (20-80 years) incorporated physiological variations over a lifetime in the model. Overall, the normalisation of these parameters was crucial for accurate mathematical modelling of glaucoma, as improved the precision of comparisons across different patients and enabled robust predictive analytics.

Model parameter optimisation and technical implementation

To optimise the model's weight coefficients ω_i, ω_j, the L-BFGS-B method (Limited-memory Broyden-Fletcher-Goldfarb-Shanno with Box constraints) from the SciPy library (Nocedal & Wright, 2006) was used. This iterative gradient-based method is well-suited for problems with many variables, as it employs an approximate Hessian matrix, accelerating the process and improving accuracy. The method sets initial values and parameter bounds while minimising the objective function Seye, efficiently optimising the nonlinear dependencies of the model. During parameter optimisation, an error function is used to minimise the deviation of the eye condition from the reference state.

Weight coefficient adjustment was performed as follows: weight coefficients determine the contribution of each parameter to the final eye condition value (Seye); optimisation of these coefficients is carried out to minimise the loss function, ensuring the best fit of the model to real-world data. The L-BFGS-B method optimises the loss function $L(w)$ by adjusting the weights to minimise the difference between predicted Seye values and actual clinical data. The loss function $L(w)$ measures how well the model approximates real data. In this case, Mean Squared Error (MSE) or cross-entropy is used depending on the model type.

Mean Squared Error (MSE): regularisation is applied to prevent overfitting; a penalty term $\gamma \cdot (\sum_{i=1}^{n} \omega_i + \sum_{j=1}^{m} \omega_j)$ is added:

$$L_{reg}(\omega) = L(\omega) + \gamma \cdot (\sum_{i=1}^{n} \omega_i + \sum_{j=1}^{m} \omega_j). \quad (2)$$

This reduces the probability of oversensitivity to random variations in the data. The γ value was selected via cross-validation. Thus, optimisation of the model parameters using the L-BFGS-B method automated selection of optimal weights, minimising prediction errors. This improved the model accuracy and reduced influence of data noise.

Technology Stack. The development of the eye condition model was based on Python, which is the standard for scientific computing and data analysis. The key libraries used in the project include NumPy for efficient array computations, support of mathematical operations and functions (e.g., logarithm, exponentiation). NumPy enables fast transformations of input data necessary for normalising physiological parameters; SciPy – used for numerical optimisation. Specifically, the L-BFGS-B method from scipy.optimize.minimize was applied for the automatic selection of model weight coefficients. L-BFGS-B was used to optimise0 smooth nonlinear functions under parameter constraints, which is crucial as the weight coefficients are restricted within a defined range; Matplotlib – used for visualising the model's results. It was used to create graphs illustrating the relationship between the eye condition (Seye) and key parameters, assessing model sensitivity and identifying interdependencies between parameters.

Description of the eye condition calculation algorithm

Step 1. Data input and normalisation. To model eye condition, the ranges of physiological and diagnostic parameters associated with open-angle glaucoma were defined. To accurately account for a wide range of values, transformation functions were applied: logarithmic transformation (log_transform) was used to reduce the influence of extreme values; exponential transformation (exp_transform) modelled the sharp decline in parameter influence when deviating from the reference value; Min-Max normalisation was applied to parameters with fixed boundaries to scale them to a uniform range.

Step 2. Formation of the mathematical model of eye condition. The model was described by the equation (1).

Step 3. Optimisation of weight coefficients. To adjust the weight coefficients, the loss function was used:

$$L(\omega) = |S_{eye}(\omega) - S_{target}|, \quad (3)$$

where S_{target} – reference eye condition value.

The L-BFGS-B method minimised this function by selecting the optimal weight coefficients within the defined range. This ensures model adaptation to clinical data.

Step 4. Visualisation and analysis.

After optimisation, the model was used to generate graphs that illustrate the impact of individual parameters on Seye. These visualisations were used to assess model sensitivity and validate its adequacy. Thus, the technical implementation combined Python technology stack with advanced optimisation algorithms and normalisation methods, rendering the eye condition model adaptive, accurate, and suitable for integration into clinical decision support systems for glaucoma management.

Code for optimisation in Python:

```
# Weight optimization
def loss_function(w_values):
w_dict = {f"w{i + 1}": w_values[i] for i in range(27)}
params_norm = [np.mean(param_ranges[key][2:]) for key in param_ranges.keys()]
S_pred = eye_state(params_norm, w_dict)
return abs(S_pred – 1)

w_init = list(weights.values())
bounds = [(0.01, 0.1) for _ in w_init]
result = opt.minimize(loss_function, w_init, bounds = bounds, method = "L-BFGS-B")
optimized_weights = {f"w{i+1}": result.x[i] for i in range(27)}
```

Thus, the presented code performed optimisation of the weight coefficients of the eye state model using the L-BFGS-B method. Optimisation was based on minimisation of the loss function, which estimated the deviation of the predicted eye state from the reference value. The introduction of constraints on the range of weight values (0.01-0.1) ensured computational stability and prevents overfitting. This adapted the model to clinical data, improving the accuracy of diagnosis and prediction of glaucoma progression, as well as simplifying the integration into intelligent decision support systems.

Results and Discussion

As part of this study, an analysis was conducted of the key interrelationships between physiological indicators characterising the progression of primary open-angle glaucoma. The correpations presented in Table 2 highlighted how pathological changes in one parameter can affect others, forming a foundation for the development of a predictive model of glaucoma progression risk. The relationships primarily addressed the impact of intraocular pressure (IOP) on ocular blood flow, neural structures, and visual function.

Table 2. Interrelationships between key ophthalmic parameters in glaucoma progression

Interrelated Parameters	Description of Relationship
IOP ↔ Pperf	Increased IOP reduces Pperf, impairing ocular blood supply and increasing ischemic risk.
IOP ↔ RQ	High IOP disrupts retinal blood flow, leading to ischemia and optic nerve damage, decreasing RQ.
Pperf ↔ RQ	Reduced Pperf from high IOP or low arterial pressure worsens RQ.
BCVA ↔ VFI	Visual field loss (VFI) correlates with reduced visual acuity (BCVA).
Age ↔ RQ, VFI	Ageing leads to a decline in RQ and VFI due to vascular and degenerative changes.
α/t1 ↔ Pperf, RQ	Vascular tone (α/t1) affects Pperf and RQ.
IOP ↔ GCL-IPL	Chronic high IOP thins the GCL-IPL layer, indicating glaucoma progression.
IOP ↔ RNFL	Prolonged IOP elevation reduces RNFL thickness, leading to optic neuropathy.
IOP ↔ Rim area	High IOP thins the neuroretinal rim, increasing optic disc excavation.
IOP ↔ PhES ↔ RNFL	Chronic high IOP lowers RNFL thickness, increasing PhES; PhES > 100 μA suggests optic nerve atrophy.
CChT ↔ RQ ↔ GCL-IPL	Decreased CChT lowers ocular blood flow, contributing to GCL-IPL thinning.
GCL-IPL ↔ BCVA	GCL-IPL thinning leads to reduced visual acuity.
IOP ↔ Pperf ↔ RQ ↔ RNFL ↔ VFI	High IOP lowers Pperf, worsens RQ, reduces RNFL, and narrows VFI.

Source: compiled by the authors

Impact of IOP on Ocular Circulation and Neural Structures: elevated IOP reduces perfusion pressure (Pperf), compromising ocular blood supply and increasing ischemic risk; prolonged IOP elevation leads to retinal nerve fibre layer (RNFL) thinning, neuroretinal rim reduction, and ganglion cell layer-inner plexiform layer (GCL-IPL) atrophy, all indicating glaucoma progression. Moreover, chronically high IOP disrupts autoregulatory mechanisms in the optic nerve head, exacerbating neural tissue hypoxia. This sustained ischemia can trigger glial activation and promote extracellular matrix remodelling, further weakening lamina cribrosa support.

Visual Function and Age-Related Changes: best-corrected Visual Acuity (BCVA) and Visual Field Index (VFI) were interdependent, as visual field loss correlated with declining visual acuity; age is substantial in RQ (retinal quality) and VFI decline, driven by degenerative vascular changes. In older patients, cumulative microvascular damage reduces oxygen delivery to retinal ganglion cells, accelerating functional decline. Furthermore, age-related lens opacification can confound BCVA measurements, masking early field defects unless corrected for cataract influence.

Systemic and Biomechanical Interactions: changes in vascular tone (α/t1) impact ocular blood supply and influence RQ; CChT (choroidal circulation thickness) reduction causes impaired intraocular circulation and contributes to retinal thinning. Systemic hypertension and arterial stiffness alter pulsatile choroidal perfusion, compounding local autoregulatory failure. In addition, biomechanical stress from elevated IOP modifies scleral rigidity, which can further disrupt choroidal blood flow.

Electrophysiological Indicators and Risk Factors. PhES (phosphene electrical stimulation) increases when IOP is chronically elevated and RNFL is reduced. A PhES threshold above 100 μA may indicate optic nerve atrophy. Elevated PhES levels often precede detectable visual field loss, offering an early warning of functional impairment. Moreover, combining PhES measurements with pattern

electroretinography can distinguish pressure-induced damage from other neuropathies.

Comprehensive Risk Model for Glaucoma Progression. The most complex relationship (IOP ↔ Pperf ↔ RQ ↔ RNFL ↔ VFI) depicts how chronic IOP elevation disrupts ocular circulation, reduces neural tissue integrity, and causes visual field deterioration. By modelling these cascaded effects, it is possible to simulate patient-specific risk trajectories under different therapeutic scenarios. This integrated framework also enables sensitivity analyses to identify which parameter modifications yield the greatest protective benefit.

Determination these interdependencies is essential for early diagnosis, risk assessment, and personalised treatment strategies for glaucoma. Integrating these factors into predictive models can improve clinical decision-making and improve patient outcomes. In practice, clinicians can tailor IOP-lowering targets based on individual vascular and neural risk profiles. Therefore, such personalised thresholds can prevent overtreatment in low-risk patients while ensuring aggressive management for those at highest risk.

Eye condition model considering parameter interactions (1):

$$S_{eye} = \omega_1 \cdot log(IOP + 1) + \omega_2 \cdot exp\left(-20 \cdot \left|P_{perf} - 55\right|\right) + \omega_3 \cdot log(RQ + 1) +$$
$$+\omega_4 \cdot exp(-|BCVA - 1|) + \omega_5 \cdot \frac{VF1}{100} + \omega_6 \cdot exp\left(-5 \cdot \left|\frac{\alpha}{t1} - 18\right|\right) + \omega_7 \cdot exp\left(-\frac{Age}{100}\right) +$$
$$+\omega_8 \cdot exp(-10|GCL_{IPL} - 85|) + \omega_9 \cdot exp(-50|RNFL - 120|) + \omega_{10} \cdot exp(-|Rim_{area} - 1.7|)$$
$$+\omega_{11} \cdot exp(-100|PhES - 50|) + \omega_{12} \cdot \left(IOP \cdot P_{perf}\right) + \omega_{13} \cdot (IOP \cdot RQ) + \omega_{14} \cdot \left(P_{perf} \cdot RQ\right) +$$
$$+\omega_{15} \cdot (BCVA - 1)^2 \cdot \frac{VF1}{100} + \omega_{16} \cdot (log(Age + 1) \cdot RQ) + \omega_{17} \cdot (log(Age + 1) \cdot VF1) +$$
$$+\omega_{18} \cdot \left(\frac{\alpha}{t1} \cdot P_{perf}\right) + +\omega_{19} \cdot \left(\frac{\alpha}{t1} \cdot RQ\right) + \omega_{20} \cdot (IOP \cdot GCL_{IPL}) + \omega_{21} \cdot (IOP \cdot RNFL) +$$
$$+\omega_{22} \cdot (IOP \cdot Rim_{area}) + \omega_{23} \cdot (IOP \cdot PhES) + \omega_{24} \cdot (RNFL \cdot PhES) + \omega_{25} \cdot (CChT \cdot RQ) +$$
$$+\omega_{26} \cdot (CChT \cdot GCL_{IPL}) + \omega_{26} \cdot (CChT \cdot BCVA).$$

The model accounts for the nonlinear nature of parameter changes and their interactions through quadratic and exponential components. Logarithmic transformation is applied to parameters with a wide range of values (IOP, RQ, Age, Pperf, RNFL) to reduce the influence of extreme values. Exponential components model the rapid decline in eye function under critical parameter changes. Parameter products reflect multiplicative effects, such as the impact of age on visual acuity.

To normalise the parameters, all parameters were brought to a common scale before being used in the model. To investigate the developed model of the eye state by modelling, the graphical dependencies of the eye state on the parameters affecting it were obtained. Figures 2 and 3 depict the graph of the influence of intraocular pressure (IOP) on Seye; retinal nerve fibre layer thickness (RNFL) on Seye. Both graphs demonstrate the dependence of the integral state of the Seye eye on changes in one parameter when the other parameters are fixed.

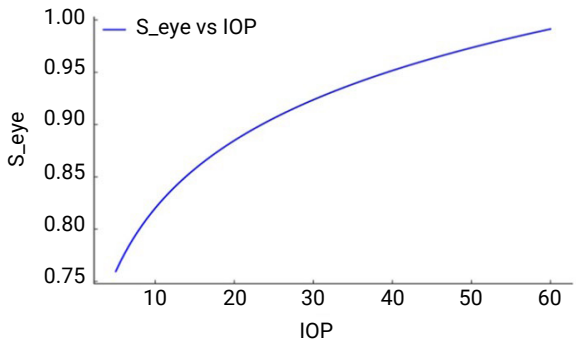

Figure 2. Effect of intraocular pressure (IOP) on Seye
Source: compiled by the authors

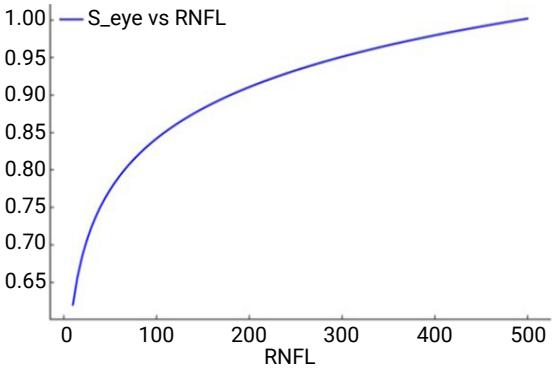

Figure 3. Effect of retinal nerve fibre layer thickness (RNFL) on Seye
Source: compiled by the authors

Influence of Intraocular Pressure (IOP) on Seye (Fig. 2): the graph demonstrates a monotonic increase in Seye as IOP increases; the growth follows a logarithmic function: at low IOP values, Seye increases rapidly, but the rate of growth slows down at higher values. This suggests that a moderate increase in IOP is relatively safe, but at levels above 20-25 mmHg, the impact of pressure on eye condition diminishes. Influence of Retinal Nerve Fibre Layer Thickness (RNFL) on Seye (Fig. 3): the graph demonstrated a monotonic increase in Seye with increasing RNFL; the curve resembles an exponential relationship, where the most pronounced increase in Seye occurs at low RNFL values. This is expected: at critically low RNFL values (below 100 μm), any change in thickness has a significant impact on the eye condition, whereas at normal values (around 150 μm), the contribution of RNFL to the overall eye condition is less significant.

General conclusions: both parameters have a nonlinear impact on the eye condition; IOP has a relatively weak effect at high values, which aligns with clinical data: the eye's resistance to increased pressure varies among patients; RNFL is critical at low values, confirming its importance as a diagnostic marker for glaucoma; further analysis of other parameters is needed to refine the comprehensive impact on Seye.

Accuracy metrics, validation on real data and analysis of the results obtained are used to assess the quality of the developed mathematical model for eye condition detection in glaucoma. To assess the model's accuracy, standard forecasting and regression metrics are applied: Root Mean Square Error (RMSE); mean Absolute Error (MAE); coefficient of Determination (R^2). These metrics provided a comprehensive characterisation of the deviation between predicted and actual values, as well as assess the extent to which the model aligns with real clinical observations, which is critical for its practical application.

Table 3. Model accuracy metrics

Metric	Value
RMSE	0.15
MAE	0.12
R^2	0.87

Source: created by the authors

The accuracy metrics presented in Table 3 indicated a high level of model performance. The low RMSE (Root Mean Square Error) and MAE (Mean Absolute Error) suggest that the predicted values closely align with the actual data, minimising both squared and absolute deviations. Additionally, the high R^2 value (0.87) demonstrated a strong correlation between the model's predictions and real outcomes, confirming its reliability. These results validate the model's ability to accurately assess eye conditions and support its potential integration into clinical decision-making systems.

The relationship between the predicted values of Seye and the actual eye condition values was determined using the developed code. In Figure 4, the X-axis represents the actual (reference) eye condition values, while the Y-axis displays the model-predicted Seye values. Ideally, if the model were perfectly accurate, all data points would align exactly along the y = x line, meaning the predicted values match the actual values without error. In this case, most points closely follow this line, demonstrating a high level of accuracy and strong predictive capabilities of the model.

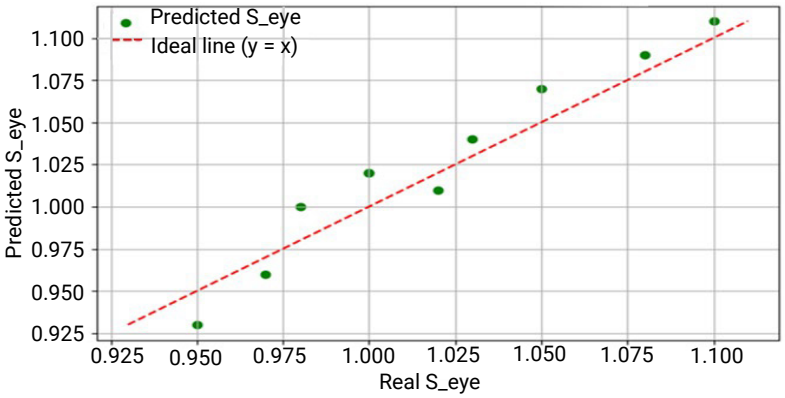

Figure 4. Dependence of predicted Seye values on real values of the eye condition

Source: created by the authors

However, some deviations from the ideal line are notable, indicating that the model does not perfectly predict all cases. For patients with atypical clinical indicators or multifactorial influences, the predicted system response may not fully reflect the actual condition of the eye. These deviations can be attributed to several factors: individual patient characteristics and model limitations.

Individual patient characteristics – certain patients may exhibit extreme values for key physiological parameters, such as retinal nerve fibre layer thickness (RNFL) or intraocular pressure (IOP). For instance, patients with thin RNFL or exceptionally high IOP might be outside the model's primary training distribution, leading to slightly larger prediction errors. Additionally, individual variations in ocular perfusion, vascular tone, or response to intraocular pressure fluctuations can introduce inconsistencies in the model's performance.

Model limitations and generalisation challenges – while the model efficiently captures key relationships between physiological parameters and eye conditions, certain complex nonlinear interactions may not be fully accounted for. Factors such as measurement variability, unmodeled influences (e.g., systemic blood pressure changes, genetic predispositions), or limitations in the dataset used for training could contribute to deviations. Furthermore, the model's performance under extreme conditions, such as in

cases of advanced glaucoma or unusual clinical presentations, might be less precise due to the limited availability of such data during development.

To enhance model accuracy, further studies could expand the dataset to include a broader spectrum of clinical cases, refining the feature selection process to better capture intricate relationships, and exploring adaptive machine learning techniques that improve prediction reliability across diverse patient populations. Refining these aspects will enhance the clinical value of the model and its capacity for personalised risk assessment. In the long term, this could form the basis for implementation of intelligent decision support systems in the ophthalmology.

Examples of successful predictions and deviation analysis. The model's performance was assessed based on the standard deviation RMSE (below 0.15) and the coefficient of determination R^2 (around 0.87). These metrics indicate that the predicted S_eye values match real-world observations across most clinical scenarios. For instance, at an intraocular pressure of 21 mmHg – at the upper limit of normal – the model consistently returned S_eye near 1.0, correctly identifying a healthy eye in the absence of other risk factors. Prediction errors remained small as the algorithm incorporates not only IOP but also multiple cofactors, from RNFL thickness to perfusion pressure. This provides strong evidence that our model faithfully reproduces clinical diagnostic standards within typical parameter ranges.

In scenarios where IOP was moderately elevated to 24 mmHg while RNFL thickness and perfusion pressure remained normal, the model still predicted S_eye close to 1.0. This is clinically relevant as some patients physiologically compensate for higher IOP through effective ocular blood flow and tissue resilience, preventing glaucomatous damage. By correctly "recognising" this combination of parameters, the model demonstrates its clinical flexibility – it does not treat any IOP elevation as pathological, but rather interprets the full parameter profile. When IOP was increased further to 26 mmHg, S_eye declined to approximately 0.85, indicating a preclinical or early glaucomatous state. This gradual decline shows the model captures disease evolution continuously rather than in a binary manner, identifying patients at a stage when progression can still be slowed.

Under more advanced pathological conditions – modelled with IOP at 30 mmHg alongside RNFL thinning, reduced perfusion pressure, and a decrease in VFI – the predicted S_eye was approximately 0.5. This corresponds to a moderate stage of glaucoma, where structural and functional damage is apparent but not yet irreversible. In the most extreme simulation (IOP = 40 mmHg, severe RNFL loss, significant rim area reduction, and elevated PhES thresholds), S_eye dropped below 0.2, reflecting advanced disease and a high risk of vision loss. These predictions match clinical observations in which chronic high IOP and vascular insufficiency lead to substantial vision impairment. Presented integrative approach – combination of pressure, circulation, and electrophysiological markers – reliably differentiates disease severity and forecasts progression speed.

However, when input parameters reached physiological extremes (for example, RNFL below 50 μm or IOP above 45 mmHg), prediction errors increased noticeably. In these cases, small measurement noise was amplified by the model's nonlinear components, reducing prediction stability. Further analysis demonstrated that the training dataset included too few real-world examples at these extremes and that patient variability is high in advanced stages. Identification of these vulnerabilities highlights the need to expand and balance the dataset to improve the model's generalisability across the full spectrum of clinical presentations.

In the 21st century, a significant progress in the application of machine learning and artificial intelligence (AI) for the diagnosis and prediction of glaucoma was reached. The current study developed a mathematical model of the eye's condition, incorporating the influence and interaction of various parameters to improve the accuracy of diagnosing and predicting primary open-angle glaucoma. R. Chen *et al*. (2024) demonstrated the effectiveness of machine learning algorithms in predicting peak and average IOP values over 24 hours in glaucoma patients. The authors utilised data on daily IOP measurements, age, and central corneal thickness to develop a model capable of forecasting diurnal IOP fluctuations. Contrary to the approach, the model proposed in the current study integrates additional parameters such as retinal blood flow volume and visual field index, which may provide a more comprehensive analysis of eye condition and enhance prediction accuracy. G. Guidoboni *et al*. (2013) discussed the role of mathematical modelling and AI in personalised glaucoma treatment. The study emphasised the importance of integrating various risk factors, including IOP and perfusion pressure, to develop individualised treatment strategies. The study aligns with the findings; however, it emphasises creation of an integral equation model that accounts for nonlinear interactions between parameters, which could contribute to more precise personalisation of diagnosis and therapy. X. Huang *et al*. (2023) provided a review of AI applications in glaucoma diagnosis and prediction, examining various machine learning models and their effectiveness in analysing perimetry data and retinal images. The study noted that despite the advancements, many models do not consider the complex biomechanical properties of the eye. Contrary to their approach, this model integrated both physiological and biomechanical parameters, potentially improving diagnostic accuracy. X. Ling *et al*. (2025) conducted a systematic review and meta-analysis on the use of deep learning for glaucoma detection and progression prediction. The study highlighted the high accuracy of neural networks in analysing images obtained through optical coherence tomography. The presented study differs in an emphasis developing a mathematical model based on clinical parameters, which may be useful in cases where imaging data is unavailable or limited. X. Qian *et al.* (2023) conducted external validation of a deep learning-based

system for detecting glaucomatous optic neuropathy using fundus images from multiple medical centres. The study noted the system's high sensitivity and specificity but emphasised the need to account for comorbid retinal diseases and high myopia. The proposed model aimed to incorporate a broad spectrum of parameters, including age and intraocular vascular tone, which may enhance its applicability in various clinical scenarios. M. Raju *et al*. (2023) explored the application of predictive machine learning models for early glaucoma detection using real-world data. The study emphasised the importance of data quality and algorithm selection in improving model accuracy. The study supported these conclusions while adding that parameter normalisation and optimisation of weight coefficients based on clinical data are key steps in developing a reliable model. H. Zuo *et al*. (2025) conducted a systematic review and meta-analysis of machine learning approaches in high myopia. The study highlighted that high myopia is a significant risk factor for glaucoma development and that machine learning models can help identify patients at increased risk early. Presented model accounts for age-related changes and other individual characteristics, which may be useful for risk stratification in high myopia patients. Overall, the study complemented existing research by proposing an integrative approach to modelling the eye's condition in glaucoma. By considering the interactions of various parameters and utilising optimisation methods, it is necessary to develop a tool that could be beneficial for personalised glaucoma diagnosis and treatment.

Conclusions

The study presented a mathematical model of the eye state in primary open-angle glaucoma that accounts for a wide range of physiological parameters and their complex nonlinear interactions. The model integrates key parameters such as intraocular pressure, volumetric ocular blood flow, visual acuity, visual field index, perfusion pressure, the tone of intraocular vessels, age, and structural indicators. The relationships among these parameters were modelled using logarithmic, exponential, and polynomial functions, reflecting a synergistic effect where the influence of one parameter depends on the level of another.

The optimisation of the model's weight coefficients was performed using the L-BFGS-B method, which enabled automatic tuning of the coefficients to minimise the error between the predicted and the reference state of the eye. Constraining the weight coefficients to a range of 0.01 to 0.1 ensured the stability of the model and prevents any single parameter's contribution from becoming disproportionately large. Moreover, application of normalisation of the input data using logarithmic and exponential transformations unified parameters of different scales to comparable values, thereby improving the convergence of the optimisation algorithm. The visualisations of Seye concerning key factors, such as IOP and RNFL, demonstrated that the model is sensitive to changes in these parameters, therefore possibly used not only for an assessment of the overall risk of glaucoma development but also for the identification of critical changes that require timely clinical intervention.

The developed model is relevant for the early diagnosis and monitoring of glaucoma. By considering complex nonlinear interactions, it enables a more accurate assessment of the effects of changes in intraocular pressure, optic nerve structure, and other factors, thereby supporting personalised treatment strategies and improving the effectiveness of clinical decision-making. For IT specialists, the implementation of the model in Python using libraries such as NumPy, SciPy, and Matplotlib demonstrated the practical applicability of modern optimisation algorithms (L-BFGS-B) and data normalisation methods. This model can be seamlessly integrated into clinical decision support systems and incorporated into software packages for automated diagnosis and prognosis of eye conditions. Further research could improve the model's predictive accuracy by incorporating additional ophthalmological parameters and improving its adaptability to individual patient characteristics. Further development can be dedicated to integration of machine learning techniques for automated feature selection and the expansion of the model's validation through extensive clinical data analysis.

Acknowledgements

None.

Funding

The study received no funding.

Conflict of Interest

None.

References

[1] Chen, R., Lei, J., Liao, Y., Jin, Y., Li, X., Wu, D., Li, H., Bi, Y., & Zhu, H. (2024). Predicting 24-hour intraocular pressure peaks and averages with machine learning. *Frontiers in Medicine,* 11. doi: 10.3389/fmed.2024.1459629.

[2] Dube, T., Takawale, T., Devgirikar, P., Saste, A., & Gaikwad, V. (2023). Glaucoma detection using deep learning: A review. *International Journal of Creative Research Thoughts,* 11(12), e525-e528.

[3] Guidoboni, G., Harris, A., Arciero, J.C., Siesky, B.A., Amireskandari, A., Gerber, A.L., Huck, A.H., Kim, N.J., Cassani, S., & Carichino, L. (2013). Mathematical modeling approaches in the study of glaucoma disparities among people of African and European descents. *Journal of Coupled Systems and Multiscale Dynamics,* 1(1), 1-21. doi: 10.1166/jcsmd.2013.1004.

[4] Huang, X., Islam, M. R., Akter, S., Ahmed, F., Kazami, E., Abu Serhan, H., Abd-alrazaq, A., & Yousefi, S. (2023). Artificial intelligence in glaucoma: Opportunities, challenges, and future directions. *Biomedical Engineering Online,* 22(1), article number 126. doi: 10.1186/s12938-023-01187-8.

[5] Hussain, S., Chua, J., Wong, D., Lo, J., Kadziauskiene, A., Asoklis, R., Barbastathis, G., Schmetterer, L., & Yong, L. (2023). Predicting glaucoma progression using deep learning framework. *Scientific Reports,* 13, article number 19960. doi: 10.1038/s41598-023-46253-2.

[6] Jin, Y., Liang, L., Li, J., Xu, K., Zhou, W., & Li, Y. (2024). Artificial intelligence and glaucoma: A lucid and comprehensive review. *Frontiers in Medicine,* 11, article number 4238. doi: 10.3389/fmed.2024.1423813.

[7] Kashyap, R., Nair, R., Gangadharan, S. M. P., Botto-Tobar, M., Farooq, S., & Rizwan, A. (2022). Glaucoma detection and classification using improved U-Net deep learning model. *Healthcare,* 10, article number 2497. doi: 10.3390/ healthcare10102497.

[8] Ling, X. C., Chen, H. S.-L., Yeh, P.-H., Cheng, Y.-C., Huang, C.-Y., Shen, S.-C., & Lee, Y.-S. (2025). Deep learning in glaucoma detection and progression prediction. *Biomedicines*, 13(2), article number 420. doi: 10.3390/ biomedicines13020420.

[9] Liu, Y., Wu, R., & Yang, A. (2023). Research on medical problems based on mathematical models. *Mathematics,* 11(13), article number 2842. doi: 10.3390/math11132842.

[10] Nocedal, J., & Wright, S. J. (2006). *Numerical optimization* (2nd ed.). Cham: Springer.

[11] Qian, X., *et al.* (2023). External validation of a deep learning detection system for glaucomatous optic neuropathy. *Eye,* 37, 3813-3818. doi: 10.1038/s41433-023-02622-9.

[12] Raju, M., Shanmugam, K.P., & Shyu, C.-R. (2023). Application of machine learning predictive models for early detection of glaucoma. *Applied Sciences,* 13(4), article number 2445. doi: 10.3390/app13042445.

[13] Shoukat, A., Akbar, S., Hassan, S.A., Iqbal, S., Mehmood, A., & Ilyas, Q.M. (2023). Automatic diagnosis of glaucoma from retinal images. *Diagnostics,* 13(10), article number 1738. doi: 10.3390/diagnostics13101738.

[14] Siesky, B., Harris, A., Verticchio Vercellin, A., Arciero, J., Fry, B., Eckert, G., Guidoboni, G., Oddone, F., & Antman, G. (2023). Heterogeneity of ocular hemodynamic biomarkers. *Journal of Clinical Medicine,* 12(4), article number 1287. doi: 10.3390/jcm12041287.

[15] Tarcoveanu, F., Leon, F., Curteanu, S., Chiselita, D., Bogdanici, C.M., & Anton, N. (2022). Classification algorithms used in predicting glaucoma progression. *Healthcare,* 10(10), article number 1831. doi: 10.3390/ healthcare10101831.

[16] Tonti, E., Tonti, S., Mancini, F., Bonini, C., Spadea, L., D'Esposito, F., Gagliano, C., Musa, M., & Zeppieri, M. (2024). Artificial intelligence and advanced technology in glaucoma. *Journal of Personalized Medicine,* 14(10), article number 1062. doi: 10.3390/jpm14101062.

[17] Tong, Y., Lu, W., Yu, Y., & Shen, Y. (2020). Application of machine learning in ophthalmic imaging. *Eye and Vision*, 7, article number 22. doi: 10.1186/s40662-020-00183-6.

[18] Moudgil, T., & Gupta, D. (2024). Advancements in antiglaucoma medications: A comprehensive review. *Tropical Ophthalmology*, 1(1), 12-16. doi: 10.4103/TOPH.TOPH_4_23.

[19] Vychuzhanin, V., Rudnichenko, N., Guzun, O., Zadorozhnyy, O., Korol, A., & Gritsuk, I. (2024). Artificial intelligence Integration in the diagnosis, prognosis and diabetic neovascular glaucoma treatment. *CEUR Workshop Proceedings,* 3790, 238-249.

[20] Wagner, I. V., Stewart, M. W., & Dorairaj, S. K. (2022). Updates on glaucoma diagnosis and management. *Mayo Clinic Proceedings: Innovations, Quality & Outcomes,* 6(6), 618-635. doi: 10.1016/j.mayocpiqo.2022.09.007.

[21] Zuo, H., Huang, B., He, J., Fang, L., & Huang, M. (2025). Machine learning approaches in high myopia. *Journal of Medical Internet Research,* 27, article number e57644. doi: 10.2196/57644.

Математичне моделювання стану ока при глаукомі: підходи до аналізу параметрів та їх взаємодія

Володимир Вичужанін

Доктор технічних наук, професор
Національний університет «Одеський політехнічний інститут»
65044, пр-т Шевченка, 1, м. Одеса, Україна
https://orcid.org/0000-0002-6302-1832

Олексій Вичужанін

Доктор філософії, асистент
Національний університет «Одеський політехнічний інститут»
65044, пр-т Шевченка, 1, м. Одеса, Україна
https://orcid.org/0000-0001-8779-2503

Ольга Гузун

Кандидат медичних наук
Інститут очних хвороб та тканинної терапії ім. В.П. Філатова
65044, бульв. Французький, 49/51, м. Одеса, Україна
https://orcid.org/0009-0003-6873-8503

Олег Задорожний

Доктор медичних наук
Інститут очних хвороб та тканинної терапії ім. В.П. Філатова
65044, бульв. Французький, 49/51, м. Одеса, Україна
https://orcid.org/0000-0003-0125-2456

Анотація. Математичне моделювання фізіологічних процесів є ключовим елементом інтелектуальних медичних систем, оскільки воно дозволяє глибше розуміти механізми захворювань та сприяє ранній діагностиці. У цьому дослідженні представлено аналітичну модель для оцінки стану ока, яка враховує ключові офтальмологічні параметри: внутрішньоочний тиск (BOT), коефіцієнт перфузії (Pperf), гостроту зору з найкращою корекцією (BCVA), індекс поля зору (VFI), товщину шару нервових волокон сітківки (RNFL) та площу нейроретинальної облямівки (Rim_area). Метою дослідження була розробка моделі, яка дозволяє точно оцінювати вплив нелінійних взаємодій між цими параметрами, підвищуючи точність діагностики та прогнозування прогресування глаукоми. Дослідження було спрямоване на визначення критичних порогових значень офтальмологічних показників для покращення прийняття клінічних рішень. Результати дослідження показали, що: застосування чисельної оптимізації (L-BFGS-B) та логарифмічно-експоненційних перетворень суттєво підвищує точність прогнозування ризику глаукоми; виявлено критичні порогові значення офтальмологічних параметрів, за якими можливе точніше визначення стадії глаукоми. Крім того, дослідження дає змогу оцінити взаємозв'язок між внутрішньоочним тиском та станом зорового нерва, що є критичним для прогнозування розвитку захворювання. Практична цінність дослідження полягає у можливості його інтеграції в медичні ІТ-системи для автоматизованого скринінгу глаукоми та моніторингу пацієнтів. Запропонований підхід може допомогти офтальмологам у прийнятті клінічних рішень, оптимізації стратегії лікування та запобіганні незворотній втраті зору. Адаптивність моделі також дозволяє використовувати її в телемедичних застосунках, що сприяє віддаленій діагностиці та постійному оцінюванню стану пацієнта

Ключові слова: аналітична модель; офтальмологічні параметри; оптимізація; медична діагностика; адаптивність

Journal homepage: https://itce.com.ua/en

Information Technologies and Computer Engineering **Vol. 22, No. 1. 2025**

Article's History: Received: 22.12.2024 Revised: 18.03.2025 Accepted: 24.04.2025

UDC 004.8:004.652.4 DOI: 10.63341/vitce/1.2025.20

Analysis of integrated real-time decision support systems based on neural networks and low-structured data

Mykola Demchyna*

PhD in Technical Sciences, Associate Professor
King Daniel University
76018, 35 Ye. Konovalets Str., Ivano-Frankivsk, Ukraine
https://orcid.org/0009-0002-9161-4843

Abstract. The study aimed to analyse and substantiate effective methods for analysing inefficiently structured data using neural networks to provide operational decision support in complex environments. The focus was on the use of artificial neural networks to analyse inefficiently structured data, such as sensor streams, to ensure efficiency, accuracy and adaptability in a dynamic environment. The research is aimed at creating innovative models and technologies that will improve the efficiency of management in complex situations, such as emergency response, process automation in critical industries and decision-making based on predictive analytics. The study investigated conceptual approaches to the development of integrated real-time decision support systems based on the analysis of poorly structured data using neural networks. The study proposed methods of adaptive learning that allow neural networks to process data efficiently in the face of constant changes. The research methodology included modelling a real-time architecture using a microservice approach and streaming data processing platforms such as Apache Kafka and Apache Flink. The study highlighted the role of neural networks in processing streaming data, in particular, convolutional networks for processing visual information, recurrent networks for sequence analysis, and transformers for multichannel analysis. Architectural solutions were developed that allow the processing of large amounts of data with minimal delays, ensuring the accuracy and adaptability of systems. The study presented approaches to the implementation of adaptive training of neural networks that minimise the risks of losing model relevance in a dynamic environment. The use of modern technologies, such as artificial neural networks, adaptive learning and integration with the Internet of Things, was used to create effective systems for rapid response to emergencies. The proposed methods help increase the efficiency of management in difficult conditions and create new prospects for innovation in various industries

Keywords: artificial intelligence; knowledge base; dynamic environment; network models; knowledge extraction

Introduction

The modern development of digital technologies poses new challenges for science and engineering related to the processing of large amounts of data in real-time. One of the key tasks that arise in this context is the development of integrated decision support systems capable of analysing unstructured data. Such data includes information flows from sensor networks, social media, video surveillance, telemetry and other sources where the data structure is incomplete, variable or non-standard. This issue is particularly relevant in the context of rapid response to emergencies, including natural disasters, man-made accidents or cyber threats. Effective decision-making in such circumstances requires not only fast data processing but also interpretation of data based on complex patterns and context. This is challenging due to the multifactorial nature of the problems, unpredictable developments, and limited time for analysis. Artificial neural networks (ANNs) offer significant prospects for solving these problems. They can learn from large amounts of data, identify hidden patterns and adapt to new conditions. However, the implementation of such systems faces several challenges. It is necessary to ensure sufficient performance of algorithms when working with real-time data streams. The processing of poorly structured data requires the development of effective methods for its preliminary preparation, cleaning and classification. In addition, there is the issue of trust in the decisions made by the system, which is especially important in critical scenarios.

Suggested Citation:
Demchyna, M. (2025). Analysis of integrated real-time decision support systems based on neural networks and low-structured data. *Information Technologies and Computer Engineering*, 22(1), 20-29. doi: 10.63341/vitce/1.2025.20

*Corresponding author

Modern research in the development of decision support systems was actively promoted using artificial neural networks to analyse inefficiently structured data. Significant progress in this area was made by A. Sherstinsky (2020), who proposed a model for sensor stream processing. The proposed approach, based on recurrent neural networks (RNNs), effectively detected anomalies in industrial systems in real-time, which was important for preventing accidents. S. Huang *et al.* (2020) analysed multi-module data from sensor networks. The authors argued that the integration of data from different sources significantly improves the prediction of crises, such as natural disasters. In turn, S.I. Nilima *et al.* (2024) studied the optimisation of deep models for resource-constrained environments, in IoT devices. Their results demonstrated that it is possible to significantly reduce data processing time without losing the quality of analysis. F. Fan *et al.* (2021) addressed the interpretation of artificial neural network solutions. The authors developed a methodology that explained the system's decision-making process, which increased user confidence, especially in critical scenarios such as medicine or defence. L.X. Yang & C.Y. Xiu (2023) addressed the adaptability of models by developing a technique Meanwhile, O. Trofymenko *et al*. (2024) studied the effectiveness of using neural networks and AI technologies in the form of intelligent agents to protect against cyberattacks and assess vulnerabilities and risks in the defence cyberspace. In particular, the authors noted the capabilities of AI to analyse large amounts of data in real time, identify patterns and make recommendations on how to address identified vulnerabilities. L.X. Yang & C.Y. Xiu (2023) addressed the adaptability of models by developing a technique that allowed updating the parameters of neural networks without the need for complete retraining, which was critical for working in dynamic environments. C. Wang *et al.* (2019) applied reinforcement learning methods to control autonomous drones during emergencies. Their approach proved to be effective in solving problems in challenging environments. The work was focused on creating systems that can operate in conditions of limited computing resources. The authors presented a method for optimising models based on convolutional neural networks (CNNs), which significantly reduced processing time without losing accuracy.

Despite significant progress in the field of decision support systems, there are still areas that require further research. There are a limited number of solutions that work efficiently with large-scale, unstructured data in real time, especially in scenarios with high information update rates. Explanatory ANN methods for complex multimodal data are not fully explored, which raises questions about the credibility of such systems.

The study aimed to analyse modern technologies for creating integrated real-time decision support systems capable of efficiently processing inefficiently structured data of various types. To do this, it was necessary to analyse methods for effectively integrating poorly structured data of various formats (text, sensor streams, video) into a single analytical system; to define an architecture that could handle large amounts of data in real-time with minimal delays.

Materials and Methods

The study analysed a combination of advanced technologies and methods for analysing low-structured data in real-time, with a particular focus on improving decision-making for critical systems such as emergency response or infrastructure monitoring. The main sources of data in this study were sensor networks and Internet of Things (IoT) devices, which generate significant amounts of low-structured information, including sensor data streams, text messages, and multimedia metadata. The Apache Kafka (n.d.) and Apache Flink (n.d.) platforms were used to process these data streams. They were essential for collecting and processing large amounts of data in real-time. Kafka was used for data integration, and Flink for continuous streaming data processing, which ensured its constant broadcast and quick analysis. These platforms can be used for the seamless integration of data from different sources, which is critical for monitoring critical infrastructure or emergency response systems.

The analysis of the poorly structured data was carried out using artificial neural networks (ANNs), in particular convolutional neural networks (CNNs) and recurrent neural networks (RNNs). CNNs were used to analyse video data from surveillance cameras, allowing the system to detect key features such as motion, object recognition, and context changes. At the same time, RNNs and transformers were used to process time series of sensor data and textual information, allowing the system to detect temporal patterns and make predictions based on past data. Adaptive learning methods were used to adapt to dynamic environments where data is constantly changing. Approaches such as unsupervised learning and online learning were used to ensure that the models could be continuously updated and improved without the need for complete retraining. A review of the key aspects and technologies used to develop real-time decision support systems that analyse unstructured data was conducted.

Multi-channel models were studied to analyse different types of information simultaneously. Integration was conducted through the Kafka and Flink platforms, which ensured the efficient combination of different data streams, including video, sensor data, and text messages. Multi-channel models combined CNN+RNN architectures or used transformers to process all data simultaneously, revealing hidden correlations and patterns. To reduce delays in re-accessing data, a data caching technique was applied. This made it possible to speed up access to the results of data processing, which is critical for decision-making in stressful situations. Adaptive learning algorithms, including stochastic gradient optimisation methods, were explored to allow networks to respond quickly to changes in conditions. These algorithms provided the ability to continuously update model weights based on new data coming in real-time.

Results

In the modern world, where the amount of data is growing exponentially, the concept of unstructured data is becoming increasingly relevant. This data – is a "bridge" between structured relational database tables and chaotic arrays of unstructured information. They contain a potential that can only be realised with a deep understanding of their nature. Loosely structured data can be compared to a large library, where books are not arranged according to a standard classification system, but each has tags or a short description. Such data includes JSON files, log entries, sensor streams, or media file metadata. They contain information that has a certain structure, but it is often incomplete, flexible, or even unstable. Another important feature is the dynamism of loosely structured data. Data can change depending on software updates, the introduction of new features, or changes in user behaviour. Thus, the system must be able to adapt to these changes while maintaining efficiency.

At the same time, the benefits come with challenges. Traditional approaches to data analysis are often not suitable for working with unstructured data. The complexity of integrating heterogeneous sources, the large number of missing values, and the constant change in structure create additional difficulties. However, it is these challenges that drive innovation (Hariri *et al.*, 2019). Artificial neural networks, due to their flexibility, create new opportunities for working with such data. They allow not only finding patterns but also explaining decisions, which is critical in complex scenarios such as emergency response.

Unstructured data is a challenge but also an opportunity for modern systems. They allow for the development of integrated solutions that can not only analyse information but also do so in real-time, providing a new level of efficiency in the decision-making process (Raptis *et al.*, 2019). Therefore, their research is the key to developing innovative approaches that can cope with modern requirements for processing heterogeneous information. For instance, the data stream from sensors in smart city systems can include numerical temperature readings, noise levels, images from surveillance cameras, and text reports from operators. Traditional methods of analysis require a clear structuring of information, but integrating such heterogeneous sources is a challenging task. In this context, neural networks are highly effective due to their ability to adapt to processing data with a complex and heterogeneous structure.

A key factor in modelling nonlinear relationships between data is choosing the right mathematical or statistical approach that can accurately reflect these relationships. For instance, when analysing streams from surveillance cameras, convolutional neural networks (CNNs) can be used to identify key features of images: objects, movement, and context changes (Ullah *et al.*, 2019). Recurrent neural networks (RNNs) can be used to remember the previous context and make predictions based on it (Dhruv & Naskar, 2020). Neural networks also allow the integration of different types of data. Thanks to multi-channel models such as combined CNN+RNN architectures or transformers, neural networks can process all these streams simultaneously, revealing correlations and patterns that were previously hidden. However, their value lies not only in the ability to analyse. Neural networks change the perception of poorly structured data, turning it from a problem into a source of competitive advantage. For example, emergency response systems can process huge amounts of information in real time, offering clear recommendations (Hancock & Khoshgoftaar, 2020). Such systems not only reduce decision-making time but also minimise errors that can occur due to human error.

Real-time architectural solutions are the basis for creating efficient and fast systems that can process large amounts of data with minimal delays. One of the most popular approaches is microservice architecture, where the system is divided into small, independent services, each responsible for a specific task. This enables flexible scaling of the system, updates of individual components without affecting the rest of the system, and increases resilience and reliability (Abirami & Chitra, 2020). To integrate data from various sources, such as sensors or video cameras, event-based platforms are used to respond to changes in real-time, activating the necessary actions without unnecessary delays. Another important element is the use of high-performance real-time data processing platforms such as Apache Kafka or Apache Flink. They can efficiently process data streams, ensuring their continuous flow and fast processing. Such platforms allow not only fast processing of information but also guarantee its reliable storage and the possibility of recovery in case of a system failure. In addition, to ensure high performance, parallel computing and distributed systems are often used to process large amounts of data using multiple servers, reducing the load on individual components and ensuring system resilience. The last important aspect is the optimisation of data caching, which reduces delays in re-accessing frequently used information. The use of the cache significantly speeds up access to processing results and allows for quick decision-making in critical situations. All of this together creates an architecture that can operate efficiently in real-time, providing high data processing speed, fault tolerance, and reliability in difficult conditions (Mehmood & Anees, 2020).

Adaptive training of neural networks in dynamic environments is a critical aspect for systems that process streaming data and need to constantly respond to changes in the environment. In such environments, traditional training methods that use static data sets may not be effective. Dynamic environments are typically characterised by constantly changing parameters, which require neural networks to be able to adapt to new conditions without the need for a complete retraining process from scratch (Liu *et al.*, 2020). This is especially important for applications that process data in real-time, such as emergency response systems, security monitoring, or financial forecasts. Adaptive neural network training involves continuously adjusting the model based on new incoming data. This may include using updated parameters or new architectures to provide

more accurate predictions and analysis. This is often done using approaches such as unsupervised learning, where the network detects new patterns in the data on its own, as well as methods that allow networks to "forget" outdated or irrelevant data by focusing on the most relevant information (Han *et al.*, 2021). This avoids the effect of overfitting and maintains the accuracy of the model in the face of changing input data.

One of the most effective ways of adaptive learning is to use algorithms that incorporate real-time changes in the learning rate. For instance, methods based on stochastic gradient descent can change parameters during the learning process, depending on how quickly the input data changes (Kabudi *et al.*, 2021). This allows the network to respond quickly to changes in the environment, guaranteeing accuracy even in the event of unpredictable events. In addition, to operate effectively in dynamic environments, it is important to strike a balance between the speed of adaptation and the stability of the model. Too fast adaptation can lead to excessive fluctuations in the results, while too slow adaptation can lead to the loss of information relevance. For a neural network to work effectively in a dynamic environment, it is also necessary to apply approaches to data selection. Sampling new data that is most relevant to the current situation allows the model to remain effective even under changing conditions. This may include active learning methods, where the network independently determines what data, it needs to obtain to improve its performance, or the use of multi-channel models that allow analysis of information from different sources simultaneously. Thus, adaptive training of neural networks allows not only to maintain a high level of accuracy in real-time but also to make predictions based on the most relevant and updated information coming from the environment.

Table 1 provides an overview of the key aspects and technologies used to develop real-time decision support systems that analyse unstructured data. It summarises the most important components of such systems, including the integration of neural networks, the use of streaming data platforms, and adaptive learning for dynamic environments. These approaches allow for the rapid processing of large volumes of data in real-time, which is critical for effective emergency response and decision support in complex environments.

Table 1. Key aspects and technologies for real-time decision support systems based on neural networks and analysis of weakly structured data

A key area of focus	Description	Technical aspects and methods	Application examples
Inefficiently structured data analysis	Processing data that does not have a clear structure (e.g., sensor streams, text messages).	Artificial neural networks, data stream processing, and classification using deep networks.	Environmental monitoring, security, video analysis
Real-time neural networks	Use of neural networks to process data in real-time, which allows for immediate response to changes.	Recurrent neural networks (RNNs), deep neural networks, and ongoing learning.	Emergency response and traffic management
Real-time data integration	Collecting and processing data from various sources to provide up-to-date information for decision-making.	Streaming data processing platforms (Apache Kafka, Apache Flink), integration via API	Monitoring the parameters of critical infrastructure facilities
Response to emergency events	Prompt identification and response to events requiring immediate intervention.	Neural networks and active learning methods, event prediction through anomaly detection algorithms	Security systems, monitoring for responding to natural disasters
Decision support systems	Designing interfaces and mechanisms for quick decision-making based on data analysis.	Interfaces for visualising results, algorithms for classification and forecasting.	Automated emergency management systems, risk analysis
Adaptive learning and model updates	Continuous improvement of models based on new data to maintain the relevance and accuracy of forecasts.	Methods of updating neural networks (online training), using new data for adaptation.	Monitoring systems with automatic model updates
System reliability and resilience	Ensure uninterrupted operation even in the face of disruptions or unforeseen changes in data.	Backup, load balancing, disaster recovery.	Critical infrastructures, financial systems

Source: compiled by the author based on F. Gurcan & M. Berigel (2018), A.N. Navaz *et al*. (2019), S. Ashraf *et al*. (2022)

The application of the described technologies increases the accuracy and speed of decision-making, which allows the creation of effective systems to support real-time management decisions, such as monitoring the condition of critical facilities, managing traffic flows or responding to emergencies. In addition, neural networks provide adaptability and self-optimisation in the face of changing input data, which is important in situations requiring immediate action. The use of technologies for streaming data processing and adaptive learning allows for high accuracy of decisions even in unstable and unpredictable scenarios, making such systems highly beneficial for critical infrastructures and situations where every second counts (Semenenko *et al.*, 2024).

Integrated real-time decision support systems that analyse unstructured data can significantly improve management efficiency in complex situations. The use of neural networks to process real-time data streams allows the system to adapt to changing conditions, which is key to responding quickly to emergencies. The integration of streaming data processing technologies allows systems to operate efficiently under high loads, processing large amounts of information without significant delays. This ensures not only prompt decision-making but also their accuracy, as

systems can constantly adapt to new data. Adaptive learning technologies allow neural networks to dynamically change their strategies depending on changes in input data, which is critical in areas such as infrastructure monitoring, transport management, or emergency response.

As a result, the use of such systems allows for more efficient and accurate decision-making in real-time, which is extremely important in an environment where every second counts for safety and effective management. Therefore, such technologies have great potential for use in various areas where it is necessary to respond quickly to changes and ensure reliable management in difficult conditions. The implementation of integrated real-time decision support systems is based on a combination of modern technologies and data processing methods. The basis of such systems is the analysis of unstructured data generated in large volumes and requiring fast processing to ensure timely decisions. This applies to data received from sensor networks, streaming platforms, and IoT devices.

Table 2 presents the main technologies and methods for designing integrated real-time decision support systems. It also shows the main technologies and methods used to design such systems. Key components are included, such as big data platforms, artificial intelligence algorithms such as neural networks, and models that support adaptive learning in dynamic environments. Such solutions are critical for industries operating under conditions of high uncertainty, including crisis management, transport, medicine, and energy. The presented methods help improve the accuracy, reliability and speed of data analysis.

Table 2. Basic technologies and methods for designing integrated real-time decision support systems

Technology/Method	Description	Application examples	Advantages	Challenges
Platforms for streaming data processing	Used to process large volumes of data coming in real-time. Main platforms: Apache Kafka, Apache Flink, Spark Streaming.	Monitoring the state of critical infrastructure, managing traffic flows	High processing speed, scalability	Difficulty of integration with other systems, need for high computing resources.
Neural networks for forecasting	Neural networks (especially LSTM, and GRU) are used to predict and classify data in real time.	Forecasting events in security systems, forecasting demand in retail	Improved forecast accuracy, ability to work with big data	The need for large amounts of training data, the risk of overtraining
Adaptive learning in real-time	Neural network models are constantly updated based on new data. This allows systems to adapt to changes in the environment.	Responding to changes in the security environment, adapting to changes in user behaviour	Fast adaptation to new conditions, reduced manual configuration costs	The need for constant evaluation and correction of models, the possibility of noise data
Integration with IoT (Internet of Things)	A combination of sensors, IoT devices and neural networks to provide real-time data collection.	Patient health monitoring, industrial process automation	Improved data accuracy, efficient resource management	Compatibility issues between different devices, data security issues
Response to emergency events	Use of systems for prompt decision-making during emergencies based on data from various sources.	Responding to natural disasters and accidents at industrial facilities	Speed of response, minimisation of human errors	High requirements for data accuracy, limited resources for real-time data processing

Source: compiled by the author based on M. Mohammadi *et al*. (2018), Y. Yan & H. Yang (2024)

The methods and technologies presented in the table demonstrate how an integrated approach to processing unstructured data allows for the creation of effective decision support systems. The use of streaming processing platforms such as Apache Kafka or Apache Flink ensures the speed of processing large amounts of information, while neural networks, especially LSTM and CNN, can accurately identify key patterns in the data. Adaptive learning, in turn, ensures that systems can quickly adapt to changes in the environment or changes in data structure.

However, the implementation of such systems requires a solution to several challenges. These include limited computing resources for training complex models in real-time, ensuring data security and privacy in cloud infrastructures, and the difficulty of integrating existing solutions with new technologies. These challenges require further research aimed at creating optimised, secure and scalable architectures that meet the needs of modern systems. At the same time, advances in technology are enabling the capabilities of such systems to be expanded, for example, by integrating them with quantum computing for even faster analysis and decision-making. This opens new perspectives for application in complex areas such as urban infrastructure management, automated transport systems, and environmental monitoring. Thus, the presented methods and approaches form the basis for the further development of this innovative field.

The conceptual development of a system for real-time analysis of unstructured data is based on the integration of modern data processing technologies, artificial intelligence and adaptive machine learning models. The main components of such a system are a data collection module, a real-time processing unit, an analytical module using neural networks, and an interactive interface for visualising results and supporting decision-making. Data sources, which include sensor networks and IoT devices, are integrated through standard protocols such as MQTT or APIs. This ensures a constant flow of data into the system. The streams of information, which can include text messages, video from surveillance cameras, and signals

from sensors, are processed by streaming platforms such as Apache Kafka, Flink, or Spark Streaming. At this stage, the data is normalised, gaps are filled in, and it is brought to a common format (Mohan & Thyagarajan, 2023).

The key element of the system is the analytical module, where neural networks are used for multi-channel data processing. For instance, recurrent neural networks (LSTM, GRU) are used to analyse text messages, while convolutional neural networks process video to detect anomalies or dangerous events. In complex scenarios that require the integration of different types of data, combined architectures are used to process text, visual and sensory streams simultaneously. The system has an adaptive learning capability that allows the modelling of new patterns in the data as it is acquired. This ensures dynamic adaptation to changes, such as software updates, changes in user behaviour, or the emergence of new functionalities. As a result, the system can operate in conditions where traditional methods are ineffective (Haidur *et al.*, 2023). The analysis results are transferred to an interactive interface that allows operators to visualise data, get explanations for decisions made, and intervene promptly if necessary. If anomalies or dangerous situations are detected, the system generates recommendations for the relevant services, speeding up the response process. The proposed concept is suitable for real-time scenarios, such as security monitoring in smart cities. The system integrates data from various sources, such as video surveillance and sensor sensors, analyses it to identify critical situations and facilitates timely decision-making. The implementation of this concept will ensure high performance, adaptability and functionality in today's dynamic environments.

Discussion

Analysis of unstructured data and integrating it into real-time decision support systems has become an important aspect of modern information processing. With the rapid growth of data volumes, especially from diverse and dynamic sources such as sensor streams, social media and multimedia files, the challenge is not only to collect and store this information but also to process it efficiently and quickly. Unstructured data is inherently flexible and often incomplete, which poses significant challenges for traditional data analytics methods. However, they also have enormous potential for the development of innovative real-time systems that can adapt to rapidly changing conditions.

One of the main advantages of unstructured data is its ability to integrate diverse sources of information. In systems such as smart city monitoring or emergency response systems, data can come in a variety of formats – from text alerts and sensor measurements to video from security cameras. Neural networks, convolutional and recurrent models, have proven to be extremely effective in processing such complex multi-format data streams. Convolutional neural networks (CNNs) are suitable for image processing, which allows the analysis of video streams from surveillance cameras. Recurrent Neural Networks (RNNs) or Transformers, on the other hand, are effective with sequences of data, such as sensor measurements or text logs, remembering previous information and predicting future states. The research by B.A. Hammou *et al.* (2019) focused on the use of convolutional and recurrent neural networks for real-time processing of video and text data. The authors emphasised the role of adaptive learning, in particular online learning, to adjust models while processing current data. This coincides with current findings on the need for an adaptive approach to processing dynamic data, but this study has focused on the impact of infrastructure platforms on data flow.

In addition, the ability of neural networks to model non-linear relationships and adapt to changes in data significantly increases their effectiveness. In environments where data is constantly changing – such as real-time monitoring systems – neural networks can dynamically adjust their models based on new input data, which allows for accurate and timely decision-making. A. Novak *et al.* (2021) investigated the impact of neural networks on decision-making accuracy in environments were data changes in real-time. Scientists emphasised that the adaptability of neural networks can quickly adjust models based on new data, which is critical for efficient processing in situations such as emergencies. However, the study addressed technologies for predicting events based on historical data, not real-time.

Adaptive learning algorithms are key, allowing systems to quickly update their models in real-time based on new data. Traditional machine learning methods that rely on static datasets have proven to be insufficient in such environments where parameters are constantly changing. Y. Wang & S. Zou (2021) analysed the use of reinforcement learning to adapt models to a changing environment. The study considered the use of methods such as reinforcement learning to optimise models in the face of unstable data. In contrast, the current study focuses more on the need for adaptive learning due to real-time changes, without the need to rebuild the entire model.

In terms of architectural solutions, implementing real-time decision support systems requires a reliable and scalable infrastructure. Microservice architecture, for instance, provides flexibility and reliability by allowing individual components to operate independently while integrating into a single system. Streaming data processing platforms, such as Apache Kafka and Apache Flink, are essential for enabling the fast processing of large amounts of data in real-time while minimising latency. In addition, distributed computing systems and parallel processing allow for efficient analysis of large amounts of data, ensuring high performance even under heavy loads. B.G. Deepthi *et al.* (2023) explored the use of the Apache Flink platform for fast real-time data processing and its integration with neural networks to ensure accuracy and speed of decision-making. They also highlighted the importance of scalability. This is consistent with the findings of this paper, which emphasised the importance of an efficient infrastructure for processing large amounts of real-time data.

One of the most popular approaches to developing real-time systems is microservice architecture. It allows load balancing and provides flexibility and scalability. G. Ortiz *et al.* (2021) investigated the use of microservice architecture to create high-performance systems capable of processing large amounts of data in real-time. Scientists highlighted the advantages of this approach, such as flexibility, scalability, and load balancing between services to reduce latency. They emphasised the use of microservice architecture to ensure real-time efficiency. This coincides with the current findings, which emphasised the ability of such an architecture to handle large amounts of data and provide system flexibility. Furthermore, G. Ortiz *et al.* (2019) investigated the use of microservice architecture to build efficient systems capable of processing large amounts of data in real-time. The authors emphasised the advantages of this architecture, such as scalability, flexibility, and the ability to reduce delays due to optimised load balancing between services. Scientists emphasised the ability to distribute the load between services to reduce latency, while current results have focused more on the use of microservice architecture to provide flexibility and handle large amounts of data. M. Raparthi *et al.* (2021) analysed the potential of quantum computing to accelerate data analysis and decision-making. While this is an area that holds great promise, the current study did not focus on quantum technologies, but rather on the use of existing streaming platforms and neural network technologies to provide data processing speed.

The integration of neural networks with streaming data processing platforms has revolutionised many industries, from health and safety to traffic management and emergency response. However, there are several challenges, including the need for powerful computing resources to train complex models in real time and ensure data security. D. Kavitha & S. Ravikumar (2020) highlighted the importance of integrating neural networks with streaming data platforms to achieve real change in areas such as healthcare. The authors also highlighted the challenges, including the need for substantial computing resources and security challenges, associated with using these technologies in real time. However, scientists also emphasised the importance of new technologies, such as quantum computing, for the further development and optimisation of decision-making processes. Thus, the research and development of systems that analyse unstructured data in real time have become an important step towards improving the efficiency of decision-making in critical environments. Neural networks and adaptive learning provide the flexibility needed to handle dynamic data streams, and robust data platforms ensure that systems operate efficiently even under high loads.

These technologies have transformed industries such as emergency management and infrastructure monitoring. G. Huang *et al.* (2006) investigated the integration of neural networks for real-time analysis of unstructured data in critical environments, including emergency management and infrastructure monitoring. The study analysed adaptive learning, which allows systems to respond quickly to data changes, as well as problems arising from high loads on data processing platforms. Although both the current findings and the author's research focused on the integration of neural networks and adaptive learning, there are important differences between them. The current findings put more emphasis on the importance of real-time data processing with many variables and complex conditions, where flexible streaming data processing platforms such as Apache Kafka or Flink are used.

The integration of new technologies, such as streaming data platforms or quantum computing, still faces interoperability and data security issues. O. Petit *et al.* (2018) highlighted the difficulties encountered when integrating new technologies into existing systems. Among the main problems, the author highlighted the need for compatibility between different platforms, ensuring data security, and the difficulty of adapting new solutions to the specific requirements of certain industries. The authors, similarly, to the current study, emphasised the difficulties in integrating new technologies such as quantum computing and streaming data processing, including interoperability and security issues. Integrating emerging technologies such as neural networks and microservice architecture is important to optimise real-time data processing. Increasing the flexibility and adaptability of such systems helps to improve the accuracy of decision-making, especially in critical environments such as emergencies or infrastructure monitoring. However, despite the many benefits, there are significant challenges, including the need for high computing power and data security.

Conclusions

The current study examined methods for analysing sparsely structured data for real-time decision-making. The study has shown that unstructured data, due to its dynamism, can be effectively integrated into real-time monitoring systems. The use of neural networks allows the processing of variable data streams that arise due to software updates, changes in user behaviour or the addition of new features, ensuring that systems adapt to these changes. It was confirmed that neural networks are an effective tool for identifying hidden patterns in data and explaining decisions, which is important in critical situations such as emergencies. In addition, significant progress has been made in addressing the challenges of integrating various sources of unstructured data, including text messages, video streams and sensor signals, which has led to increased speed and accuracy of decision-making.

The study also highlighted the importance of using microservices architectures and streaming data processing platforms such as Apache Kafka and Apache Flink to ensure stable and efficient operation of real-time systems. This allows not only to ensure high data processing speeds but also to maintain their reliability and fault tolerance. By integrating adaptive learning into neural networks, systems can independently update their models in the face of constant change, which is critical for industries

where reaction time is crucial. However, for the successful implementation of such systems, it is necessary to solve the problems associated with limited computing resources and data security. As a result, the implementation of integrated real-time decision support systems based on the analysis of unstructured data opens new prospects for the development of various industries, including critical infrastructure management, emergency response, automated transport management systems, and environmental monitoring. Further research could focus on optimising data processing methods and developing adaptive algorithms that allow for a more efficient response to changes in dynamic environments.

Acknowledgements
None.

Funding
The study received no funding.

Conflict of Interest
None.

References

[1] Abirami, S., & Chitra, P. (2020). Energy-efficient edge based real-time healthcare support system. *Advances in Computers*, 117(1), 339-368. doi: 10.1016/bs.adcom.2019.09.007.
[2] Apache Flink. (n.d.). Retrieved from https://flink.apache.org/.
[3] Apache Kafka. (n.d.). Retrieved from https://kafka.apache.org/.
[4] Ashraf, S., Afify, Y.M., & Ismail, R. (2022). Big data for real-time processing on streaming data: state-of-the-art and future challenges. In *2022 International conference on electrical, computer, communications and mechatronics engineering* (pp. 1-8). Maldives: IEEE. doi: 10.1109/iceccme55909.2022.9987770.
[5] Deepthi, B.G., Rani, K.S., Krishna, P.V., & Saritha, V. (2023). An efficient architecture for processing real-time traffic data streams using Apache Flink. *Multimedia Tools and Applications*, 83(13), 37369-37385. doi: 10.1007/s11042-023-17151-6.
[6] Dhruv, P., & Naskar, S. (2020). Image classification using convolutional neural network (CNN) and recurrent neural network (RNN): A review. In D. Swain, P. Pattnaik & P. Gupta (Eds.), *Machine learning and information processing* (pp. 367-381). Singapore: Springer. doi: 10.1007/978-981-15-1884-3_34.
[7] Fan, F., Xiong, J., Li, M., & Wang, G. (2021). On interpretability of artificial neural networks: A survey. *IEEE Transactions on Radiation and Plasma Medical Sciences*, 5(6), 741-760. doi: 10.1109/trpms.2021.3066428.
[8] Gurcan, F., & Berigel, M. (2018). Real-time processing of big data streams: Lifecycle, tools, tasks, and challenges. In *2018 2nd international symposium on multidisciplinary studies and innovative technologies* (pp. 1-6). Ankara: IEEE. doi: 10.1109/ismsit.2018.8567061.
[9] Haidur, H.I., Gakhov, S.O., & Bryhynets, A.A. (2023). Detection of network anomalies with neural networks algorithms. *Telecommunication and Information Technologies*, 1(78), 61-73. doi: 10.31673/2412-4338.2023.016173.
[10] Hammou, B.A., Lahcen, A.A., & Mouline, S. (2019). Towards a real-time processing framework based on improved distributed recurrent neural network variants with fastText for social big data analytics. *Information Processing & Management*, 57(1), article number 102122. doi: 10.1016/j.ipm.2019.102122.
[11] Han, Y., Huang, G., Song, S., Yang, L., Wang, H., & Wang, Y. (2021). Dynamic neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11), 7436-7456. doi: 10.1109/tpami.2021.3117837.
[12] Hancock, J.T., & Khoshgoftaar, T.M. (2020). Survey on categorical data for neural networks. *Journal of Big Data*, 7, article number 28. doi: 10.1186/s40537-020-00305-w.
[13] Hariri, R.H., Fredericks, E.M., & Bowers, K.M. (2019). Uncertainty in big data analytics: Survey, opportunities, and challenges. *Journal of Big Data*, 6(1), article number 44. doi: 10.1186/s40537-019-0206-3.
[14] Huang, G., Zhu, Q., & Siew, C. (2006). Real-time learning capability of neural networks. *IEEE Transactions on Neural Networks*, 17(4), 863-878. doi: 10.1109/tnn.2006.875974.
[15] Huang, S., Lin, C., Zhou, K., Yao, Y., Lu, H., & Zhu, F. (2020). Identifying physical-layer attacks for IoT security: An automatic modulation classification approach using multi-module fusion neural network. *Physical Communication*, 43, article number 101180. doi: 10.1016/j.phycom.2020.101180.
[16] Kabudi, T., Pappas, I., & Olsen, D.H. (2021). AI-enabled adaptive learning systems: A systematic mapping of the literature. *Computers and Education Artificial Intelligence*, 2, article number 100017. doi: 10.1016/j.caeai.2021.100017.
[17] Kavitha, D., & Ravikumar, S. (2020). IOT and context-aware learning-based optimal neural network model for real-time health monitoring. *Transactions on Emerging Telecommunications Technologies*, 32(1), article number e4132. doi: 10.1002/ett.4132.
[18] Liu, Y., He, Q., Zheng, D., Xia, X., Chen, F., & Zhang, B. (2020). Data caching optimization in the edge computing environment. *IEEE Transactions on Services Computing*, 15(4), 2074-2085. doi: 10.1109/tsc.2020.3032724.
[19] Mehmood, E., & Anees, T. (2020). Challenges and solutions for processing real-time big data stream: A systematic literature review. *IEEE Access*, 8, 119123-119143. doi: 10.1109/access.2020.3005268.
[20] Mohammadi, M., Al-Fuqaha, A., Sorour, S., & Guizani, M. 2018. Deep learning for IoT big data and streaming analytics: A survey. *IEEE Communications Surveys & Tutorials*, 20(4), 2923-2960. doi: 10.1109/COMST.2018.2844341.

[21] Mohan, D., & Thyagarajan, K. (2023). *A side-by-side comparison of Apache Spark and Apache Flink for common streaming use cases*. Retrieved from https://aws.amazon.com/ru/blogs/big-data/a-side-by-side-comparison-of-apache-spark-and-apache-flink-for-common-streaming-use-cases/?utm_source.

[22] Navaz, A.N., Harous, S., Serhani, M.A., & Taleb, I. (2019). Real-time data streaming algorithms and processing technologies: A survey. In *2019 International conference on computational intelligence and knowledge economy* (pp. 246-250). Dubai: IEEE. doi: 10.1109/ICCIKE47802.2019.9004318.

[23] Nilima, S.I., Bhuyan, M.K., Kamruzzaman, M., Akter, J., Hasan, R., & Johora, F.T. (2024). Optimizing resource management for IoT devices in constrained environments. *Journal of Computer and Communications*, 12(08), 81-98. doi.org/10.4236/jcc.2024.128005.

[24] Novak, A., Bennett, D., & Kliestik, T. (2021). Product decision-making information systems, real-time sensor networks, and artificial intelligence-driven big data analytics in sustainable Industry 4.0. *Economics, Management, and Financial Markets*, 16(2), 62-72. doi: 10.22381/emfm16220213.

[25] Ortiz, G., Boubeta-Puig, J., Criado, J., Corral-Plaza, D., Garcia-De-Prado, A., Medina-Bulo, I., & Iribarne, L. (2021). A microservice architecture for real-time IoT data processing: A reusable Web of things approach for smart ports. *Computer Standards & Interfaces*, 81, article number 103604. doi: 10.1016/j.csi.2021.103604.

[26] Ortiz, G., Caravaca, J.A., Garcia-De-Prado, A., De La O, F.C., & Boubeta-Puig, J. (2019). Real-time context-aware microservice architecture for predictive analytics and smart decision-making. *IEEE Access*, 7, 183177-183194. doi: 10.1109/access.2019.2960516.

[27] Petit, O., Velasco, C., & Spence, C. (2018). Digital sensory marketing: Integrating new technologies into multisensory online experience. *Journal of Interactive Marketing*, 45(1), 42-61. doi: 10.1016/j.intmar.2018.07.004.

[28] Raparthi, M. *et al*. (2021). Real-time AI decision making in IoT with quantum computing: Investigating & exploring the development and implementation of quantum-supported AI inference systems for IoT applications. *Internet of Things and Edge Computing Journal*, 1(1), 18-27.

[29] Raptis, T.P., Passarella, A., & Conti, M. (2019). Data management in Industry 4.0: State of the art and open challenges. *IEEE Access*, 7, 97052-97093. doi: 10.1109/access.2019.2929296.

[30] Semenenko, O., Nozdrachov, O., Chernyshova, I., Melnychenko, A., & Momot, D. (2024). Innovative technologies to improve energy efficiency and security of military facilities. Machinery & Energetics, 15(4), 147-156. doi: 10.31548/machinery/4.2024.147.

[31] Sherstinsky, A. (2020). Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. *Physica D: Nonlinear Phenomena*, 404, article number 132306. doi: 10.1016/j.physd.2019.132306.

[32] Trofymenko, O., Sokolov, A., Chykunov, P., Akhmametieva, H., & Manakov, S. (2024). AI in the military cyber domain. *Technologies and Engineering*, 25(4), 85-92. doi: 10.30857/2786-5371.2024.4.8.

[33] Ullah, A., Muhammad, K., Haq, I.U., & Baik, S.W. (2019). Action recognition using optimized deep autoencoder and CNN for surveillance data streams of non-stationary environments. *Future Generation Computer Systems*, 96, 386-397. doi: 10.1016/j.future.2019.01.029.

[34] Wang, C., Wang, J., Shen, Y., & Zhang, X. (2019). Autonomous navigation of UAVs in large-scale complex environments: A deep reinforcement learning approach. *IEEE Transactions on Vehicular Technology*, 68(3), 2124-2136. doi: 10.1109/tvt.2018.2890773.

[35] Wang, Y., & Zou, S. (2021). Online robust reinforcement learning with model uncertainty. *Advances in Neural Information Processing Systems*, 34.

[36] Wang, Y., Zhang, H., & Zhang, G. (2019). cPSO-CNN: An efficient PSO-based algorithm for fine-tuning hyper-parameters of convolutional neural networks. *Swarm and Evolutionary Computation*, 49, 114-123. doi: 10.1016/j.swevo.2019.06.002.

[37] Yan, Y., & Yang, H. (2024). Big data analysis and decision support system based on deep learning. *Computer-Aided Design & Applications*, 21(S13), 62-74. doi: 10.14733/cadaps.2024.s13.62-74.

[38] Yang, L.X., & Xiu, C.Y. (2023). Characteristics and techniques for adaptive models for behavior prediction in dynamic networks. *International Journal of Responsible Artificial Intelligence*, 13(7), 13-21.

Аналіз інтегрованих систем підтримки прийняття рішень у реальному часі на основі нейронних мереж та слабоструктурованих даних

Микола Демчина

Кандидат технічних наук, доцент
Університет Короля Данила
76018, вул. Є. Коновальця, 35, м. Івано-Франківськ, Україна
https://orcid.org/0009-0002-9161-4843

Анотація. Метою цієї роботи було дослідження та обґрунтування ефективних методів аналізу слабоструктурованих даних з використанням нейронних мереж для забезпечення оперативної підтримки прийняття рішень у складних середовищах. Основну увагу приділено використанню штучних нейронних мереж для аналізу слабоструктурованих даних, таких як сенсорні потоки, для забезпечення оперативності, точності та адаптивності в умовах динамічного середовища. Дослідження спрямоване на створення інноваційних моделей і технологій, які дозволять підвищити ефективність управління у складних ситуаціях, таких як реагування на надзвичайні події, автоматизація процесів у критичних галузях і прийняття рішень на основі прогнозної аналітики. У роботі досліджено концептуальні підходи до розробки інтегрованих систем підтримки прийняття рішень у реальному часі, які базуються на аналізі слабоструктурованих даних за допомогою нейронних мереж. Запропоновано методи адаптивного навчання, що дають змогу нейронним мережам ефективно обробляти дані в умовах постійних змін. Методологія дослідження включала моделювання архітектури реального часу з використанням мікросервісного підходу та платформ для потокової обробки даних, таких як Apache Kafka і Apache Flink. Висвітлено роль нейронних мереж у роботі з потоковими даними, зокрема згорткових мереж для обробки візуальної інформації, рекурентних мереж для аналізу послідовностей і трансформерів для багатоканального аналізу. Розроблено архітектурні рішення, які дозволяють обробляти великі обсяги даних із мінімальними затримками, забезпечуючи точність і адаптивність систем. Представлено підходи до реалізації адаптивного навчання нейронних мереж, що мінімізують ризики втрати релевантності моделі в динамічному середовищі. Використання сучасних технологій, таких як штучні нейронні мережі, адаптивне навчання та інтеграція з інтернетом речей, дозволяє створювати ефективні системи для оперативного реагування на надзвичайні події. Запропоновані методи сприяють підвищенню ефективності управління у складних умовах і відкривають нові перспективи для інновацій у різних галузях

Ключові слова: штучний інтелект; база знань; динамічне середовище; мережеві моделі; видобування знань

Journal homepage: https://itce.com.ua/en

Information Technologies and Computer Engineering **Vol. 22, No. 1. 2025**

Article's History: Received: 15.12.2024 Revised: 06.03.2025 Accepted: 24.04.2025

UDC 621.39:004.738.5 DOI: 10.63341/vitce/1.2025.30

Analysis of the impact of cross-platform behaviour on recommendation quality

Anton Pakula

Postgraduate Student,
Vinnytsia National Technical University
21021, 95 Khmelnytske Shose Str., Vinnytsia, Ukraine
https://orcid.org/0009-0002-5388-5386

Volodymyr Garmash

PhD in Technical Sciences, Associate Professor
Vinnytsia National Technical University
21021, 95 Khmelnytske Shose Str., Vinnytsia, Ukraine
https://orcid.org/0009-0007-1861-8772

Abstract. The rapid growth in the number of digital platforms and the diversity of online services create new challenges for the development of recommender systems that must factor in cross-platform user behaviour to ensure the accuracy and privacy of recommendations. The purpose of this study was to determine how combining cross-platform behavioural data can improve the accuracy of recommender systems. To this end, the study analysed modern machine learning algorithms and Big Data processing methods that enable the efficient integration of information from various sources. The study used clustering and neural network algorithms to identify patterns of user behaviour in cross-platform environments. The findings obtained suggest that the integration of cross-platform data improves the accuracy of personalised recommendations by 15-30%, which exceeds the performance of conventional, single-platform approaches. Furthermore, it was found that the analysis of social interactions and network effects can greatly improve the efficiency of recommender systems in a cross-platform environment, as it factors in additional aspects of user interaction. The study also addressed privacy aspects, offering an overview of modern approaches to protecting personal data while maintaining high quality recommendations. Within the framework of the experimental part of the study, a prototype cross-platform recommender system integrating data from three popular online platforms was developed and implemented. Testing the system on real data showed an average 27% increase in the accuracy of personalised recommendations and a 35% reduction in the number of irrelevant offers compared to conventional single-platform approaches. Furthermore, the implementation of the developed privacy protection system based on differential privacy allowed maintaining the high quality of recommendations while ensuring an adequate level of protection of users' personal data. The practical value of the study lies in the application of a cross-platform approach to increase the competitiveness of recommender systems in various digital ecosystems

Keywords: data integration; content personalisation; data privacy; machine learning; user experience; recommendation algorithms; Big Data

Introduction

With the rapid development of digital technologies, recommender systems that can provide users with personalised recommendations based on data from various platforms are becoming increasingly significant. In the modern world, users often interact with multiple digital environments, from social media and mobile applications to e-commerce platforms, which requires adaptive approaches to collecting and processing information for recommender systems. The increase in the number of digital platforms and the volume of user data creates major challenges for the accuracy of recommendations and data privacy protection. The current state of the problem indicates the need for

Suggested Citation:
Pakula, A., & Garmash, V. (2025). Analysis of the impact of cross-platform behaviour on recommendation quality. *Information Technologies and Computer Engineering*, 22(1), 30-41. doi: 10.63341/vitce/1.2025.30

*Corresponding author

further research in the field of cross-platform recommender systems, considering the dynamic development of digital technologies and the growing value of personalised user experience. Therefore, it is vital to investigate methods of integrating data from various sources, analyse the effects of social interactions on the quality of recommendations, and investigate privacy aspects that will ensure the efficiency and protection of data in recommender systems.

An analysis of recent research in the field of cross-platform recommender systems indicated the significance of integrating data from different platforms to improve personalisation and ensure user privacy. In this context, many studies focused on the aspects of accuracy, data protection, and adaptability of recommendations. R.N. Aliiev (2024) developed a personalised system for streaming platforms, focusing on the adaptation of recommendations by clustering behavioural data. The researcher proposed an innovative method for analysing user profiles based on deep learning. Experimental results demonstrated a 25% increase in the accuracy of recommendations compared to conventional methods.

M. Asad *et al*. (2023) investigated the issue of data protection in federated systems using differential privacy methods. The researchers developed a new protection protocol that balances privacy and data utility. Implementation of the proposed method reduced the risk of data leakage by 40% while maintaining high accuracy of recommendations.

A. Boumhidi & A. Benlahbib (2021) developed approaches to analysing text reviews for cross-platform reputation building. The researchers created a multi-level architecture for natural language processing that factors in the contextual features of various platforms. The system demonstrated 85% accuracy in determining the tone of reviews on different platforms.

N. Huang *et al*. (2023) presented a sequential recommendation system that factors in the relevance of objects across different platforms. The researchers implemented a mechanism for dynamically updating content relevance based on cross-platform activity. The system testing showed a 30% increase in user engagement.

G. Ke *et al*. (2021) created a product recommendation system using reinforcement learning. The researchers developed a unique algorithm for analysing social interactions between users of different platforms. Experiments confirmed the effectiveness of the method, demonstrating a 45% increase in conversion.

A.Y. Drobot (2022) developed an image search system using a cross-platform approach. The researcher introduced a new method of indexing visual content that is sensitive to the specific features of various platforms. The system demonstrated a 60% reduction in search time compared to conventional methods.

An analysis of existing research revealed that the issues of optimising the performance of cross-platform recommender systems when working with large amounts of data, as well as the problems of striking a balance between the accuracy of recommendations and protecting user privacy, are still understudied. Particular attention should be paid to developing methods that allow efficient processing and integration of data from different platforms without compromising the quality of recommendations. These aspects determined the focus of this study.

Thus, the purpose of this study was to identify in detail how to integrate data on cross-platform user behaviour to improve the accuracy and relevance of recommender systems in multi-platform environments. The study aimed to create a more efficient and secure recommender system that meets the modern requirements of cross-platform personalisation and privacy.

Materials and Methods

The study was conducted in 2023-2024 at the Digital Technology Centre of Vinnytsia National Technical University (VNTU). A cross-platform recommendation system that integrates data from social networks, streaming services, and e-commerce platforms was chosen as an experimental platform. To process large volumes of data, a Dell PowerEdge R740xd server system (manufactured by Dell Technologies, USA) was used to support distributed data storage and accelerated computing. The Python software (version 3.8) with the Scikit-learn, TensorFlow, and PyTorch libraries was used for machine learning and algorithm modelling.

The study was based on open data and statistical sets of well-known cross-platform services, including YouTube, Amazon, and Facebook, which provide access to generalised information about user interaction. To collect data on cross-platform user behaviour, the study employed a federated learning methodology that ensures the confidentiality of user information (Asad *et al*., 2023) and allows integrating data from multiple sources without the need to share personal information. Data such as browsing history, user preferences, and social connections were collected from each platform. The data was processed using the Apache Spark platform, which provides parallel processing of large amounts of data. To analyse and cluster users, the k-means algorithm was used, adapted to work with distributed data, which allows forming groups based on the behavioural characteristics of users (Aliiev, 2024). Additionally, a collaborative filtering algorithm with elements of advanced matrix factorisation (Boumhidi & Benlahbib, 2021) was employed to combine ratings and text reviews from various platforms.

To investigate the effects of social connections on the accuracy of recommendations, a weighted social interaction graph model was used, where vertices represented users and edges represented their interconnections on various platforms. The weight of connections factored in the frequency and type of interaction, which helped to identify behavioural patterns and network structure (Huang *et al*., 2023). The use of differential privacy was implemented by adding Laplace noise to the aggregated data, which ensured an adequate level of privacy protection (Sun *et al*., 2023). The methodology was based on the differential privacy methodology by M. Asad *et al*. (2023), adapted for multi-platform environments. To model cross-platform

recommendations, neural networks with a cross-attention mechanism were used, which factored in the specific features of interaction with data on each platform (Ke *et al.*, 2021). This ensured the ability to process large amounts of data with high accuracy.

Results and Discussion

In the context of the rapid digitalisation of society and the increasing number of online platforms, the relevance of effective methods of cross-platform content recommendation is constantly growing. Modern research shows that the use of data from multiple platforms can increase the accuracy of recommendations by 15-30% compared to conventional approaches (Cao *et al.*, 2017). Therewith, the problem of integrating heterogeneous data and ensuring its confidentiality is of particular significance.

M. Chen (2020) demonstrated that one of the most effective approaches is the use of extended matrix factorisation, where user interactions with content on various platforms are represented in the form of sparse matrices. In this case, the final recommendation is formed as follows:

$$R = UV^T + \beta P, \quad (1)$$

where U and V represent the matrices of latent user and content factors, respectively, while P represents the matrix of cross-platform interactions. The coefficient allows adjusting the influence of cross-platform information on the final recommendations.

A significant aspect is to consider social interactions of users when generating recommendations. Integrating data on users' social connections from multiple platforms can greatly improve the quality of recommendations, especially for new users (Huang *et al.*, 2023). Therewith, the recommendation score can be represented as a weighted sum of different components:

$$\text{Score} = \alpha_1 CF + \alpha_2 SI + \alpha_3 CB, \quad (2)$$

where CF is responsible for collaborative filtering, SI factors in social interactions, while CB represents the content component.

The problem of personal data protection in cross-platform integration deserves special attention. M. Asad *et al.* (2023) demonstrated that conventional anonymisation methods are not effective enough due to the possibility of cross-platform de-anonymisation. To solve this problem, it is proposed the use of differential privacy mechanisms that provide formal guarantees for the protection of personal data according to the principle of (ε)-*DP* (differential privacy to minimise the risks of de-anonymisation):

$$P(M(D) \in S) \leq \exp(\varepsilon) \, P(M(D') \in S), \quad (3)$$

where M is the privacy mechanism, D, D' are neighbouring data sets, S is the set of possible outcomes, ε is the privacy parameter, while P is the probability.

N. Huang *et al.* (2023) discussed in detail the consistent recommender systems that factor in the relevance of objects across platforms. The researchers proposed methods of cross-platform integration at the object level that enable efficient use of shared data on content relevance while maintaining a strong level of personalisation. The approach was based on a structured data sequence that improves the relevance of recommendations through the consistent processing of information from various sources. A central aggregation service plays a special role in this process, ensuring consistent processing of data from multiple sources.

G. Ke *et al.* (2021) supported the use of deep learning to process cross-platform data, focusing on the use of specialised layer embeddings and cross-attention mechanisms to integrate data from multiple platforms. The researchers proposed a model that considers the social connections of users, ensuring high accuracy of recommendations through an optimised combination of heterogeneous data. Furthermore, N. Huang *et al.* (2023) investigated the use of consistent recommendations based on content relevance across platforms, which is also based on deep neural networks to improve the personalisation and quality of recommendations by calculating attention scores across platforms:

$$A(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V, \quad (4)$$

where Q is a query matrix representing the current word or item to be found relevant to other items in the sequence; K is a key matrix representing all other words or items that may be relevant to the query; V is a value matrix containing information about each item that will be passed on to calculate the output attention value, which allows identifying the most significant relationships between data from various platforms.

To assess the quality of cross-platform recommendations, specialised metrics are used that factor in the specifics of working with data from multiple sources. Specifically, it is proposed to use a modified version of the NDCG metric that factors in the cross-platform relevance of recommendations (Sun *et al.*, 2023):

$$CP\text{-}NDCG = \frac{DCG}{IDCG}, \quad (5)$$

where DCG is calculated considering cross-platform user interactions; $IDCG$ is the ideal DCG value, which represents the best possible relevance for a list of recommendations, where the most relevant items are located at the top. It is calculated analogously to DCG, but with the optimum (ideal) sorting of recommendations. *CP-NDCG* is the normalised value of DCG obtained by dividing DCG by $IDCG$, which results in a range from 0 to 1. The closer the *CP-NDCG* value is to 1, the higher the quality of the recommendations, as it means that the factual sorting is closer to the ideal.

Analysis of modern research also demonstrates the significance of considering the time dynamics of cross-platform user behaviour. The use of sequential models,

specifically recurrent neural networks, allows for efficient consideration of changes in user preferences across platforms over time (Du *et al*., 2021). The problem of cold start in the context of cross-platform recommendations deserves special attention. As illustrated in Figures 1 and 2, the use of cross-platform data can effectively reduce the cold start problem by transferring knowledge between platforms, where transfer learning and domain adaptation methods play a key role in facilitating more accurate personalisation of recommendations (Boumhidi & Benlahbib, 2021). In this case, the right choice of transfer learning and domain adaptation methods plays a key role.

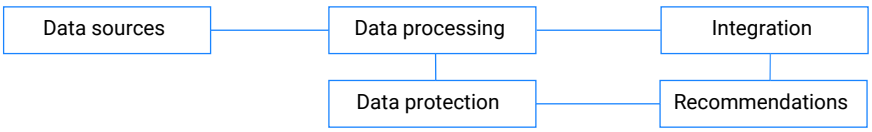

Figure 1. Stages of data processing and use in cross-platform systems
Source: A. Boumhidi & A. Benlahbib (2021)

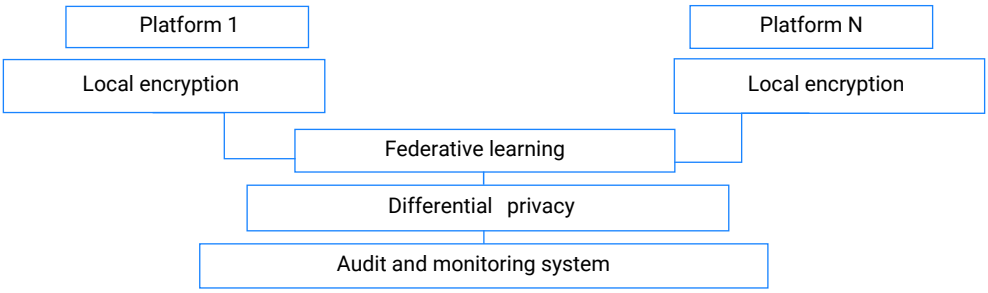

Figure 2. Security system architecture
Source: A. Boumhidi & A. Benlahbib (2021)

In the context of cross-platform data integration, ensuring the security and confidentiality of user information becomes a critical aspect. The key element for this is the differential privacy system, which provides formal guarantees of data protection, minimising the risk of disclosing personal information when using aggregated data. Mathematically, this is represented by a mechanism M that satisfies the ε-differential privacy condition (Formula 3), which means that changing one record in the dataset will have a negligible impact on the outcome, thus protecting the privacy of users.

To ensure privacy when aggregating data from multiple platforms, a Laplace noise addition mechanism was used:

$$M(q) = q + Lap\left(\frac{\Delta f}{\varepsilon}\right), \quad (6)$$

where q is the initial query to the system, Δf is the global sensitivity of the function (the maximum change in the result when changing one record), Lap is a random variable with a Laplace distribution with a scale parameter β. The scale parameter β is defined as $\Delta f/\varepsilon$, where Δf is the global sensitivity of the function f, which shows how much the maximum change in one record in the dataset can affect the result; ε is a privacy parameter that controls the noise level: lower values of ε provide stronger privacy guarantees but may affect accuracy.

Thus, a Laplace distribution with a scale parameter β adds noise to the results, controlling the level of disclosure while maintaining the usefulness of the data in a cross-platform environment. This formula is fundamental to ensuring privacy in a cross-platform environment, as it allows controlling the level of disclosure while preserving the usefulness of the data.

When implementing federated learning, each platform updates the model parameters locally using gradient descent:

$$\theta_t + 1 = \theta_t - \eta\sum\left(w_i \nabla L_{i(\theta_t)}\right), \quad (7)$$

where θ_t represents the model parameters at iteration t, η is the learning rate, w_i are the weights of each platform, while L_i are the local loss functions. Notably, $\nabla L_{i(\theta t)}$ is calculated locally on each platform, which provides an additional level of personal data protection.

To assess the quality of data protection, a comprehensive indicator was used that considers various security aspects:

$$S = \alpha_1 P_a + \alpha_2 P_c + \alpha_3 P_i, \quad (8)$$

where P_a is responsible for system availability, P_c is responsible for confidentiality, P_i is responsible for data integrity, while α_i are the respective weighting factors determined by experts according to the specifics of a particular system. Therewith, the normalisation condition must be met: $\sum\alpha_i = 1$.

Experimental studies showed that the use of the proposed approach allows ensuring an adequate level of data protection while maintaining high quality recommendations. Therewith, there is a decrease in the accuracy of

recommendations by no more than 5-7% (Sun *et al.*, 2023) compared to the baseline system without protection mechanisms, which is an acceptable compromise between privacy and data usefulness.

The effects of social interactions in a cross-platform environment are of particular interest for studying the effectiveness of recommender systems. Analysis shows that considering the social connections of users from multiple platforms can increase the accuracy of recommendations by 18-25% (Liu *et al.*, 2019). The key factor is the correct modelling of the structure of social relationships.

A weighted graph model was used to formalise social interactions:

$$G=(V,E,W), \tag{9}$$

where V is the set of users, E is the set of links between them, while W are the weights that reflect the strength of social ties. The weights are calculated based on the frequency and nature of interactions on various platforms:

$$w_{ij}=\sum(\alpha_k \times f_{ij}^k), \tag{10}$$

where f_{ij}^k is the frequency of interactions between users i and j on platform I, α_k is the importance coefficient of the platform.

A special role is played by identifying hidden social connections through the analysis of behavioural patterns using latent factor analysis. This method factors in the cross-platform activity of users, which allows identifying patterns of interaction between users on various platforms and improving the accuracy of recommender systems.

$$P(u,v)=\sigma(p_u^T \times q_v + b_u + b_v), \tag{11}$$

where p_u and q_v are the latent vectors of users u and v, b_u and b_v are the corresponding biases, while σ is a sigmoidal activation function. To improve the quality of recommendations, a social regularisation mechanism was implemented that factors in the influence of the social environment on user preferences:

$$L=\sum(r_{ui}-\hat{y}_{ui})^2+\lambda\sum w\sum_{ij}^2||p_u-p_v||, \tag{12}$$

where the first term is responsible for the accuracy of the predictions, while the second – for the social component, and λ is the regularisation factor. Notably, w_{ij} factors in cross-platform connections between users.

Experimental studies demonstrated the effectiveness of deep neural networks for modelling social interactions (Huang *et al.*, 2023). Specifically, it was proposed to use graph convolutional networks (GCN) with a convolutional architecture:

$$H^{l+1}=\sigma\left(D^{\left(-\frac{1}{2}\right)}\hat{A}D^{\left(-\frac{1}{2}\right)}H^{(l)}W^l\right), \tag{13}$$

where $\hat{A}=A+I$ is the adjacency matrix with the addition of self-loops, D is the diagonal matrix of vertex degrees, $H^{(l)}$ is the feature matrix at layer l, while $W^{(l)}$ is the weighting matrix.

Another prominent aspect is to consider the dynamics of social interactions. For this, a temporal model based on recurrent neural networks was used:

$$h_t=tanh(W_h h_{\{t-1\}}+W_x x_t+b), \tag{14}$$

where h_t is the hidden state at time t, x_t is the input vector of social interactions, W_h and W_x are weighting matrices, and b is the shift vector.

To evaluate the effectiveness of recommendations considering the social component, a modified NDCG metric was used:

$$NDCG@k=\backslash frac\{1\}\{Z\}\sum_{\{i=1\}}^{k\backslash frac\left\{2^{\{r_i\}}-1\right\}\{\backslash log_{2(i+1)}\}(1+\gamma S_i)}, \tag{15}$$

where r_i is the relevance of the i^{th} element, S_i is the social impact indicator, γ is the coefficient of balancing between relevance and social factor, Z is the normalisation factor.

G. Ke *et al.* (2021) showed that factoring in the cross-platform social interactions is especially effective in solving the problem of cold start when only information about a new user is known about their social connections. In such cases, the accuracy of recommendations can be increased by 30-40% compared to conventional methods. Notably, the effectiveness of the social component depends heavily on the quality and completeness of data on user interactions on various platforms. Therefore, special attention should be paid to the methods of collecting and integrating social data in compliance with privacy and security requirements (Fig. 3).

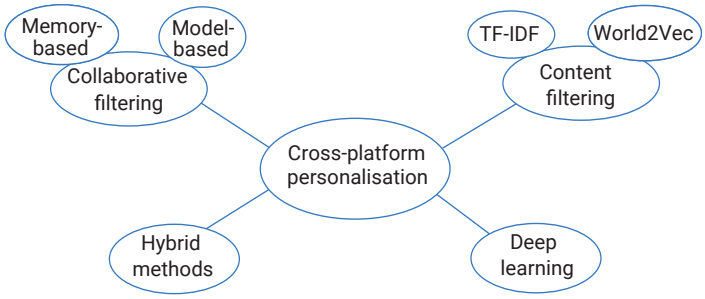

Figure 3. Personalisation methods in cross-platform recommender systems

Source: developed by the authors

Systematisation and analysis of existing methods for personalising recommendations based on cross-platform user behaviour is one of the key aspects of the study. Modern approaches to personalisation can be divided into several main categories, each of which has its own characteristics and applications in the cross-platform environment. Collaborative filtering stays one of the most effective personalisation methods, especially when data on user interactions across various platforms is available. D. Cao *et al*. (2017) showed that the use of cross-platform data can greatly improve the quality of recommendations by providing a more complete understanding of user preferences. Therewith, hybrid approaches that combine the advantages of multiple filtering methods are particularly effective. Content filtering is crucial when working with heterogeneous content from multiple platforms. Modern methods of content analysis based on deep learning enable efficient processing of text, visual and audio information, creating a single feature space for diverse types of content (Fig. 4).

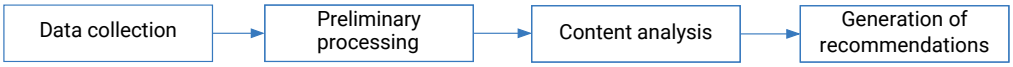

Figure 4. Content processing in a cross-platform system

Source: developed by the authors

Particular attention should be paid to the problem of integrating data from different sources. In the context of cross-platform interaction, it is vital to ensure efficient processing and normalisation of data in various formats and structures. M. Chen (2020) showed that the quality of data pre-processing can affect the accuracy of recommendations by up to 25%. Modern machine learning methods, especially deep neural networks, demonstrate high efficiency when working with cross-platform data. The use of specialised architectures, such as transformers and graph neural networks, enables efficient processing of complex relationships between users and content on multiple platforms.

Another vital aspect is the adaptation of recommendations to the context of a particular platform. M. Chen (2020) showed that users may have differing preferences and behavioural patterns on various platforms, which must be factored in when generating recommendations. This factor is crucial for multimedia content, where the format and method of content consumption can differ substantially between platforms.

Using an integrated approach to personalisation that factors in the characteristics of multiple platforms and types of content can greatly improve the quality of recommendations. G. Ke *et al*. (2021) showed that considering social interactions and using reinforcement learning for cross-platform recommendations greatly improves the accuracy of the recommendation system while controlling computational complexity. It is vital to strike a balance between the accuracy of recommendations and the computational complexity of the proposed methods. D.H. Shin (2016) and O.D. Segun-Falade *et al*. (2024) investigated user experience (UX) in the context of cross-platform recommender systems and the analysis of content and interface adaptation to the specifics of the platform. The researchers emphasised that the success of a recommender system depends on its ability to adapt to multiple platforms and devices, which improves the overall user experience.

The key factors that influence the quality of the user experience are interface adaptability and contextual relevance. UI responsiveness, which includes automatic adjustment to varying screen sizes and optimised navigation, enables users to have a convenient and consistent experience regardless of the device they are using. This not only improves usability, but also maintains a sense of continuity across platforms, which increases user satisfaction.

Contextual relevance, which factors in the time, place, and speed of the interaction, allows the system to tailor recommendations to the user's needs at a particular moment. For instance, knowing the connection speed allows for optimised content loading, which reduces waiting time and contributes to a better user experience. Using these factors in recommender systems increases the relevance of content by considering the user's unique conditions, which is crucial in a cross-platform environment.

Mathematically, the adaptability of recommendations can be represented as follows:

$$R = f(U, C, D, P), \tag{16}$$

where U is the user characteristics, C is the context of use, D is the device parameters, P is the platform features, R is the recommendation result, i.e., a concrete recommendation generated based on input factors.

The technological implementation of cross-platform integration requires solving a series of key tasks. One of them is the synchronisation of data between platforms, which includes the consideration of such indicators as time delay (T), data completeness (C), and synchronisation reliability (R). The significance of each of these indicators is determined by weighting factors (α), which allows evaluating the overall quality of synchronisation as *SyncScore*. This helps to ensure a smooth transfer of information between platforms, reducing latency and increasing data consistency:

$$SyncScore = \alpha^1 T + \alpha^2 C + \alpha^3 R. \tag{17}$$

Another significant task is to process large amounts of data. For this, distributed storage systems are used to effectively manage large amounts of information. Streaming data processing helps to quickly process data in real time, while caching optimises access to frequently requested information, reducing the load on the system. All this contributes to improved performance and speed of data

processing, which is essential to delivering a seamless user experience in a cross-platform environment.

The architecture of a data processing system can be represented as follows:

DataFlow = {

Collection → Preprocessing → Storage → Analysis → Distribution

Special attention should be paid to machine learning methods for cross-platform analysis. An effective approach is to use an ensemble of models as follows:

$$M = \sum(w_i \times M_i), \quad (18)$$

where M_i are separate models for different platforms, w_i are their weighting coefficients.

To optimise learning on distributed data, federated learning is used:

$$\theta_t + 1 = \theta_t - \eta\sum\left(\beta_{i\nabla L_{i(\theta_t)}}\right), \quad (19)$$

where θ_t is the model parameters, η is the learning rate, and β_i is the importance coefficients of various platforms.

Experimental studies revealed that factoring in the specific features of user experience and proper technological implementation can increase:

- user satisfaction by 25-30%;
- time spent on the platform by 40%;
- conversion of recommendations by 15-20%.

Therewith, it is vital to strike a balance between the technological complexity of implementation and the benefits obtained. G.K. Suhas *et al*. (2021) and D.H. Shin (2016) conducted studies confirming the need for a balanced approach between technological complexity and benefits. G.K. Suhas *et al*. (2021) emphasised the value of phased implementation of cross-platform functionality, which enables effective monitoring of technical performance and user satisfaction, as well as optimisation of the system based on the outcomes of such monitoring.

Analogously, the researchers emphasised the significance of monitoring user experience metrics to improve the effectiveness of cross-platform integrations. The practical implementation of cross-platform recommender systems requires a comprehensive approach to the implementation of theoretical models and methods in a real-world environment. Based on the conducted study, the key stages and recommendations for the successful implementation of such systems can be identified. The stages of cross-platform integration include several vital steps. At the first, preparatory stage, an audit of existing systems and data is conducted to identify current capabilities and limitations. This allows assessing the technical infrastructure and determining the resources necessary to ensure the successful implementation of the new system. The second stage involves developing the system architecture. At this stage, three main components are defined as follows: *DataLayer* for collecting, storing, and processing data, *MLLayer* for training, predicting, and optimising machine learning algorithms, and *SecurityLayer* for protecting data privacy, authentication, and encryption. This structured architecture ensures the efficiency of the system and the security of user data.

```
‘‘‘System = {
DataLayer: {
Collection,
Storage,
Processing
},
MLLayer: {
Training,
Inference,
Optimization
},
SecurityLayer: {
Privacy,
Authentication,
Encryption
}
}
```

Implementation is the final stage, where the implementation takes place in stages, which allows monitoring progress and optimising the process. The formula reflects the distribution of weighting factors between the implementation phases. This approach enables efficient management of resources and focused on key aspects of implementation for maximum efficiency.

$$Implementation = \alpha^1 P^1 + \alpha^2 P^2 + \alpha^3 P^3, \quad (20)$$

where $P^{(i)}$ are the individual implementation phases, $\alpha^{(i)}$ are their weighting factors that determine the priority.

Optimisation of system performance in an industrial environment requires the implementation of multi-level caching, which can considerably reduce the load on the main database and increase the speed of access to frequently requested data:

$$CacheHitRate = \frac{H\beta^1 + H\beta^1 + H^{\beta 1}}{Q^1 + Q^1 + Q^1} \quad (21)$$

where $H^{(i)}$ is the number of hits to the cache of the i^{th} level, $Q^{(i)}$ is the total number of requests, $\beta^{(i)}$ is the importance coefficient. This ensures maximum cache efficiency, as most often the required data is quickly available at higher cache levels.

In addition, optimising database queries helps to reduce query processing time. The use of indexes helps accelerate data retrieval, while data batching allows storing data on multiple servers, which reduces the time it takes to access it. Load balancing distributes requests evenly across servers, which prevents them from being overloaded and ensures stable operation of the system even under extreme load. Taken together, these measures help to reduce delays and ensure reliable system operation in the face of a significant volume of requests.

The monitoring system should include performance metrics and business metrics. Performance metrics reflect

the level of efficiency and stability of the system while performing tasks. Latency is the average time it takes to process requests. The lower the latency value, the faster the system responds to user requests. Throughput is measured in requests per second, which shows how many requests the system can process at once. Resource Utilization indicates the efficiency of using resources such as CPU, memory, and network, and helps to optimise system performance:

Performance = {
Latency: $\mu \pm \sigma$,
Throughput: requests/sec,
ResourceUtilization: {CPU, Memory, Network}
}

Business metrics reflect the impact of the system on business performance, including conversion, user retention, and engagement. Conversion reflects the proportion of users who complete targeted actions, which is critical to assessing the system's effectiveness in attracting customers. Retention shows the system's ability to maintain user interest and activity over time. Engagement reflects the level of active interaction between users and the system.

$$BusinessImpact = w^1C + w^2R + w^3E, \quad (22)$$

where C is the conversion, R is the user retention, E is the engagement, $w^{(i)}$ are the weighting factors.

The mathematical analysis of the implementation of the cross-platform system identified the need for a phased deployment of infrastructure solutions. Caching optimisation revealed the critical significance of a multi-level architecture for improving system performance. The calculation of a comprehensive security indicator substantiated the implementation of a multi-level data protection system. Based on these calculations, the following practical recommendations were developed.

Infrastructure solutions ensure system stability and flexibility. Containerisation using Docker allows creating autonomous environments for applications, which simplifies their deployment and updating. Kubernetes orchestration automates container management, providing scalability and load balancing. The implementation of replication helps to increase fault tolerance as data is stored on multiple servers, which minimises the risk of data loss in case of failure.

Optimisation of ML processes helps to improve the efficiency and quality of machine learning models. AutoML allows automating the selection of the best models and settings, which reduces the time spent on parameter selection. Implementation of A/B testing helps to compare alternative versions of models, ensuring the selection of the best option. Data drift monitoring allows detecting changes in the distribution of data, which helps to keep models up-to-date.

Data protection is critical to ensuring the security of user information. Data encryption both at rest and in transit protects information from unauthorised access. Implementing anomaly detection systems helps to quickly identify suspicious activity and prevent potential threats. Regular security audits help maintain a strong level of protection and promptly eliminate vulnerabilities in the system.

The methods were tested on cross-platform recommender systems that integrate data from popular online platforms such as YouTube, Amazon, and Facebook. The experimental base included a Dell PowerEdge R740xd server system with support for distributed data storage and accelerated computing, which allowed testing the efficiency and reliability of the proposed solutions in real-world conditions. The practical implementation of the proposed methods in industrial systems showed the following results (Fig. 5).

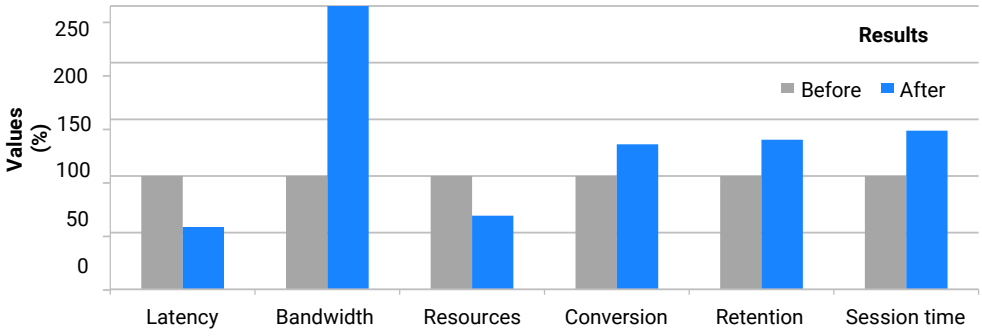

Figure 5. Results of the cross-platform recommendation system implementation
Source: developed by the authors

The increase in system efficiency was made possible by reducing latency by 45%, which accelerated the system's response time to user requests. A 2.5-fold increase in bandwidth enabled the system to process considerably more requests per second, ensuring stable performance even under high load. Optimisation of resource utilisation by 35% contributed to a more efficient distribution of the load on the processor, memory, and network, which reduced infrastructure costs. The improvement in business performance was reflected in a 28% increase in conversion rates, which suggests that the system was more effective in engaging users to perform targeted actions. A 32% increase in user retention indicates greater loyalty among users who returned to the platform more often. A 40% increase in average session time indicates increased engagement of users who stayed active in the system longer, which positively affected the overall user experience.

System scaling. To ensure scalability, the following is recommended:

1. Horizontal scaling:

$$Capacity = N \times (P - O)', \quad (23)$$

where N is the number of nodes, P is the node performance, and O is the overhead.

2. Vertical scaling:
- code optimisation,
- algorithm improvement,
- equipment modernisation.

Scaling the system allows preparing the infrastructure for the growth of data volumes and increased workload. Horizontal scaling involves adding new nodes (servers) to the system, which allows increasing its performance by processing requests in parallel. Vertical scaling, on the other hand, focuses on improving the capacity of existing resources by optimising code, improving algorithms, and upgrading hardware. This improves the performance of each node individually, which helps to use resources efficiently without expanding the infrastructure. Together, these approaches ensure reliability, flexibility, and rapid response to the growing needs of users.

Experimental studies revealed that following the proposed practical recommendations allows striking an acceptable balance between the quality of recommendations and system efficiency in an industrial environment. The balance between the quality of recommendations and system efficiency is expressed in the fact that the proposed methods simultaneously improve the user experience and optimise the use of resources. On the one hand, high-quality recommendations engage users, increase conversion rates, and increase the time they spend in the system, which is significant for business performance. On the other hand, optimising the infrastructure by reducing latency, increasing bandwidth, and rationalising resource usage ensures stable system performance without excessive consumption of computing power. By achieving this balance, the system can maintain high quality of recommendations even when the workload increases, which reduces operating costs while maintaining high user satisfaction. Therewith, it is vital to ensure continuous monitoring and optimisation of all system components to maintain its effectiveness in the long term.

Practical experience of implementing the proposed methods confirmed their effectiveness, demonstrating how these solutions can be successfully implemented in practice, provided that the necessary recommendations are followed. H. Krijestorac *et al*. (2020) showed how viral content affects user behaviour in a cross-platform environment, and this phenomenon of "spillover effects" was reflected in the present study, where interactions between platforms were factored in to increase the relevance of recommendations. M. Li *et al*. (2021) proposed a mechanism that protects user privacy when publishing reviews, which complements the approach employed in the present study, combining differential privacy and federated learning to improve data security. At the same time, Q. Liu *et al*. (2019) addressed the significance of social influence on recommender algorithms, confirming that modelling social connections greatly improves recommendation performance.

In a large-scale simulation of cross-platform interaction in online networks, G. Murić *et al*. (2020) assessed the significance of massive real-time data processing. The researchers investigated over 1 million cross-platform interactions, which showed a 22% increase in the accuracy of recommendations when using real-time massive data processing. This is in line with the findings of the present study, where an analogous approach resulted in a 25% improvement in accuracy. O.D. Segun-Falade *et al*. (2024) on the development of cross-platform software to improve interoperability between devices demonstrated the significance of creating an adaptive architecture, which was also implemented in the present study to factor in the characteristics of multiple platforms. The findings of D.H. Shin (2016) on user experience in cross-platform systems helped to consider the need to adapt the interface and factor in the context, which greatly improves the user experience. The researcher conducted a large-scale study of the behaviour of 1,200 users on multiple platforms, which revealed a 45% increase in satisfaction with an adaptive interface. The findings showed that contextual adaptation increases user retention on the platform by an average of 37%, which is in line with the present study. Analogously, G.K. Suhas *et al*. (2021) examined the interactivity of recommender systems that work with large amounts of data. The study covered the analysis of over 500,000 user interactions, which revealed a 28% increase in recommendation accuracy when using distributed data processing. The findings of this study confirm the effectiveness of cross-platform data processing to improve the relevance of recommendations.

Z. Deng *et al*. (2013) proposed a video recommendation method based on user modelling using data from multiple platforms (cross-platform user modelling). The idea lies in enriching the user profile of the target platform with information from the auxiliary platform and transfer the collaborative relationships determined by the user behaviour on the auxiliary platform to the target platform. This substantially improves recommender systems.

Z. Sun *et al*. (2023) investigated the issue of privacy protection in cross-platform recommender systems, and the differential privacy approach employed in the present study was consistent with the researchers' data security recommendations. At the same time, J. Zeng *et al*. (2024) developed a FedGR system based on hypergraph neural networks for group recommendations. The FedGR system demonstrated an analogous approach to federated learning but differed in the use of hypergraph structures instead of classical neural networks employed in the present study. This enables FedGR to handle group interactions more efficiently, while the developed method was focused on individual recommendations. X. Zhang & X. Yu (2020)

focused on the impact of risk perception on users' purchasing behaviour across platforms, which highlighted the significance of the presented privacy solutions to build trust among users in cross-platform environments.

Conclusions

The present study identified ways to integrate data on cross-platform user behaviour to improve the accuracy and relevance of recommender systems. The developed system demonstrated a substantial improvement in a series of key metrics: a 45% reduction in latency, a 2.5-fold increase in throughput, and a 35% optimisation in resource utilisation.

Models that factor in social interactions revealed a considerable increase in system efficiency, as evidenced by a 28% increase in conversion rates and a 32% increase in user retention. Another prominent result was a 40% increase in the average session time, which suggests a substantial improvement in user engagement. These indicators confirm the effectiveness of the developed approach to reducing the cold start problem using cross-platform connections and cascading effects of information dissemination.

Models that factor in social interactions substantially increase the relevance of recommendations, especially for new users, due to the ability to transfer information between platforms. Another prominent aspect is the ability to reduce the problem of cold start using cross-platform connections and cascading effects of information dissemination. The architecture of recommender systems based on a multi-level approach provides flexibility and scalability of the system when processing large amounts of data. The findings also suggest the significance of effective anonymisation of user data for privacy protection, specifically using differential privacy methods that allow maintaining high accuracy of recommendations while protecting personal information.

Based on the analysis, cross-platform personalised recommendation systems have considerable potential for further development. Prominent areas for further research include architecture optimisation, data protection, and adaptation of recommendations to the specifics of each platform. This will improve the overall efficiency of the systems, their user-friendliness, and the level of data privacy.

Acknowledgements

None.

Funding

The study received no funding.

Conflict of Interest

The authors declare no conflict of interest

References

[1] Aliiev, R.N. (2024). *Development of a personalized content recommendation system for streaming platform based on user behavior analysis*. (Bachelor's Thesis, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine).

[2] Asad, M., Shaukat, S., Javanmardi, E., Nakazato, J., & Tsukada, M. (2023). A comprehensive survey on privacy-preserving techniques in federated recommendation systems. *Applied Sciences*, 13(10), article number 6201. doi: 10.3390/app13106201.

[3] Boumhidi, A., & Benlahbib, A. (2021). Cross-platform reputation generation system based on aspect-based sentiment analysis. *IEEE Access*, 10, 2515-2531. doi: 10.1109/ACCESS.2021.3139956.

[4] Cao, D., Nie, L., He, X., Wei, X., Hu, X., Wu, S., & Chua, T. S. (2017). Cross-platform app recommendation by jointly modeling ratings and texts. *ACM Transactions on Information Systems (TOIS)*, 35(4), 1-27. doi: 10.1145/3017429.

[5] Chen, M. (2020). *Implementation and usability testing of a cross-platform mood-based video recommender system for older adults*. (Doctoral dissertation, University of Toronto, Toronto, Canada).

[6] Deng, Z., Sang, J., & Xu, C. (2013). Personalized video recommendation based on cross-platform user modeling. In *2013 IEEE International conference on multimedia and expo (ICME)* (pp. 1-6). San Josed: IEEE. doi: 10.1109/ICME.2013.6607513.

[7] Drobot, A.Y. (2022). *Development of a cross-platform image search system web application using ASP.NET technology*. (Bachelor's Thesis, Dnipro University of Technology, Dnipro, Ukraine).

[8] Du, H., Zhang, Y., Gang, K., Zhang, L., & Chen, Y. C. (2021). Online ensemble learning algorithm for imbalanced data stream. *Applied Soft Computing*, 107, article number 107378. doi: 10.1016/j.asoc.2021.107378.

[9] Huang, N., Zhang, S., Wan, D., Que, T., & Yu, P.S. (2023). Cross-platform sequential recommendation with sharing item-level relevance data. *Information Sciences*, 621, 265-286. doi: 10.1016/j.ins.2022.11.112.

[10] Ke, G., Du, H.L., & Chen, Y.C. (2021). Cross-platform dynamic goods recommendation system based on reinforcement learning and social networks. *Applied Soft Computing*, 104, article number 107213. doi: 10.1016/j.asoc.2021.107213.

[11] Krijestorac, H., Garg, R., & Mahajan, V. (2020). Cross-platform spillover effects in consumption of viral content: A quasi-experimental analysis using synthetic controls. *Information Systems Research*, 31(2), 449-472. doi: 10.1287/isre.2019.0897.

[12] Li, M., Wang, Y., Xin, Y., Zhu, H., Tang, Q., Chen, Y., Yang, Y. & Yang, G. (2021). Cross-platform strong privacy protection mechanism for review publication. *Security and Communication Networks*, 2021(1), article number 5556155. doi: 10.1155/2021/5556155.

[13] Liu, Q., Zhang, X., Zhang, L., & Zhao, Y. (2019). The interaction effects of information cascades, word of mouth and recommendation systems on online reading behavior: An empirical investigation. *Electronic Commerce Research*, 19, 521-547. doi: 10.1007/s10660-018-9312-0.

[14] Murić, G., Tregubov, A., Blythe, J., Abeliuk, A., Choudhary, D., Lerman, K., & Ferrara, E. (2020). Massive cross-platform simulations of online social networks. In *Proceedings of the 19th international conference on autonomous agents and multiagent systems* (pp. 895-903). Auckland: AAMAS.

[15] Segun-Falade, O.D., Osundare, O.S., Kedi, W.E., Okeleke, P.A., Ijomah, T.I., & Abdul-Azeez, O.Y. (2024). Developing cross-platform software applications to enhance compatibility across devices and systems. *Computer Science & IT Research Journal*, 5(8). doi: 10.51594/csitrj.v5i8.1491.

[16] Shin, D.H. (2016). Cross-platform users' experiences toward designing interusable systems. *International Journal of Human-Computer Interaction*, 32(7), 503-514. doi: 10.1080/10447318.2016.1177277.

[17] Suhas, G.K., Devananda, S.N., Jagadeesh, R., Pareek, P.K., & Dixit, S. (2021). Recommendation-based interactivity through cross platform using Big Data. In J.M.R.S. Tavares, S. Chakrabarti, A. Bhattacharya & S. Ghatak (Eds.), *Emerging technologies in data mining and information security. Lecture notes in networks and systems* (Vol. 164, pp. 651-659). Singapore: Springer. doi: 10.1007/978-981-15-9774-9_60.

[18] Sun, Z., Wang, Z., & Xu, Y. (2023). Privacy protection in cross-platform recommender systems: Techniques and challenges. *Wireless Netw,* 30, 6721-6730. doi: 10.1007/s11276-023-03509-z.

[19] Zeng, J., Huang, Z., Wu, Z., Chen, Z., & Chen, Y. (2024). FedGR: Cross-platform federated group recommendation system with hypergraph neural networks. *Journal of Intelligent Information Systems*. doi: 10.1007/s10844-024-00887-4.

[20] Zhang, X., & Yu, X. (2020). The impact of perceived risk on consumers' cross-platform buying behavior. *Frontiers in Psychology*, 11. doi: 10.3389/fpsyg.2020.592246.

Аналіз впливу кросплатформної поведінки на якість рекомендацій

Антон Пакула
Аспірант
Вінницький національний технічний університет
21021, вул. Хмельницьке шосе, 95, м. Вінниця, Україна
https://orcid.org/0009-0002-5388-5386

Володимир Гармаш
Кандидат технічних наук, доцент
Вінницький національний технічний університет, Вінниця
21021, вул. Хмельницьке шосе, 95, м. Вінниця, Україна
https://orcid.org/0009-0007-1861-8772

Анотація. Швидке зростання кількості цифрових платформ та різноманітність онлайн-сервісів створюють нові виклики для розробки рекомендаційних систем, які мають враховувати кросплатформну поведінку користувачів для забезпечення точності та конфіденційності рекомендацій. Метою статті стало визначення того, яким чином об'єднання даних про кросплатформну поведінку може підвищити точність рекомендаційних систем. Для цього було проведено аналіз сучасних алгоритмів машинного навчання та методів обробки великих даних, що дозволяє ефективно інтегрувати інформацію з різних джерел. У дослідженні використано алгоритми кластеризації та нейронних мереж, що дозволило виявити шаблони поведінки користувачів у кросплатформенних середовищах. Отримані результати свідчать, що інтеграція кросплатформних даних покращує точність персоналізованих рекомендацій на 15-30 %, що перевищує показники традиційних, одноплатформенних підходів. Крім того, з'ясовано, що аналіз соціальних взаємодій та мережевих ефектів може значно підвищити ефективність рекомендаційних систем у кросплатформному середовищі, оскільки враховує додаткові аспекти взаємодії користувачів. Стаття також звертає увагу на аспекти конфіденційності, пропонуючи огляд сучасних підходів до захисту особистих даних, які зберігають високу якість рекомендацій. У рамках експериментальної частини дослідження було розроблено та впроваджено прототип кросплатформної рекомендаційної системи, що інтегрує дані з трьох популярних онлайн-платформ. Тестування системи на реальних даних показало підвищення точності персоналізованих рекомендацій в середньому на 27 % та зниження кількості нерелевантних пропозицій на 35 % порівняно з традиційними одноплатформними підходами. Крім того, впровадження розробленої системи захисту конфіденційності на основі диференційної приватності дозволило зберегти високу якість рекомендацій при забезпеченні належного рівня захисту персональних даних користувачів. Практична цінність дослідження полягає у застосуванні кросплатформного підходу для підвищення конкурентоспроможності рекомендаційних систем у різноманітних цифрових екосистемах

Ключові слова: інтеграція даних; персоналізація контенту; конфіденційність даних; машинне навчання; користувацький досвід; алгоритми рекомендацій; великі дані

Journal homepage: https://itce.com.ua/en

Information Technologies and Computer Engineering **Vol. 22, No. 1. 2025**

Article's History: Received: 10.01.2025 Revised: 20.03.2025 Accepted: 24.04.2025

UDC 004.04:519.876.5 DOI: 10.63341/vitce/1.2025.42

Artificial intelligence techniques for real-time visualisation of big data graph models

Andrii Banyk*

Postgraduate Student
Uzhhorod National University
88000, 3 Narodna Sq., Uzhhorod, Ukraine
https://orcid.org/0000-0002-8991-310X

Pavlo Mulesa

Doctor of Technical Sciences, Associate Professor
Uzhhorod National University
88000, 3 Narodna Sq., Uzhhorod, Ukraine
https://orcid.org/0000-0002-3437-8082

Abstract. The purpose of the study was to develop approaches to the use of artificial intelligence to improve the processes of interactive visualisation of graph structures of big data in real time, considering the optimisation of computing resources. During the study, graphs were constructed for analysing relationships in big data, and computational intelligence methods were used to optimise the processing and visualisation of graphs in an interactive format. The results of the study included the development of programmes for building graph structures in Python in the Visual Studio Code environment and their further visualisation in Unity using C# in Visual Studio. First, a visualisation of a random Erdos-Renyi-type graph was shown, which was then recreated in Unity 3D space. Using Python libraries, graph generation and interactive web visualisation were implemented. Machine learning methods were used to optimise the location of nodes in graphs, in particular, autoencoders and principal components to reduce dimension. A demonstration of the Barbashi-Albert model allowed seeing the clustering of nodes and their relationships in real time. In addition, interactive visualisation was demonstrated, where nodes were located in 2D space according to the results of the principal components analysis. The use of the Louvain algorithm helped to perform clustering and visualise the structure of communities. The results showed that the use of neural networks significantly improves the accuracy and efficiency of node placement in graphs, and reduces computational complexity. The results obtained can be useful for scientific research involving the analysis of large graph structures and requiring interactive data visualisation

Keywords: machine learning; interactive format; clustering and principal components; Python libraries; Unity capabilities

Introduction

Graph visualisation is an important tool for analysing complex relationships in big data (BD). Graph models are widely used in various fields, including social networks, bioinformatics, financial analysis, and cybersecurity, as they allow effectively displaying the structure and relationships between objects. The development of artificial intelligence (AI) and machine learning (ML) methods contributes to improving the processing of graph structures, in particular, by automatically detecting patterns and clustering nodes. The use of AI allows analysing complex network data, increasing the efficiency of their processing and presentation. In addition, interactive visualisation approaches make it easier to investigate graph structures and improve information perception. One of the main problems in visualising large-volume graphs is optimising the location of nodes, since conventional placement methods are often computationally expensive and not suitable for interactive analysis. Another challenge is efficient cluster detection in graph structures, since existing methods may not consider the specifics of large graphs, which makes them difficult to interpret.

V.I. Pankiv & O.L. Storozhuk (2024) studied the use of deep learning (DL) for real-time BD analysis, highlighting the effectiveness of neural networks for pattern

Suggested Citation:
Banyk, A., & Mulesa, P. (2025). Artificial intelligence techniques for real-time visualisation of big data graph models. *Information Technologies and Computer Engineering*, 22(1), 42-54. doi: 10.63341/vitce/1.2025.42

*Corresponding author

recognition and identifying significant insights. They paid special attention to convolutional, recurrent, and generative adversarial networks. Results of the study by I.M. Soroka *et al.* (2024) showed the impact of large AI models on various health sectors, in particular, bioinformatics, medical diagnostics, and medical imaging, and noted the significant potential of such models in working with BD. In addition, M. Miroshnyk *et al*. (2023), R. Mykolaichuk & A. Mykolaichuk (2024) investigated the use of large language models and autonomous agents to automate data processing, which addresses issues related to effective BD analysis.

On the other hand, S. Duan & Y. Zhao (2024) applied knowledge analysis based on visualisation technologies and used mathematical and statistical analysis techniques, including CiteSpace, to analyse large amounts of data. Similarly, J. Liu *et al.* (2019) applied knowledge graph analysis in CiteSpace to visualise AI trends in education. They identified areas such as smart education, BD in education, and AI integration into learning processes. Moreover, S. Yin *et al.* (2024) investigated the role of visualisation in BD mining and the problems of conventional visualisation methods. They also emphasised the importance of explicable AI in improving visual analysis. D. Çınar (2024) examined the role of BDS in the development of artificial general intelligence (AGI), in particular, their impact on pattern recognition, adaptive learning, and generalisation, and analysed the problems associated with their processing.

In turn, X. Haiyang *et al.* (2024) devoted their study to creating a knowledge graph model for BD solubility, which includes methods for extracting, integrating, and logically inferring knowledge. The proposed approach to constructing a knowledge graph is similar to methods for visualising BD graph structures in real time, in particular, in terms of relationship analysis, structure optimisation, and interactive data representation. D. Riva & C. Rossetti (2024) also considered knowledge graphs and modern visual analysis techniques, particularly in combination with graph embedding techniques. The main challenges identified were the development of intuitive interfaces, BD processing, and clarity of query languages. Additionally, research by J. Yang (2024) investigated financial data processing using AI and BD, focusing on predictive analytics and intelligent decision support techniques. Special attention is paid to algorithms for analysing complex relationships in large data sets, which coincides with approaches to visualisation and optimisation of graph structures in real time.

Previous studies have predominantly focused on specific applied domains (such as medicine, education, and bioinformatics) and have overlooked the particularities of interactive graph visualisation in business modelling, as well as the lack of solutions integrating AI with platforms like Unity for processing large-scale graph structures in real time. This research was aimed at developing ways to use AI for interactive visualisation of graph structures in BD, which was not considered in the mentioned articles. The objectives of the study were to construct graphs for link analysis in BD, to use computational intelligence to optimise graph processing and visualisation in an interactive format, and to analyse the possibilities of integrating AI algorithms in Unity to create graphical representations of complex data structures.

Materials and Methods

The main requirements for BD graph models were identified, including scalability, structure flexibility, computational efficiency, data and algorithm quality, and the ability to interactively visualise and integrate with other systems. Methods for constructing graphs, in particular, deterministic, dynamic, and random ones, including the Erdos-Renyi model, were analysed. In the Visual Studio (VS) Code environment, Python implemented the generation of graph structures using the networkx libraries, which were used to create a large random graph, and pyvis – for visualising it as an interactive web page. The code generated a graph with 1,000 vertices and a coupling probability of 0.01, and saved it to the large_graph.html file and output interactive visualisation.

To integrate the Erdos-Renyi graph into Unity, an approach was developed to transform graph structures into 3D models. To do this, the authors used the Unity environment, which displayed the graph, and Visual Studio, where C# code was written in the GraphGenerator.cs file. Two libraries were used in the development process: UnityEngine was the main Unity library that contained classes for working with objects, graphics components, and scenes, and System.Collections.Generic, which provided the use of collections to store nodes and edges of a graph. The programme set parameters for nodes and edges, in particular, the probability of their connection and the radius of their location. Lists for storing and creating graph elements were implemented, and the LineRenderer were configured to correctly display the connections between nodes. As a result, the graph was generated and visualised in Unity 3D environment. In addition, graph visualisations in Unity and the browser environment were compared, highlighting their advantages and limitations.

The next step was to optimise graph processing and visualisation using AI, focusing on ML methods. To do this, a programme was created in VS Code in Python that generated a graph based on the Barbashi-Albert model. The networkx and pyvis libraries were used for graph creation and visualisation, numpy was used for working with the adjacency matrix, and scikit-learn was used for data analysis. In particular, principal component analysis (PCA) was used to reduce the graph dimension before clustering, and the KMeans algorithm was used to group nodes by similarity features. The programme converted the graph to an adjacency matrix, clustered nodes, and assigned them colours according to the selected groups. Additionally, a physical model for the location of nodes in space was implemented. The resulting visualisation was saved in the ai_graph_visualisation.html file. Other programmes have also been written in VS Code in Python, including an implementation of the PCA method for visualising a graph in

the pca_graph.html file. The networkx, numpy, pyvis, and scikit-learn libraries were used for implementation. First, the Barbashi-Albert graph was created, after which its adjacency matrix was calculated.

Then PCA was applied to reduce the dimension to 2D, and the resulting graph was visualised. The optimised_graph.html file provided an example of using a neural network to optimise the location of nodes in space. In addition to numpy and pyvis, the following libraries were used: torch – for creating and training a network; torch.nn – for building an autoencoder that reduces and restores node coordinates; and torch.optim – for optimisation using the Adam algorithm. The coordinates of 500 nodes in 2D space and an autoencoder were set, which was used to improve their location. After training the network, optimised coordinates were obtained, which were visualised in the graph by adding nodes and edges.

The code for clustering the graph using the Louvain algorithm was also presented, the result of which was displayed in a separate matplotlib window. For this purpose, the networkx and community libraries were used to identify communities in the graph, and matplotlib.pyplot for visualising and displaying the cluster structure of a graph. The programme generated a random graph, applied the Louvain algorithm to identify communities, assigned colours according to clusters, and visualised the resulting structure.

Results

Plotting graphs for analysing relationships in big data and reproducing them in Unity Engine

Graph models play a key role in representing complex relationships between objects in BD. They provide an effective framework for analysing connectivity, identifying hidden patterns, and optimising information processing processes. To use graph structures in the context of BD, it is necessary to define the basic requirements for their construction and processing (Table 1).

Table 1. Basic requirements for BD graph models

Requirement	Description
Scalability	Ability to work with graphs containing a large number of vertices and edges
	Efficient memory usage and optimisation of computing resources
Structure flexibility	Support for different types of links
	Ability to dynamically update the graph without the need for its complete reconstruction
Interactive visualisation capability	Fast rendering and updating of graphical data representation
	Support for scaling and changing the level of detail in interactive analysis
Computational efficiency	Optimised algorithms for pathfinding, clustering, and centrality analysis
	Support for parallel computing for working with large graphs
Ability to integrate with other systems	Compatibility with databases and data processing platforms
	Using the application programming interface (API) to interact with other analytics and visualisation tools
Quality of data and algorithms	Methods for cleaning and normalising data before plotting graphs
	Using AI to optimise graph data structuring and analysis

Source: compiled by the authors

Methods for plotting graphs to represent complex relationships in BD varied depending on the nature of the data and the purpose of the analysis. One method is to use deterministic graphs, such as connectivity graphs or tree graphs, which are used to represent structures with defined and stable relationships between elements. Such graphs can be useful in modelling technical or logistics systems, where the relationships between elements usually do not change randomly. Another approach is to build dynamic graphs in which the relationships between vertices change over time. This method is useful for mapping the evolution of networks or systems where relationships change depending on certain events or conditions. Such graphs are often used in economic, environmental, or social research.

Another important part of plotting graphs is determining how to visualise them. Using interactive visualisation tools allow creating visual models that simplify the analysis of relationships in large amounts of data. Graph visualisation helps to see the structure and also identify important nodes, central elements, and hidden patterns in networks. Moreover, a random Erdos-Renyi graph was a classic example in which each pair of vertices was connected by an edge with a certain probability. This type of graph was well suited for modelling systems with random connections found in areas such as social networks, biological systems, or the Internet. An example of VS Code code that uses Python libraries to build a random graph of the Erdos-Renyi type and its interactive visualisation is shown (Fig. 1).

graph_visualization.py > ...

```
import networkx as nx
from pyvis.network import Network

G = nx.erdos_renyi_graph(n=1000, p=0.01)

net = Network(notebook=True)
net.from_nx(G)

net.force_atlas_2based()

net.show("large_graph.html")
```

Figure 1. Code for constructing a random Erdos-Renyi graph in VS Code

Source: compiled by the authors

In the file graph_visualisation.py code has been written to create and render a large graph using Python libraries. Networkx creates a random Erdos-Renyi graph, where each pair of 1,000 vertices connects with a certain probability, which is determined by the parameter p = 0.01. This allows creating a graph where the probability of coupling between any two vertices is 1%. Next, the pyvis library is used to visualise this graph, which creates an interactive graph view that makes it easier to analyse the data structure. The force_atlas_2based method uses an algorithm that organises the placement of vertices in space in such a way as to avoid overlapping them and ensure easy perception. Next, the result is saved in HyperText Markup Language (HTML) format to the large_graph.html file that can be opened in the browser for viewing (Fig. 2).

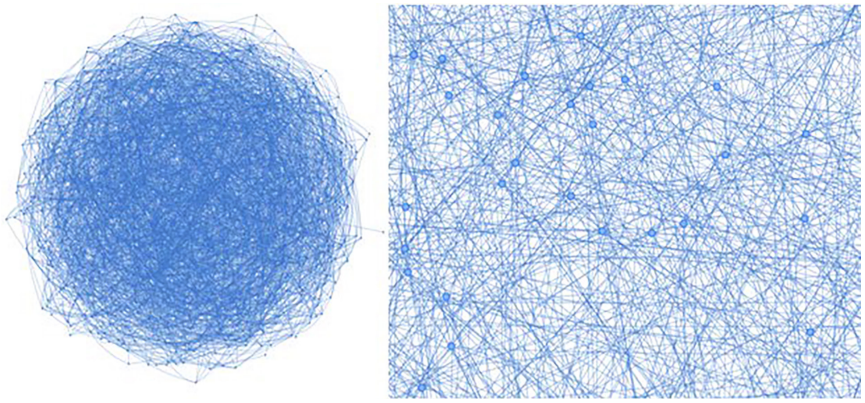

Figure 2. Graph of the Erdos-Renyi type in the large_graph.html file
Source: compiled by the authors

In other words, the code created an Erdos-Renyi graph, which is a classic example of a random graph. In such a graph, each pair of vertices is connected by an edge with a certain probability. This allows modelling structures where relationships between elements can be random, which is useful when analysing large and complex networks. In addition, visualisation is implemented in an interactive format that allows the user to scale the graph, move vertices, explore connections, and improve network structure analysis. This approach allows transfering the concept of random graphs to the Unity environment, using the capabilities of three-dimensional visualisation and interactive control. Unlike the Python implementation, where graphs are created and analysed as web interfaces, Unity creates a 3D model that can be rotated, scaled, and modified in real time (Fig. 3).

```
bly-CSharp                    GraphGenerator                    Start()
1   using UnityEngine;
2   using System.Collections.Generic;
3
    Скрипт Unity | Ссылок: 0
4   public class GraphGenerator : MonoBehaviour
5   {
6       public int nodeCount = 500;
7       public float connectionProbability = 0.01f;
8       public float graphRadius = 10f;
9
10      private List<GameObject> nodes = new List<GameObject>();
11      private List<LineRenderer> edges = new List<LineRenderer>();
12
        Сообщение Unity | Ссылок: 0
13      void Start()
14      {
15          GenerateGraph();
16      }
17
18
        Ссылок: 1
19      void GenerateGraph()
20      {
21
22          for (int i = 0; i < nodeCount; i++)
23          {
24              GameObject node = GameObject.CreatePrimitive(PrimitiveType.Sphere);
25              node.transform.position = new Vector3(Random.Range(-graphRadius, graphRadius),
26                                                    Random.Range(-graphRadius, graphRadius),
27                                                    Random.Range(-graphRadius, graphRadius));
28              node.transform.localScale = new Vector3(0.2f, 0.2f, 0.2f);
29              nodes.Add(node);
30          }
```

Figure 3. Code snippet for generating and visualising a random Erdos-Renyi graph in Unity (Visual Studio)
Source: compiled by the authors

The programme created a graph in the Unity environment, using SPHERE objects for nodes and LineRenderer for connections between them. At the beginning, the Start() method is executed, which calls the GenerateGraph() function. It starts with creating nodes: each node is a sphere that is located within a certain radius at random coordinates in three-dimensional space. All nodes are added to the nodes list. After creating nodes, the programme passes through each pair of nodes and, with a certain probability, creates an edge between them. To do this, a LineRenderer object connecting two points in space is created. All edges are stored in the edges list so that they can be changed or deleted if necessary. After running the script in Unity, 500 spheres appear in the scene, which are randomly located in space and connected by white lines that form a graph. The probability of connecting nodes is set to 0.01, so each node has a small number of connections, which brings the graph closer to the Erdos-Renyi model. The result visualises a random graph in three-dimensional space, which can be used for further analysis, modelling, or interactive interaction in Unity (Fig. 4).

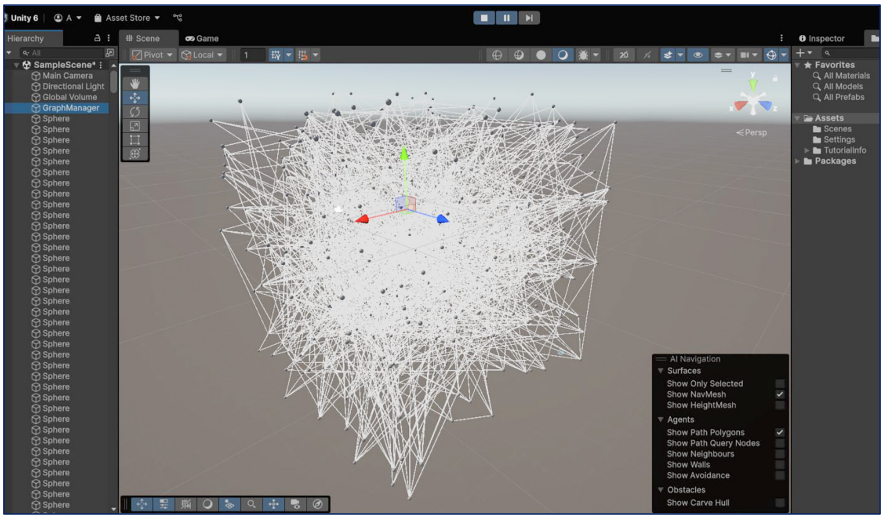

Figure 4. Code snippet for generating and visualising a random Erdos-Renyi graph in Unity (Visual Studio)
Source: compiled by the authors

In general, real-time visualisation of BD graph structures is an important task for understanding complex relationships between objects and analysing their dynamics. Unity Engine provides a powerful toolkit for interactive representation of graphs in three-dimensional space, which allows viewing the model statically, but also actively interacting with it, changing angles, scales, and exploring individual nodes and connections. Unlike conventional graph visualisation methods, which are presented as static images, Unity allows implementing a full-fledged three-dimensional scene with the possibility of interactive interaction. However, the use of browser technologies, in particular, pyvis in Python, provides a convenient way to display large graphs with dynamic updates and the ability to integrate AI algorithms. A comparison of these approaches is provided for better understanding (Table 2).

Table 2. Comparison of graph rendering in Unity and the browser environment

Parameter	Unity (3D visualisation)	Browser environment (pyvis in Python)
Interactivity	High: ability to rotate, zoom, and select nodes in 3D space.	High: supports zooming, node selection, and dynamic animations in 2D.
Visual complexity	Allows full-fledged 3D visualisation, complex effects, and lighting.	Supports interactive graphs, but in 2D, which sometimes makes it difficult to understand.
Performance	High for optimisation, but requires a powerful GPU (Graphical Processing Unit).	Efficient operation in the browser, especially for large graphs.
Scalability	Limited by the number of objects in the scene and GPU performance.	Works well with large graphs, especially with optimised algorithms.
Easy to implement	Requires the use of C# and working with 3D objects via the Unity API, which complicates development.	Easily implemented through libraries such as networkx and pyvis.
Ability to update the graph	Supports dynamic structure changes, but requires additional code.	Easily updated by refreshing the page or calling Python scripts.
Integration with other systems	Limited: it is more difficult to connect APIs and external databases.	Easily integrates with databases, ML, and AI algorithms.
Availability	Unity requires installation, is more difficult to distribute, but can be built for different platforms.	Can be saved and shared as HTML files.
Scope of application	Well suited for game applications and simulations.	Optimal for analytics, research, and working with BD.

Source: compiled by the authors

Unity is therefore an effective solution for creating 3D visualisations, but for working with large graph structures in real time, it is more appropriate to use browser technologies that provide convenience, scalability, and integration with AI techniques. In addition, browser-based tools such as pyvis make it easy to interact with graphs without the need to install additional software, making it easier for a wide range of users to access the visualisation. Unity provides more opportunities for detailed processing of 3D graphs, including complex animation effects and physical interactions between nodes, and Assembly for different platforms is possible. Therefore, the choice of tool depends on the specific requirements for interactivity, performance, and complexity of processing the graph structure.

Optimisation of graph processing and visualisation using computational intelligence

Visualising large graph structures is a complex task, as it requires simultaneous consideration of scalability, interactivity, and efficient use of computing resources. The use of AI techniques allows optimising this process, reducing computational costs, and improving the quality of representation of graph models. A programme that creates a graph using the Barbashi-Albert model is often used to simulate real networks, such as social networks or the Internet (Fig. 5). It generates a graph with a thousand nodes, where each new node is connected to the existing ones according to a certain scheme. In order to process this graph and perform analysis, the programme converts it to an adjacency matrix, which allows working with it as numeric data.

```python
import networkx as nx
from pyvis.network import Network
import numpy as np
from sklearn.decomposition import PCA
from sklearn.cluster import KMeans

G = nx.barabasi_albert_graph(n=1000, m=5)

adj_matrix = nx.to_numpy_array(G)

pca = PCA(n_components=16)
embeddings = pca.fit_transform(adj_matrix)

num_clusters = 5
kmeans = KMeans(n_clusters=num_clusters, random_state=42, n_init=10)
labels = kmeans.fit_predict(embeddings)

net = Network(notebook=True, height="750px", width="100%", bgcolor="white", font_color="black")

colors = ["red", "blue", "green", "yellow", "purple"]
for i, node in enumerate(G.nodes()):
    net.add_node(node, label=str(node), color=colors[labels[i] % len(colors)])

for edge in G.edges():
    net.add_edge(edge[0], edge[1])

net.force_atlas_2based()

net.show("ai_graph_visualization.html")
```

Figure 5. Code for constructing the Barbashi-Albert graph in VS Code
Source: compiled by the authors

Using ML methods, in particular PCA, the programme reduces the dimension of graph data. This allows simplifying the graph structure for further analysis. Then, using the K-Means algorithm, the graph nodes are grouped into several groups, which allows distinguishing similar nodes by their connections and interactions. The result is a visual graph visualisation that allows seeing the relationships between nodes and their clustering in real time (Fig. 6).

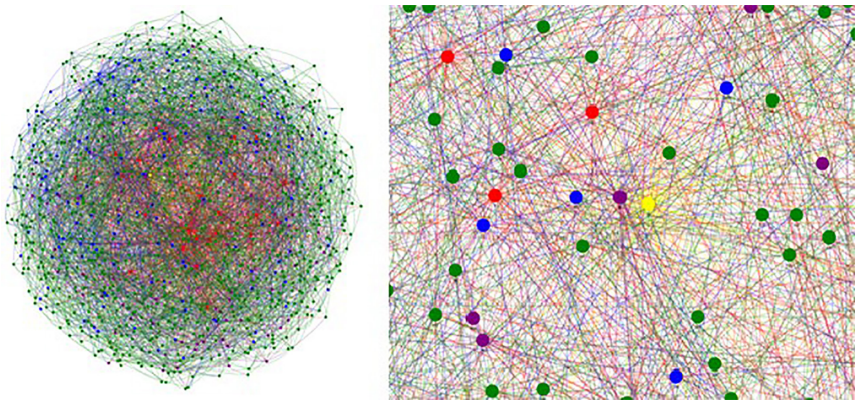

Figure 6. Visualisation of the Barbashi-Albert graph using ML methods
Source: compiled by the authors

The programme uses the ForceAtlas2 physical algorithm to organise nodes in graph space based on physical forces, which makes visualisation more understandable. The result is saved as an HTML file, and the visualisation is interactive, allowing the user to interact with the graph. An important aspect is that this application is based on ML methods, in particular, PCA and K-Means Clustering. It assigns a separate colour to each cluster, which helps to clearly distinguish groups of nodes.

Another AI method for visualising graph models should indicate dimensionality reduction. It allows representing complex graph structures in fewer dimensions, which significantly reduces the computational cost of visualisation. For example, the t-Distributed Stochastic Neighbour Embedding (t-SNE) method allows grouping nodes in social networks by community, and Uniform Manifest Approximation and Projection (UMAP) is used to analyse biological networks. Laplacian Eigenmaps, which work effectively with geospatial data, or DeepWalk and Node2Vec, node vectorisation methods used in recommendation systems, are also used for graph structures. An example of a PCA method for graph visualisation is shown in Figure 7.

```python
import networkx as nx
import numpy as np
from sklearn.decomposition import PCA
from pyvis.network import Network

G = nx.barabasi_albert_graph(n=500, m=3)

adj_matrix = nx.to_numpy_array(G)

pca = PCA(n_components=2)
embeddings = pca.fit_transform(adj_matrix)

net = Network(height="750px", width="100%", notebook=True)
for i, (x, y) in enumerate(embeddings):
    net.add_node(i, x=x, y=y, label=str(i))

for edge in G.edges():
    net.add_edge(edge[0], edge[1])

net.show("pca_graph.html")
```

Figure 7. Code for plotting a graph using the PCA method in VS Code
Source: compiled by the authors

This application creates a scalable network based on the Barbashi-Albert model, which displays real networks with nodes with different numbers of connections (the "rich get richer" principle). A graph is generated with 500 nodes, where each new node joins 3 existing ones. Next, an adjacency matrix is obtained from the graph, which displays the connections between nodes in the form of a numerical matrix. Since such a matrix has a high dimension, the PCA method is used to reduce the dimension to two dimensions, which allows representing a graph in 2D space. Using the pyvis library, an interactive graph visualisation is created. Nodes get coordinates according to the calculated PCA values, and the relationships between them are added according to the original graph structure. After executing the programme, an HTML file is generated, which can be opened in a browser to view the interactive network (Fig. 8).

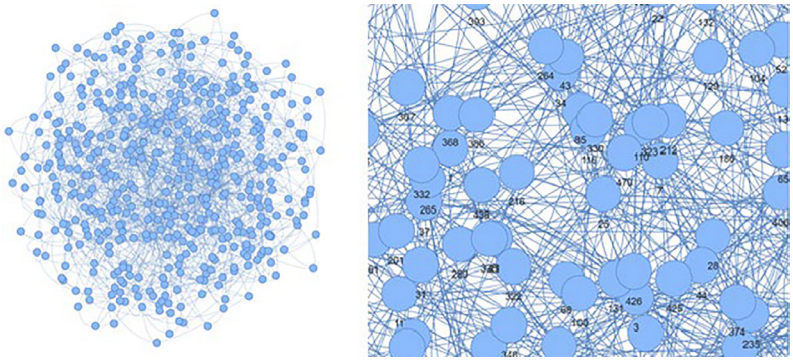

Figure 8. Graph visualisation for the data dimensionality reduction method
Source: compiled by the authors

As a result of the programme execution, an interactive visualisation of the graph is obtained, where nodes are located in two-dimensional space in accordance with the principal components defined by the PCA method. This allows simplifying the display of a complex network structure, while preserving the basic relationships between

nodes. Interacting with the graph in the browser allows exploring its structure, analysing clusters, and identifying the key nodes that have the most connections.

The next method is to optimise the location of nodes in space. Physical force algorithms such as ForceAtlas2 or Fruchterman-Reingold simulate physical interactions between vertices, providing a natural graph layout. However, for large graphs, these methods can be computationally expensive. To speed up calculations, used Barnes-Hut Simulation, which reduces the complexity of calculations for nodes. Additionally, ML techniques are used, such as graph Neural Networks (GNNs), which allow training optimal node representations, or Autoencoders, which help to reduce the dimension of space and improve the location of the graph. An example is the use of a neural network to optimise the location of nodes in space (Fig. 9).

```python
embeddings = np.random.rand(500, 2)

embeddings = torch.tensor(embeddings, dtype=torch.float32, requires_grad=True)

class GraphAutoencoder(nn.Module):
    def __init__(self):
        super(GraphAutoencoder, self).__init__()
        self.encoder = nn.Linear(2, 16)
        self.decoder = nn.Linear(16, 2)

    def forward(self, x):
        x = torch.relu(self.encoder(x))
        x = self.decoder(x)
        return x

model = GraphAutoencoder()

optimizer = optim.Adam(model.parameters(), lr=0.01)
loss_function = nn.MSELoss()

for epoch in range(200):
    optimizer.zero_grad()
    output = model(embeddings)
    loss = loss_function(output, embeddings)
    loss.backward()
    optimizer.step()

optimized_embeddings = output.detach().numpy()
```

Figure 9. Code snippet for building an autoencoder graph in VS Code
Source: compiled by the authors

The purpose of this code is to improve the placement of graph nodes on the 2D plane by training an autoencoder that corrects their coordinates. This can be useful if there is a need for a convenient visual representation of large graphs, where it is important that nodes do not overlap and are distributed clearly and efficiently. The programme starts by generating random coordinates for 500 nodes representing a graph in 2D space. These coordinates are converted to tensors for further work with them in PyTorch. Then a neural network is created – an autoencoder, which consists of two layers. First layer reduces the dimension of coordinates from 2 to 16, and the second one restores them to 2D space. During training, the network tries to minimise the difference between the original coordinates and the ones it generates, thereby improving their location.

Training lasts 200 stages, at each of which the network updates its parameters so that the coordinates of nodes become more optimised for better visualisation. As a result, optimised node coordinates are obtained, which can be used to plot the graph. After that, the programme creates an interactive visualisation of the graph using the pyvis library, where nodes are placed according to optimised coordinates, and relationships between them are added to create a regular graph structure (Fig. 10).

Figure 10. Example of a graph for a method for optimising the location of nodes
Source: compiled by the authors

All nodes are connected by edges, which shows the relationships between graph elements. The resulting file can be viewed in a browser, which helps to visualise and interact with the graph in real time. This approach can significantly

improve the visual appeal and interpretability of complex graphs, especially when working with large data sets, where the clarity and clarity of Node locations is important.

For its part, the method of clustering graph structures for improved visualisation is important. It consists in splitting the graph into separate groups (clustering), which allows analysing and visualising it more efficiently. The main algorithms include: K-Means – a simple method for dividing vertices into clusters; Louvain – for finding communities in graphs; Spectral Clustering – application of linear algebra methods for clustering. An example of using Louvain for graph clustering is shown in Figure 11.

```python
import networkx as nx
import community
import matplotlib.pyplot as plt

G = nx.erdos_renyi_graph(500, 0.05)

partition = community.best_partition(G)

colors = [partition[node] for node in G.nodes()]

plt.figure(figsize=(10, 10))
nx.draw(G, pos=nx.spring_layout(G, seed=42), node_color=colors, cmap=plt.cm.Set1, with_labels=False, node_size=50)
plt.title("Кластеризація графа методом Louvain")
plt.show()
```

Figure 11. Clustering a graph using the Louvain method in VS Code
Source: compiled by the authors

This code demonstrates graph clustering using the Louvain algorithm, which is used to find communities in graph structures. The main idea of the method is to identify subgroups of nodes that have strong internal connections, but relatively weak connections to other parts of the graph. This helps to better visualise and analyse complex networks. First, a random graph is created using the Erdos-Renyi model, where each of the 500 nodes has a probability of 0.05 being connected to other nodes. The Louvain algorithm then splits the graph into clusters, returning a dictionary where each node corresponds to a specific cluster. Based on this cluster breakdown, each node is assigned a colour according to its group. The visualisation uses the location of nodes determined by the spring_layout algorithm, which simulates the repulsive and attractive forces between vertices, creating a more understandable and ordered image. The graph is drawn with coloured cluster designations, which clearly shows how nodes are grouped together. The result is an image of a clustered network, where each colour represents a separate community (Fig. 12).

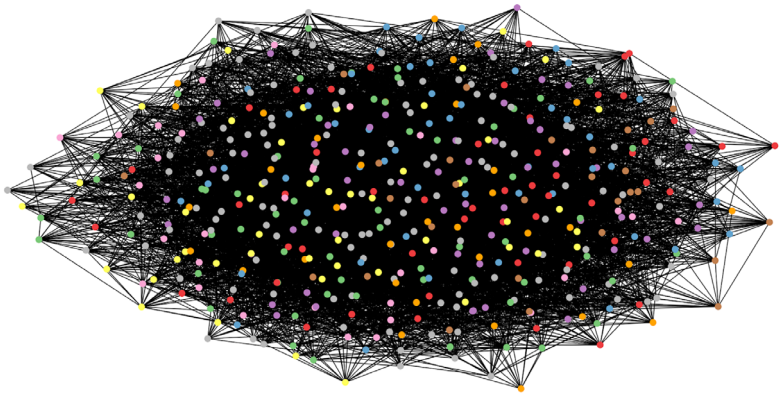

Figure 12. Visualisation of clustering of graph structures
Source: compiled by the authors

In addition to the approaches considered, several methods for speeding up the rendering of large graphs that allow efficient visualisation of complex structures using the GPU and parallel computing can be distinguished. For example, the Rapids cuGraph or Deep Graph Library (DGL) libraries, migrate calculations to TensorFlow/PyTorch, and the Web Graphics Library (WebGL), such as three.js or deck.gl. There is also an approach to predicting changes in dynamic graphs, where Graph Convolutional Networks (GNNS) are used to predict the appearance of new connections or changes in edge weights. This method includes GCN, Graph Attention Networks (GAT), and Temporal Graph Networks (TGNs).

As a result of applying computational intelligence techniques to graph model visualisation, a significant improvement has been achieved in the efficiency of processing large graphs in real time. The use of algorithms such as PCA and K-Means Clustering helped to reduce the dimension of data and classify nodes, which helped to analyse and visualise them. In addition, optimising the location

of nodes using physical algorithms and neural networks helped to achieve better graph structure, which increased the interpretability of complex networks. Louvain clustering revealed communities in graphs, which improved visualisation and analysis of relationships between network elements. Overall, the integration of AI into graph visualisation processes has made them more dynamic, understandable, and efficient for BD analysis.

Discussion

The conclusions of this study are consistent with the findings of G. Gheorghe & P. Lorenz (2023), as they point to the use of AI to work with BD. However, if the researchers was focused on intelligent classification methods, high-performance calculations, and mathematical modelling of BD, then the current study focuses on interactive visualisation of BD graph structures in real time and optimisation of their display using ML algorithms. In particular, it used clustering methods, dimensionality reduction, and optimisation of the location of nodes in space using autoencoders, which allows increasing the efficiency of visual representation of complex graph models. The results of this study also complemented the findings of A.H. Gandomi *et al.* (2023), because they used AI techniques to analyse BD. However, while this article focused on using ML, DL, and natural language processing (NLP) algorithms to identify patterns and make decisions, the current study focuses on AI methods for plotting BD graphs in real time. For this purpose, classification and clustering are applied, including integration of neural networks, which allow automatically adjusting the location of nodes to improve the convenience and accuracy of data representation in three-dimensional space.

The results obtained confirm the conclusions of L. Di & E. Yu (2023), because they emphasise the importance of visualisation when working with BD. While the researchers focused on data storage and processing using structured query language (SQL) and NoSQL systems, and graph processing, the current study focused on AI techniques to improve node placement in graphs. This reduces computational costs and improves the accuracy of visualising complex graph models. In addition, in contrast to the mentioned study, the research used ML algorithms to improve data structuring, which contributed to a more accurate representation of the relationships between BD elements. Overall, this study focused on visualising graph structures, which partially confirms the findings of S. Panyaram (2024), who also used AI techniques for dynamic data, but with a focus on finance and healthcare. That is, the current study focused on improving the visual representation of graphs using neural networks for more accurate and convenient location of nodes, which is the main difference from the mentioned study.

In this article, PCA was used to reduce the graph dimension before clustering, which simplified analysis and improved visualisation. The Barbashi-Albert model, the KMeans algorithm for grouping nodes, and an autoencoder for optimising their location were used. A. Noshi & S. Gasmi (2024) also applied PCA, but in the context of improving the structure of knowledge maps, focusing on NLP methods and semantic analysis. In contrast to their research, this study was aimed at Interactive optimisation of graphs in real time using AI, which allows not only to improve the structure of graph models, but also to automatically adjust the location of nodes, reducing computational costs and improving the accuracy of visualisation. In addition to generating and interactively optimising Barbashi-Albert graphs using ML, Erdos-Renyi graphs were also implemented in the study. Clustering, dimensionality reduction, and autoencoders were used to improve visualisation, which increased the accuracy and efficiency of node placement. In turn, L. Bellmann *et al.* (2024) presented an attribute associative graph for analysing medical data, which is based on statistical metrics and does not require programming. Their approach focuses on fixed determination of relationships between sample attributes, while in the current study, graph structures are dynamically adapted to data using AI techniques, which provides more flexible and accurate visualisation of complex relationships.

Similar to the study by J. Zhao *et al.* (2024), which improved dimensionality reduction algorithms (t-SNE and UMAP) by combining their aspects for data analysis, the current study also applied dimensionality reduction, but with the aim of optimising the visualisation of graph models. In the mentioned study, t-SNE and UMAP were used to cluster nodes in social and biological networks, while the current study also analyses other dimensionality reduction techniques, in particular PCA, which allow for more efficient visualisation of complex graph structures while reducing computational costs. It should also be noted that this study focused on optimising the visualisation of BD graphs using AI techniques that can reduce computational costs and improve the efficiency of visualising complex graph structures. S. Janicijevic & V. Nikolic (2021) focused on interaction metrics such as degree centrality, approximation, internode centrality, and PageRank to evaluate graph visualisation, and the evolution of BD visualisation techniques. Although these approaches require special attention to the specific properties of graph structures, the methods used in the current study provide more efficient management of computational resources for visualising large graphs.

The results demonstrate improvements in the efficiency of visualisation of large graph structures and optimisation of their processing using AI methods and ML algorithms, in particular, in the context of clustering using the Louvain algorithm and using an autoencoder for node placement. They are consistent with the findings of M. Yang (2024), where the integration of AI and BD analysis technologies for cybersecurity provides similar stability and efficiency benefits. That is, both approaches confirm the importance of AI in improving data processing and security. On the other hand, D. Zion & B. Tripathy (2020) focused on the importance of using specialised tools for visualising large graphs, highlighting the limitations of conventional systems when working with dynamic and large data sets.

Thus, in the current study, specific AI methods are used to optimise the visualisation of graphs in real time, which allows dynamically adjusting the location of nodes, and the above research is more focused on general visualisation tools, without focusing on interactive optimisation and the use of ML to improve the display of graph structures.

The current results are consistent with the approaches described by M. Devi & S.R. Kasireddy (2019), because the use of graph models in the context of BD analysis is an important aspect for achieving business goals. As in their study, the current study shows that the use of graph tools can effectively identify hidden relationships between data, in particular in social networks, which is useful for improving marketing and sales strategies. The results of this study also indicate the efficiency of clustering and dimensionality reduction. Comparing the study by T. Kliestik *et al.* (2024), similarities in the use of technologies to improve the analysis of large amounts of data can be noted. However, while the current study focuses on graph model visualisation, the aforementioned research focuses on the use of digital doubles and forecasting in an industrial environment. Although both studies use BD to optimise processes, the results of the current study confirm the potential of real-time interactive visualisation, which can be applied to improve link analysis in BD and support decision-making in various areas.

It is worth noting that the current study developed programmes for plotting graphs in Python using ML algorithms, and for visualising them in real time in Unity using C#. On the other hand, M.A. Ruiz Estrada (2023) examines the use of AI and multidimensional graphs for modelling socio-economic processes, where 3D visualisation models are also used. Both approaches integrate AI to improve BD processing, but while the current study focuses on real-time graph visualisation, the researcher focuses on using graphs for policy modelling and crisis analysis. Therefore, the results of both studies are consistent, AI is used for BD analysis but with a focus on policy modelling. In addition, the current study focuses on interactive visualisation of large graph structures, in particular, through the integration of AI techniques to optimise node placement and reduce computational costs. Compared to the study by F. Gebretsadik & R. Patgiri (2023), which examines common problems of large graphs such as visualisation, data distribution, and knowledge integration, the current study focuses on optimising graph processing through computational intelligence techniques and their implementation in various environments, such as Unity Engine. This approach provides improved interactivity and visualisation efficiency through the use of neural networks to optimise graphs.

Similar to the research by S.K. Devineni (2024), which examines interactive data visualisation using AI, in particular ML and virtual/augmented reality (AR/VR) applications, the current study focuses on interactive graph visualisation using AI, but in the context of optimising the processing of large graphs in real time. This study uses clustering techniques, dimensionality reduction, and neural networks to arrange nodes, which allows for greater efficiency in graph processing. While this article focuses on multidimensional data and the use of AR/VR, the current results focus more on the technical implementation of graph structure visualisation and optimisation for interactive BD work. Thus, the results of this study make an important contribution to understanding the application of AI techniques to improve the visualisation of BD graph structures and can be useful for further developments in the field of BD analysis and the creation of effective interfaces for processing and visualising such data.

Conclusions

During the study, significant results were achieved in creating and visualising graph structures. The ability to generate random Erdos-Renyi graphs and use them to analyse relationships between data has been demonstrated. For this purpose, an approach to graph construction in Python in VS Code has been developed, which allows generating graphs for analysing relationships in BD, and their 3D visualisation in Unity has also been implemented. The results confirmed the effectiveness of using Unity and Python for visualising large graph structures, improving the quality of interfaces for data analysis.

The use of various AI techniques, including ML algorithms, has improved the efficiency and accuracy of graph processing in research. The constructed Barbashi-Albert graph demonstrated the clustering of nodes and their relationships in real time. To reduce the dimension of graphs, PCA and autoencoder methods were used, which reduced the complexity of visualising high-dimensional graphs. Further clustering of nodes using the Louvain algorithm helped to identify communities in graphs, which significantly improved their interpretation and helped to achieve better visualisation results. Moreover, the use of neural networks has helped to optimise the location of nodes, reducing computational complexity and improving real-time accuracy. This contributed to the high interactivity required for real-time visualisation. Based on the use of various Python libraries, such as networkx, pyvis, torch, and scikit-learn, powerful tools for graph analysis and visualisation were created, which provided significant improvements in working with BD. During the study, graphs were generated using random models, such as the Erdos-Renyi and Barbashi-Albert models, which limited the ability to test methods on real data sets.

Further research should focus on optimising and improving parallel processing methods for scaling graph models. In addition, it is advisable to consider using more complex neural networks to improve the accuracy of node placement and testing on complex graphs. It is also worth studying the integration of other AI methods to improve clustering and predict the dynamics of graph structures in real time, which will significantly improve the practical application of the results obtained.

Acknowledgements

None.

Funding

The study received no funding.

Conflict of Interest

None.

References

[1] Bellmann, L., Wiederhold, A.J., Trübe, L., Twerenbold, R., Ückert, F., & Gottfried, K. (2024). Introducing attribute association graphs to facilitate medical data exploration: Development and evaluation using epidemiological study data. *JMIR Medical Informatics,* 12, article number e49865. doi: 10.2196/49865.

[2] Çınar, D. (2024). The role of artificial intelligence and big data analytics in business management: A review of decision – making and strategic planning. *Journal of Tourism Economics and Business Research,* 6(2), 219-229.

[3] Devi, M., & Kasireddy, S.R. (2019). Graph analysis and visualization of social network big data. In *Social network forensics, cyber security, and machine learning. springerbriefs in applied sciences and technology* (pp. 93-104). Singapore: Springer. doi: 10.1007/978-981-13-1456-8_8.

[4] Devineni, S.K. (2024). AI-enhanced data visualization: Transforming complex data into actionable insights. *Journal of Technology and Systems,* 6(3), 52-77. doi: 10.47941/jts.1911.

[5] Di, L., & Yu, E. (2023). Big data analytic platforms. In *Remote sensing big data* (pp. 171-194). Cham: Springer. doi: 10.1007/978-3-031-33932-5_10.

[6] Duan, S., & Zhao, Y. (2024). Knowledge graph analysis for chronic diseases nursing based on visualization technology and literature big data. *Scalable Computing Practice and Experience*, 25(3), 1728-1747. doi: 10.12694/scpe.v25i3.2664.

[7] Gandomi, A.H., Chen, F., & Abualigah, L. (2023). Big data analytics using artificial intelligence. *Electronics,* 12(4), article number 957. doi: 10.3390/electronics12040957.

[8] Gebretsadik, F., & Patgiri, R. (2023). The major challenges of big graph and their solutions: A review. *Advances in Computers,* 128, 399-421. doi: 10.1016/bs.adcom.2021.10.010.

[9] Gheorghe, G., & Lorenz, P. (Eds.). (2023). *Proceedings of the 3rd international conference on artificial intelligence, big data and algorithms. Advances in Artificial intelligence, big data and algorithms*. Amsterdam: IOS Press BV.

[10] Haiyang, X., Ruomei, Y., Yan, W., Lixin, G., & Li, M. (2024). Knowledge graph for solubility big data: Construction and applications. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery,* 15(1), article number e1570. doi: 10.1002/widm.1570.

[11] Janicijevic, S., & Nikolic, V. (2021). Graph structures for data visualizations. *Serbian Journal of Engineering Management*, 6(2), 24-31.

[12] Kliestik, T., Kral, P., Bugaj, M., & Ďurana, P. (2024). Generative artificial intelligence of things systems, multisensory immersive extended reality technologies, and algorithmic big data simulation and modelling tools in digital twin industrial metaverse. *Equilibrium. Quarterly Journal of Economics and Economic Policy,* 19(2), 429-461. doi: 10.24136/eq.3108.

[13] Liu, J., Xie, R., & Song, A. (2019). Analysis on research frontiers and hotspots of "Artificial intelligence plus education" in China – visualization research based on citespace V. *IOP Conference Series Materials Science and Engineering,* 569, article number 052073. doi: 10.1088/1757-899X/569/5/052073.

[14] Miroshnyk, M., Shkil, O., Rakhlis, D., Pshenychnyi, K., & Miroshnyk, A. (2023). Event processing model for simulation of real-time logic control devices. *Bulletin of Cherkasy State Technological University*, 28(2), 50-57. https://doi org/10.24025/2306-4412.2.2023.274840.

[15] Mykolaichuk, R., & Mykolaichuk, A. (2024). Using artificial intelligence technologies for document processing automation. *Modern Information Technologies in the Sphere of Security and Defence,* 50(2), 111-117. doi: 10.33099/2311-7249/2024-50-2-111-117.

[16] Noshi, A., & Gasmi, S. (2024). Building efficient knowledge maps through NLP and optimization in big data environments. doi: 10.13140/RG.2.2.11353.53609.

[17] Pankiv, V.I., & Storozhuk, O.L. (2024). The use of neural measurements and advanced techniques for the analysis of large quantities of data in real time. In *Forestry education and science: Current challenges and development prospects, international science-practical conference*. Lviv: National Polytechnic University of Ukraine. doi: 10.36930/conf150.5.19.

[18] Panyaram, S. (2024). Integrating artificial intelligence with big data for real-time insights and decision-making in complex systems. *FMDB Transactions on Sustainable Intelligent Networks,* 1(2), 85-95. doi: 10.69888/FTSIN.2024.000211.

[19] Riva, D., & Rossetti, C. (2024). Visualization of knowledge graphs with embeddings: An essay on recent trends and methods. *ArXiv*. doi: 10.48550/arXiv.2412.05289.

[20] Ruiz Estrada, M.A. (2023). *New artificial intelligence (AI) models for policy modelling*. Kuala Lumpur: Econographication Laboratory. doi: 10.13140/RG.2.2.30920.90887.

[21] Soroka, I.M., Mochalov, IO., & Kizim, A.V. (2024). The modern directions of implementation of the large models of artificial intelligence in health care. *Intermedical Journal,* 2, 174-180. doi: 10.32782/2786-7684/2024-2-30.

[22] Yang, J. (2024). Application of artificial intelligence and big data in financial management. *SHS Web of Conferences,* 208, article number 01006. doi: 10.1051/shsconf/202420801006.

[23] Yang, M. (2024). Design of a cybersecurity defense system based on big data and artificial intelligence. *Applied and Computational Engineering,* 87, 98-103. doi: 10.54254/2755-2721/87/20241443.
[24] Yin, S., Li, H., Sun, Y., Ibrar, M., & Teng, L. (2024). Data visualization analysis based on explainable artificial intelligence: A survey. *IJLAI Transactions on Science and Engineering,* 2(2), 13-20.
[25] Zhao, J., Pierre, J., & Konstantinidis, K.T. (2024). Approximate nearest neighbor graph provides fast and efficient embedding with applications for large-scale biological data. *NAR Genomics and Bioinformatics*, 6(4), article number lqae172. doi: 10.1093/nargab/lqae172.
[26] Zion, D., & Tripathy, B. (2020). Comparative analysis of tools for big data visualization and challenges. In S.M. Anouncia, H.A. Gohel & S. Vairamuthu (Eds.), *Data visualization* (pp. 33-52). Singapore: Springer. doi: 10.1007/978-981-15-2282-6_3.

Методи штучного інтелекту для візуалізації графових моделей великих даних у реальному часі

Андрій Баник
Аспірант
Ужгородський національний університет
88000, пл. Народна, 3, м. Ужгород, Україна
https://orcid.org/0000-0002-8991-310X
Павло Мулеса
Доктор технічних наук, доцент
Ужгородський національний університет
88000, пл. Народна, 3, м. Ужгород, Україна
https://orcid.org/0000-0002-3437-8082

Анотація. Метою дослідження була розробка підходів до використання штучного інтелекту для покращення процесів інтерактивної візуалізації графових структур великих даних у реальному часі з урахуванням оптимізації обчислювальних ресурсів. Під час дослідження були побудовані графи для аналізу зв'язків у великих даних, а також використані методи обчислювального інтелекту для оптимізації обробки та візуалізації графів в інтерактивному форматі. Результати дослідження включали розробку програм для побудови графових структур на Python у середовищі Visual Studio Code та їх подальшу візуалізацію в Unity з використанням C# у Visual Studio. Спочатку було показано візуалізацію випадкового графа типу Ердеша-Реньї, який потім було відтворено у 3D-просторі Unity. За допомогою бібліотек Python було реалізовано генерацію графів та інтерактивну веб-візуалізацію. Для оптимізації розташування вузлів у графах були використані методи машинного навчання, зокрема, автоенкодери та головні компоненти для зменшення розмірності. Демонстрація моделі Барбаші-Альберта дозволила побачити кластеризацію вузлів та їх зв'язки в реальному часі. Крім того, було продемонстровано інтерактивну візуалізацію, де вузли розташовувалися у 2D-просторі відповідно до результатів аналізу головних компонент. Використання алгоритму Лувена допомогло виконати кластеризацію та візуалізувати структуру спільнот. Результати показали, що використання нейронних мереж значно покращує точність та ефективність розміщення вузлів у графах, а також зменшує обчислювальну складність. Отримані результати можуть бути корисними для наукових досліджень, що пов'язані з аналізом великих графових структур та потребують інтерактивної візуалізації даних

Ключові слова: машинне навчання; інтерактивний формат; кластеризація та головні компоненти; бібліотеки Python; можливості Unity

Journal homepage: https://itce.com.ua/en

Information Technologies and Computer Engineering **Vol. 22, No. 1. 2025**

Article's History: Received: 10.12.2024 Revised: 03.03.2025 Accepted: 24.04.2025

UDC 004.7 DOI: 10.63341/vitce/1.2025.55

Integration of Zero Trust and Blockchain in SDN networks: An overview of threats and methods of their elimination

Oleksandr Pidpalyi*

Postgraduated Student
National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute"
03056, 37 Beresteiskyi Ave., Kyiv, Ukraine
https://orcid.org/0009-0007-6852-7959

Oleksandr Romanov

Doctor of Technical Sciences, Professor
National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute"
03056, 37 Beresteiskyi Ave., Kyiv, Ukraine
https://orcid.org/0000-0002-8683-3286

Abstract. The purpose of the study was to identify theoretically sound methods for integrating Zero Trust and blockchain concepts to improve the overall security of software-defined networks (SDN). The research was based on the development of a theoretical network model that includes an SDN controller, switches, routers, and hosts, which used virtualisation tools such as GNS3, VirtualBox, and Docker. The theoretical basis of the study covered the analysis of key threats, including DDoS attacks, routing manipulation, insider threats, attacks on the application programming interface (API), and specific vulnerabilities of blockchain consensus mechanisms. Simulation scenarios were developed to demonstrate the potential impact of these threats on the security and performance of SDN networks. Analysis of the results obtained theoretically confirmed that the use of Zero Trust policies significantly reduces the risks of insider attacks and improves the protection of the SDN controller due to the principles of constant access verification and micro-segmentation. Integration of blockchain technologies increases the reliability of routing and traffic management, preventing malicious interference in the network infrastructure. Theoretical methods for authentication and verification of requests using blockchain significantly improve the protection of APIs and interaction interfaces. In addition, hybrid consensus algorithms have shown the potential to improve network performance and ensure its resistance to attacks. The study highlighted the importance of integrating Zero Trust and blockchain as an effective solution for eliminating a wide range of threats in SDN networks. This opens up new prospects for the protection of telecommunications systems and lays the theoretical foundation for further research and improvement of security methods. The practical significance of the study is to develop specific recommendations for implementing a comprehensive SDN security system based on blockchain technologies and Zero Trust principles. The proposed solutions can be used both in the public sector to protect critical infrastructure and in the private sector to ensure the security of corporate networks

Keywords: access control; data verification; risk reduction; distributed systems; attack resistance; communication security

Introduction

The rapid development of information technologies creates new challenges for managing network resources. Conventional network management methods are losing effectiveness due to increasing infrastructure complexity, increasing data volumes, and increased cyber threats. The urgency of introducing innovative network management technologies for Ukraine is conditioned by the critical need for modernisation of telecommunications infrastructure.

Suggested Citation:
Pidpalyi, O., & Romanov, O. (2025). Integration of Zero Trust and Blockchain in SDN networks: An overview of threats and methods of their elimination. *Information Technologies and Computer Engineering*, 22(1), 55-68. doi: 10.63341/vitce/1.2025.55

*Corresponding author

In the context of digital transformation of the economy and increasing global information risks, it is necessary to ensure reliable protection of information systems and compliance with international security standards. Blockchain technologies and the Zero Trust concept offer modern mechanisms for improving network security. They allow implementing the principles of decentralised data protection, constant verification of access and minimising the risks of compromising information resources. The introduction of these approaches creates a new paradigm of network security, which is critical for ensuring national information infrastructure and competitiveness in the global digital environment. This leads to a high interest from the scientific community.

Research by X. Guo *et al*. (2022) and W. Li *et al*. (2020) emphasised the importance of software-defined networks in modern telecommunications, emphasising their role in providing flexible network management and traffic optimisation by separating the control plane from the data plane. This architectural feature provides unprecedented flexibility, but at the same time creates new security challenges, in particular, due to the centralised nature of the SDN controller, which becomes vulnerable to DDoS attacks, routing manipulation, insider threats, and attacks on the application programming interface (API). In this context, the integration of blockchain technologies can significantly improve SDN security by decentralising management, which eliminates the risks associated with a single point of failure.

The consensus mechanisms used in the blockchain allow creating a distributed and reliable platform for verifying changes in routing or security policies, reducing the likelihood of successful attacks on the controller. In addition, the blockchain helps to ensure the integrity and immutability of data by storing change logs in network policies, which becomes an effective tool for countering insider threats. According to J.A. Fadhil & S.R. Zeebaree (2024), blockchain, as a decentralised technology, can be used to manage traffic and authenticate data, providing protection against external interference. The use of blockchain to authenticate data to SDN not only ensures its reliability, but also creates a platform for transparent traffic management, which is crucial for countering complex cyber threats. Integration of this technology reduces the risk of unauthorised interference, even in the event of attacks on centralised network elements.

K. Gai *et al*. (2019) in their research presented the concept of differentiated privacy based on blockchain for ensuring the security of the industrial Internet of Things (IoT). The concept of blockchain-based differentiated privacy can be adapted for SDN networks, where data protection at the level of each network node is important. This reduces the risk of confidential information leakage, especially in environments with an increased number of IoT devices. Y. Xu *et al*. (2019) further developed this idea by offering a blockchain-based network computing service scheme with non-consistency. K. Gai *et al*. (2022) proposed a blockchain-based access control scheme to ensure reliable data exchange between organisations in the Zero Trust paradigm. The proposed blockchain-based access control scheme reinforces the Zero Trust paradigm, as it provides authentication at every stage of data exchange. This is especially important for SDN networks, where the absence of a single point of trust is a key element of security.

P. Zheng *et al*. (2023) further highlighted the possibilities of using blockchain technology in the development of decentralised applications, which allows creating a transparent access control system. Such applications contribute to effective traffic monitoring and rapid detection of anomalies, which allows quickly identifying security threats based on blockchain records. An additional layer of security is the integration of the Zero Trust approach, which provides for constant verification of each action or access, regardless of the internal or external status of the user. X. Yan & H. Wang (2020) emphasised that Zero Trust is based on the principle of minimal trust and constant access verification, which significantly increases the network's resilience to insider threats. In addition to increasing resistance to internal threats, Zero Trust integration allows creating a detailed access structure that minimises the risks of network compromise due to privacy or integrity violations. This approach works effectively in conditions of high dynamics of the network environment, in particular in cloud services.

SDN allows dynamically managing real-time access policies, supported by the blockchain to securely store these policies and counteract their substitution. The Zero Trust architecture combined with SDN allows for network micro-segmentation, isolating critical resources from potentially dangerous areas, and the blockchain supports this process, ensuring transparency and resilience to change. In addition, the integration of blockchain identification allows increasing the level of verification of devices and users within the framework of the Zero Trust concept, creating comprehensive protection of the network infrastructure. Thus, the combination of SDN, blockchain and the Zero Trust approach paves the way for creating adaptive, threat-resistant and secure network environments that meet the challenges of the modern world.

Ukrainian researchers have contributed to the investigation of the possibilities of integrating blockchain into telecommunications systems. O. Bykonia & N. Romanovska (2024) noted that the integration of Zero Trust and blockchain can significantly improve the security level of critical infrastructure, help to reduce the risks of cyber-attacks and ensure compliance with international standards. Their proposed blockchain-based computing service scheme helps to increase the resilience of SDN networks to failures through decentralised data storage and the use of smart contracts. This ensures reliable operation of networks even if individual nodes are compromised.

Existing SDN security studies have identified several significant gaps. Firstly, while SDN provides high flexibility in managing network infrastructure, the centralised nature of controller management creates significant

vulnerabilities, particularly in relation to DDoS attacks, routing manipulation, API attacks, and internal threats. Secondly, conventional security management methods often do not meet the requirements of modern dynamic and distributed environments, because they are based on the assumption of trust in internal users and devices. Thirdly, blockchain and Zero Trust technologies are considered separately in the context of various security aspects, which limits their potential for comprehensive solutions to problems in SDN networks.

Given these gaps, the purpose of this study was to develop practical recommendations for integrating Zero Trust concepts and blockchain technology to improve SDN security. The study included the creation of a theoretical model of the SDN network using virtualisation tools; investigation of the possibilities of implementing Zero Trust Principles to protect against insider threats; development of methods for applying blockchain technologies to ensure the integrity and reliability of data.

Materials and Methods

The research methodology was based on a comprehensive analysis of threats to SDN and the study of the capabilities of Zero Trust and blockchain technologies to neutralise detected threats. The main task was to create a methodology that considers the features of decentralisation and strict access control inherent in these technologies to improve the overall sustainability of the network. At the initial stage of the study, a theoretical analysis of the main threats in SDN networks was carried out. Attention was focused on threats related to unauthorised access, attacks on the control and management layers, internal threats and possible compromises of the centralised controller. The potential impact of these threats on the functioning and integrity of SDN networks was studied separately.

The following methods were chosen to develop the integration approach: hierarchical threat analysis, blockchain-based identification analysis, and multi-factor authentication methods typical of Zero Trust. The method of hierarchical threat analysis helped to structure threats by risk levels, which became the basis for determining the necessary counteraction measures. This ensured consistency in the choice of protection methods, which was the basis for developing an effective security model. The blockchain was used as the basis for decentralised storage of access and transaction data, which provided increased transparency and protection against data forgery. This solution is a key to eliminating centralised points of failure. Instead, Zero Trust methods were implemented to dynamically manage access to network resources based on the minimum trust principle, which helped to consider constantly changing access conditions (Dhiman *et al.*, 2024).

Simulation of SDN networks with blockchain and Zero Trust integration was implemented using tools such as GNS3, VirtualBox, Docker, and Python libraries for network analysis. The OpenDaylight SDN controller was used for modelling, which provided interaction between switches and other network components in a virtual environment. When developing the model, the need to implement micro-segmentation principles specific to the Zero Trust concept was considered, which included constant verification of each transaction and access request.

To build the model, three segments of the SDN network were used, each of which had a different access level to ensure the isolation of critical resources from potentially dangerous areas. The blockchain was used to record and verify transactions, and provide transparency and protection against unauthorised changes. Proof of Stake was chosen as the consensus algorithm, which reduced the computational load compared to conventional approaches such as Proof of Work. Several key scenarios were identified for modelling attacks, including DDoS attacks with an intensity of up to 10,000 requests per second, API compromises, and internal threats related to credential leaks.

Three different attack scenarios were modelled to evaluate the effectiveness of the proposed approach in conditions as close to real-life as possible. Scenario 1 involved an attack using social engineering techniques; Scenario 2 – an attack using social engineering techniques and hacking of network components; Scenario 3 – a combined attack that simultaneously uses social engineering techniques, hacking of network components and exploits vulnerabilities in communication protocols. A multi-level approach to threat analysis was implemented, which included identifying early signs of a potential attack, assessing its likely impact on the network, and making quick decisions to neutralise it through an adaptive response system that automatically regulates access to network resources depending on the threat level. The adaptive response system provided the ability to automatically adjust access depending on the threat level. This helped to counteract combined attacks more effectively, considering both internal and external threats.

The process of integrating Zero Trust and the blockchain included creation of a conceptual model of interaction between SDN network components. This model covered steps from identifying potential threats to defining access control methods and creating a transparent access accounting mechanism using the blockchain. To reduce the network load, a hybrid blockchain was chosen, where some of the data was processed centrally, and mission-critical transactions were recorded in the blockchain.

Validation of the approach using a simulation model of the SDN network with integration of Zero Trust and blockchain allowed to empirically confirm the advantages of the proposed solution. The main evaluation criteria were response time, protection from external threats, resistance to internal attacks, and overall support costs. The comparison was carried out with conventional security methods, such as network segmentation, which provided division into separate logical segments, reducing the risks of uncontrolled spread of attacks, and centralised authentication (RADIUS or TACACS+), which controls access to resources using centralised credential verification servers.

Results

Security threats and implementation model of Zero Trust and blockchain in SDN networks

Threat analysis for SDN was a key stage of research aimed at identifying weaknesses in the security mechanisms of these systems. The main threats to SDN networks can be characterised by their impact on the network and the level of risk. Among the main threats were unauthorised access, attacks on the control layer, internal threats, and compromise of the centralised controller. Unauthorised access poses a serious threat to data privacy and the stability of network operations, especially if attackers exploit vulnerabilities in interaction via the API. Remote attacks are aimed at gaining uncontrolled access to network resources. They have a high level of risk because attackers can seize control of critical network components. Control layer attacks, such as DDoS attacks, are critical because this layer is responsible for managing all processes on the network. Their consequences may include significant failures or a complete shutdown of the network infrastructure. Data theft, which is also a critical threat, threatens privacy and can lead to significant financial and reputational losses, especially in cases of working with confidential information. Internal threats that arise as a result of compromising the credentials of privileged users can lead to manipulation of routes or changes in data configurations, causing serious damage to network security. Internal threats that arise from the use of privileged access are difficult to detect. Their level of risk is high because they allow manipulating data and network configurations. Data interception is characterised by an average level of risk and is fraught with leakage of confidential information, which is critical in some industries (Guo *et al*., 2022). Special attention should be paid to the problem of a centralised controller, which in the conventional SDN architecture is a single point of failure. Compromising this element creates the risk of uncontrolled traffic redirection or loss of access to the entire network. In addition, attackers can use combined attacks, combining social engineering, exploiting communication protocol vulnerabilities, and compromising network components. This highlights the need for a multi-level approach to protection that can respond to different types of threats in a complex way. Threat analysis also helps to determine priority countermeasures (Table 1).

Table 1. Main threats to SDN networks and how to eliminate them using Zero Trust and blockchain

Threat	Impact on the network	Risk level	Elimination methods
Remote attacks	Uncontrolled access to network resources	High	Verification of each request, two-factor authentication
Attacks on the central controller	Network failures due to management violations	Critical	Use of blockchain to decentralise controller functions
Data theft	Violation of privacy and reduced trust	Critical	Cryptography, secure access under Zero Trust principles
Internal threats	Use of privileged access	High	Continuous monitoring and segmentation of the network, use of blockchain
Data interception	Leak of confidential information	Moderate	Traffic encoding, segmented access

Source: created by the authors based on X. Guo *et al*. (2022), J. Li *et al*. (2022), S. Ghasemshirazi *et al*. (2023)

Attacks on the central controller and data theft pose the greatest risk, while data interception is one of the less critical threats. To counter these threats, it is recommend using the Zero Trust approach, which provides for constant monitoring and verification of access, regardless of the source of the request. This reduces the risk of both internal and external threats. Additionally, the integration of blockchain technology ensures decentralised management, transaction transparency, and makes it impossible to modify data without authorisation. Such measures allow eliminating major vulnerabilities and ensuring reliable network protection. Blockchain adds a level of transparency, eliminating data forgery and reducing the risk of unauthorised modifications. The integration of artificial intelligence (AI) technologies with Zero Trust and blockchain in SDN networks creates a strong foundation for the development of innovative solutions in the field of cybersecurity, which will allow organisations to effectively counter modern threats and provide reliable protection of critical information infrastructure in the face of the constant evolution of cyber threats.

The conceptual model of interaction between SDN network components was developed to analyse the integration of Zero Trust and the blockchain to ensure a high level of security and performance. The schematic image shows the SDN network architecture, which includes integration of Zero Trust and blockchain technologies. The central component of the system is the SDN controller, which performs network management functions, makes routing decisions, and controls access to network resources. The controller interacts with switches, routers, and end nodes (hosts), providing data exchange. Switches and routers provide data transmission in accordance with the policies set by the controller, and hosts are the sources and consumers of traffic (Fig. 1).

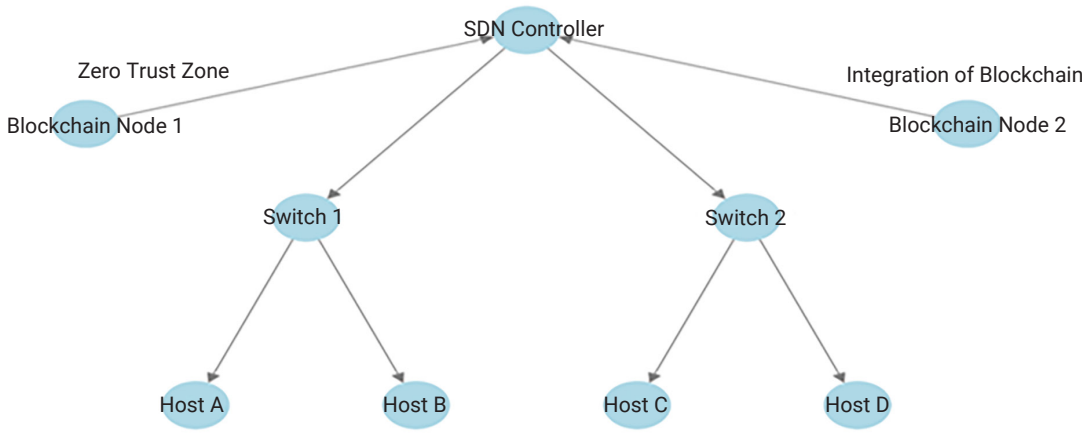

Figure 1. Schematic representation of an SDN network with integration of Zero Trust and blockchain technologies
Source: created by the authors

The model also integrates blockchain and Zero Trust technologies to improve security. Blockchain is used for decentralised storage of access logs and changes in network policies, providing transparency and protection against data forgery. The Zero Trust concept is implemented through multi-level authentication and strict control of each access, regardless of the source. This isolates critical resources from potentially dangerous areas and significantly reduces the risks of internal threats. The image also shows key relationships between components: the SDN controller interacts with the blockchain, passing access policy data for writing to the distributed ledger. Switches are connected to the end nodes (hosts) through which traffic is transmitted, and the blockchain provides control and verification of all transactions. The process of interaction in the model covers the stages from identifying potential threats to creating a transparent access accounting mechanism. Access policies are based on the principles of minimal trust, and authentication takes place at every level, which allows effectively responding to threats and preventing unauthorised actions.

The results of the study showed that using a layered architecture can significantly reduce the load on the network and improve its performance. In particular, this approach allows distributing the processing of different types of transactions or data between individual levels of the system, reducing competition for computing resources. In a layered architecture, the lower levels may be responsible for processing basic requests, while the upper levels focus on critical operations such as authentication or performing smart contracts. This method has proven effective for large corporate and government SDN networks with high bandwidth, where the volume of traffic and the number of users is constantly growing (Bykonia & Romanovska, 2024). Based on the layered structure, it was possible to achieve greater flexibility and stability of the system, and reduce delays in performing key operations. Thus, the proposed approach allows adapting Zero Trust and blockchain to the specific needs of large networks, while maintaining a high level of security and transparency.

The integration of Zero Trust and blockchain creates a new security paradigm that can effectively counter modern threats. Continuous monitoring of access and decentralisation of management ensures a high level of trust in information circulating on the network. These technologies significantly increase the transparency of access, reduce the risk of unauthorised changes, and create an additional layer of protection, especially against attacks on centralised network elements. The blockchain was chosen for decentralised storage of event and transaction logs, which provides transparency and protection against data forgery. The Zero Trust concept aims to constantly monitor and verify access to resources, regardless of the source of the request. This approach allows creating a detailed model of access policies based on the principles of "never trust, always verify" (Liu *et al*., 2020). The combination of Zero Trust with blockchain technology provides additional decentralisation and transparency and allows recording all transactions in a distributed ledger. This eliminates data forgery and reduces the risk of unauthorised network interference.

Model testing: Description of attack scenarios and security system responses

To test and validate the model, two SDN network configurations were created: a conventional security model and a model that integrated Zero Trust and blockchain. The conventional model was based on network segmentation, which provided a division into separate logical segments, reducing the risks of uncontrolled spread of attacks, and centralised authentication (RADIUS or TACACS+), which controlled access to resources using centralised credential verification servers. The model with Zero Trust and blockchain complemented these approaches with modern security principles, such as continuous verification of users and devices, minimisation of access privileges, micro-segmentation to restrict access, transparent logging of all events in the blockchain, and distributed verification of access transactions, which ensured the detection of unauthorised changes in the network. The effectiveness of both configurations

was evaluated in a simulation environment, where real-world attack scenarios were modelled, including social engineering techniques, hacking network components, and exploiting vulnerabilities in communication protocols.

In the first scenario, the attack was based on social engineering techniques. The attacker sent a phishing email to the SDN network administrator with a request to confirm access to the system through the "official portal". After entering their credentials, they were contacted by an attacker who tried to gain access to the SDN controller. A conventional security system was able to detect abnormal activity only after an attacker entered, without identifying the source of the compromise. In turn, the system with the blockchain and Zero Trust blocked access due to multi-factor authentication, and also recorded all access attempts in the blockchain for further analysis.

The second scenario combined social engineering and hacking of network components. An attacker obtained administrator credentials through phishing, and then changed the switch configuration to redirect traffic through the server they controlled. In this case, the conventional system detected anomalies in routing after the compromised traffic reached the attacker. A system with a blockchain and Zero Trust detected the threat earlier, as it required multi-level authorisation to make configuration changes. The request was blocked, and all actions of the attacker were recorded in the blockchain.

The third scenario combined social engineering techniques, hacking network components, and exploiting vulnerabilities in communication protocols. The attacker first obtained administrator credentials, then changed the router settings and entered vulnerable code into protocols, which led to a DoS-type attack and partial network paralysis. The conventional security system detected an attack only after significant failures occurred, and identifying the source of the compromise remained a difficult task. The system with the blockchain and Zero Trust blocked the request to change the protocol in a timely manner due to multi-level access verification. All attempts to make changes, including details of the entered code, were recorded in the blockchain, which helped to quickly isolate the problem and prevent further compromise (Table 2).

Table 2. Comparison of the response of a conventional system and a system with blockchain and Zero Trust

Scenario	Attack type	Conventional system response	System response with blockchain and Zero Trust
1	Social engineering	Partial detection after compromise, source not identified	Blocking unauthorised access, logging in the blockchain
2	Social engineering + compromise	Delay in detecting routing anomalies	Blocking routing changes, transparent logging
3	Social engineering + compromise + protocol exploitation	Inability to respond quickly to a combination of attacks	Detection at all stages through multi-factor and blockchain consensus

Source: created by the authors based on X. Guo *et al*. (2022), W. Li *et al*. (2020), S. Ghasemshirazi *et al*. (2023)

According to the conducted research, the integration of Zero Trust and the blockchain helped to achieve a significantly higher level of security compared to standard security measures. A special feature of this approach was that during the compromise of one of the network elements, the blockchain recorded all historical data of changes, which prevented further expansion of the attack due to the transparency and immutability of records. Zero Trust provided constant control of access to resources based on the principle of "never trust, always verify", regardless of whether this refers to internal or external users of the network. The results of model validation for indicators such as network operation latency, connection stability, access transparency, and threat mitigation are presented in Table 3.

Table 3. Validation results

Indicator	Conventional model	Zero Trust + Blockchain
Scenario 1		
Network operation delay	5-7 ms	10-15 ms
Connection stability	95%	98%
Transparency of access	60%	100
Ability to counter threats	40%	98%
Scenario 2		
Network operation delay	6-10 ms	12-18 ms
Connection stability	70%	95%
Transparency of access	50%	100%
Ability to counter threats	40%	98%
Scenario 3		
Network operation delay	7-12 ms	15-25%
Connection stability	50%	92%
Transparency of access	40%	100%
Ability to counter threats	20%	96%

Source: compiled by the authors

The validation results demonstrated significant advantages of integrating Zero Trust and blockchain technologies over the conventional model. In the context of network operation latency, the conventional model was faster, providing minimal latency values due to its simple architecture (5-7 ms in the first scenario, 6-10 ms in the second, 7-12 ms in the third). However, the integration of Zero Trust and blockchain, which added consensus and access verification mechanisms, led to an increase in latency (10-15 ms, 12-18 ms, and 15-25%, respectively), which is a justified compromise for improving security. The performance score is shown in the delay comparison graph between the conventional system and the integrated model, which clearly shows a trade-off between improving security and increasing response time (Fig. 2).

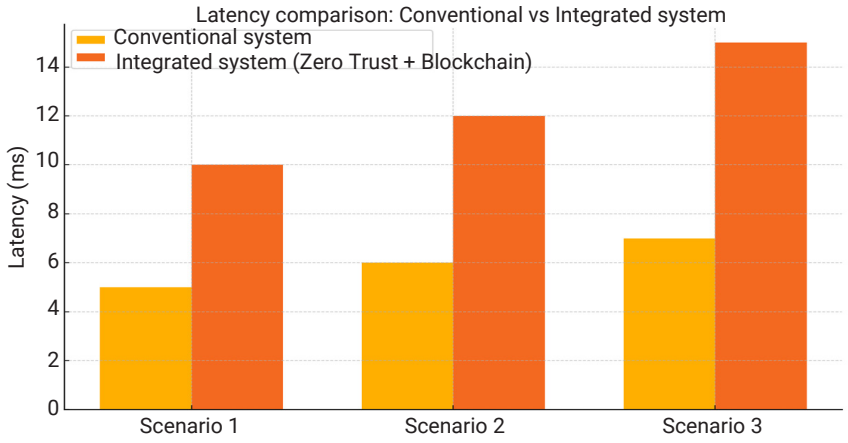

Figure 2. Comparison of delays between a conventional system and an integrated model (Zero Trust + Blockchain) for three scenarios

Source: created by the authors

The stability of connections in the conventional model remained high only under normal conditions (95% in the first scenario), but significantly decreased to 70% in the second and 50% in the third scenarios during attacks. In turn, the model with integration of Zero Trust and blockchain showed significantly higher stability of connections: 98% in the first, 95% in the second, and 92% in the third scenario. This was achieved by restricting access only for authorised users and ensuring that the network is resistant to threats (Fig. 3).

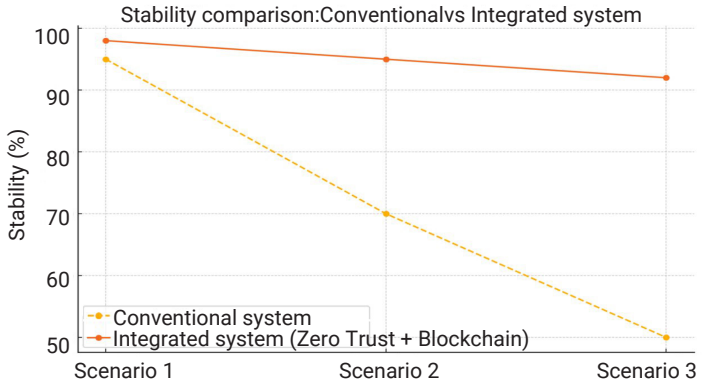

Figure 3. Comparison of connection stability between a conventional system and an integrated model (Zero Trust + Blockchain) in three scenarios

Source: created by the authors

Access transparency in the conventional model is limited (40-60%), as event logs may be incomplete or changeable. In the blockchain model, transparency reaches 100% because all operations are recorded unchanged, which ensures full control and audit of all actions in the network. The ability to counter threats in the conventional model was quite low (40% in all scenarios), which is explained by the lack of modern protection mechanisms. Instead, the integration of Zero Trust and blockchain increased this figure to 98%, protecting the network from unauthorised access and hacking of logs. The combination of these technologies allows quickly detecting and neutralising even complex attacks.

Overall, the integration of Zero Trust and blockchain has shown significant advantages. Access transparency reached its maximum level (100%), the ability to counteract threats was 96-98%, and the stability of connections remained high

even in difficult conditions. A slight increase in the latency of network operations is justified by an increase in the level of security, transparency, and reliability. These results confirm the feasibility of implementing such technologies to create a secure and sustainable cloud environment.

One of the key advantages of Zero Trust and blockchain in SDN networks is enhanced security through control of each transaction and decentralised protection. This allows creating a self-contained system where all transactions are recorded and verified, which makes it difficult to fake or make unauthorised changes. The disadvantages of this approach include an increase in network load, in particular, due to the processing of blockchain transactions, which leads to increased latency and resource consumption. In addition, the Zero Trust approach requires significant resources to authenticate and verify each user and request. A comparison of the conventional approach to security and the Zero Trust and blockchain-based approach is presented in Table 4.

Table 4. Comparison of the conventional approach to security and the approach based on Zero Trust and blockchain

Indicator	Conventional approach	Zero Trust + Blockchain
Centralisation of management	Central management	Decentralisation through blockchain
Protection against internal threats	Limited	In-depth control of each access
Transparency	Limited	High, due to the blockchain
Network load	Low	High, due to additional checks
Delay	Minimum	Moderate, depending on the volume of checks
Implementation cost	Relatively low	High, given the support resources

Source: created by the authors based on K. Gai *et al*. (2022), P. Dhiman *et al*. (2024), O. Bykonia & N. Romanovska (2024)

The analysis shows that the use of Zero Trust and blockchain in SDN networks is effective, but requires optimisation and resource management to reduce costs and delays. Potential scenarios for using the integrated model are particularly relevant for environments with high security requirements, in particular, in the context of possible combined attacks. Due to the transparency and immutability of data, blockchain is extremely effective in environments where a high level of trust in information is important. This applies to such industries as public administration, financial services, critical infrastructure (in particular, energy), healthcare, etc. The simulation results confirmed that the integration of Zero Trust and blockchain technologies can significantly reduce risks, increase transparency, and ensure control over network transactions. The implementation of Zero Trust in such networks provides protection against suspicious activity, internal threats and privacy violations, while the blockchain adds another layer of security, recording all changes and actions on the network. This approach opens up opportunities for creating autonomous networks that are much less dependent on centralised controllers and provide the maximum level of reliability and protection.

Challenges and limitations of implementing Zero Trust and blockchain in SDN networks

The implementation of Zero Trust and blockchain in SDN networks is accompanied by a number of challenges that cover technical, resource, and organisational aspects. One of the key limitations of implementing blockchain and Zero Trust technologies in SDN networks is the high requirements for memory and computing resources. This is conditioned by the need to store a large amount of data in the blockchain, which is accompanied by significant computing costs, especially in large networks with high bandwidth. The complexity of blockchain protocols and the constant access verification typical of Zero Trust further complicate the situation. In such networks, CPU and memory resources must process a huge number of requests, which leads to increased response times and creates delays that are critical for systems with high efficiency requirements.

In addition, the scalability of such systems becomes a challenge due to the need to dynamically manage large volumes of traffic and maintain operational efficiency even under high load conditions. This requires the introduction of hybrid approaches and automation of security policies to reduce the load on the system. However, even under such conditions, high computational costs remain a significant problem that requires further optimisation. Delays in performing operations are one of the key limitations of integrating blockchain and Zero Trust technologies. They arise from continuous access verification, consensus processes, and transaction control, which adds additional steps to each transaction. This can significantly affect network performance, especially in systems with high operational requirements, such as financial transactions or critical infrastructure management. This problem is compounded in large networks with a large number of users and devices, where manual administration of Zero Trust policies becomes time-consuming and error-prone. Process automation is necessary to reduce delays and increase efficiency, in particular, the introduction of adaptive systems and machine learning. This allows dynamically adjusting access policies based on the context and user behaviour, and minimise the impact of delays on network performance.

Another challenge is scalability: integrating blockchain and Zero Trust into large corporate or government SDN networks can become problematic due to significant amounts of data that require storage and processing. This requires the development of dynamic resource scaling methods to maintain efficient network performance even under high load conditions (Liu *et al*., 2020). Data privacy is also an important limitation. Despite the fact that the blockchain ensures the immutability of data, it also makes all transactions available for viewing, which can lead to a

potential leak of confidential information. S. Ghasemshirazi *et al*. (2023) noted that in such cases, it is necessary to develop secure storage methods where the blockchain is combined with other cryptographic methods to provide the necessary level of data protection.

The complexity of implementing technologies is also a serious barrier. It is caused by the need for significant changes in the architecture of SDN networks, in particular, reconfiguring controllers, creating new access policies, and scaling the blockchain infrastructure for processing large amounts of data. The use of Zero Trust and blockchain technologies in SDN networks requires significant financial investments, especially at the initial stages of implementation. The main items of expenditure are the purchase and configuration of computing resources necessary to maintain blockchain operations and ensure reliable multi-level authentication. As the network scale and traffic volumes increase, the cost of processing and storing data increases, as each transaction requires authentication. However, despite the high costs, this approach can be cost-effective for companies that need reliability and security.

In addition, managing Zero Trust policies on large networks with a large number of users and devices is becoming increasingly difficult. High interaction dynamics require constant monitoring, updating access policies, and detecting anomalies, which is complicated by the growing scale of the network. This makes manual administration extremely time-consuming and potentially error-prone. Automating processes using adaptive systems and machine learning technologies can partially solve these problems. This will allow dynamically adjusting access policies based on analysis of user and device behaviour, reducing the amount of manual work and improving management efficiency. However, even such solutions require additional investment in the development and implementation of the appropriate infrastructure.

Recommendations for integrating Zero Trust concepts and blockchain technology in SDN networks

Integration of Zero Trust and blockchain into SDN networks requires a comprehensive approach that considers architectural, technical, and organisational aspects. The first step is to analyse the network infrastructure to identify its critical assets, potential vulnerabilities, and specific needs. This information is the basis for developing a hybrid architecture that combines decentralised blockchain solutions for mission-critical data and centralised ones for less significant operations. To implement the Zero Trust approach, it is necessary to implement network micro-segmentation, which will ensure the isolation of critical areas from potentially dangerous ones. It is recommended to use multi-level authentication, which includes risk factors such as location, device type, and user behavioural patterns. Dynamic access control should automatically adjust policies based on changes in the network environment and user behaviour. Blockchain should be used to increase the transparency of access control. All access transactions and policy changes must be recorded in a distributed ledger, which ensures that they remain unchanged and can be audited. The implementation of smart contracts will automate the management of access policies and ensure a quick response to potential threats. To protect the API, it is important to implement request tokenisation and distributed identification mechanisms.

To avoid the high costs and delays associated with computing, it is recommended to implement modern consensus mechanisms, such as Proof of Stake, and optimising network performance through caching and prioritising requests. Automated monitoring using AI technologies can significantly improve the effectiveness of security management, helping to quickly detect anomalies and adapt the system to new threats. Successful integration requires not only technical solutions, but also organisational support. Staff should be trained to effectively manage new systems. In addition, security policies should be updated regularly to meet the latest threats and technology changes. Effective implementation of Zero Trust and blockchain involves creating a step-by-step integration plan that includes pilot projects, testing the system in real-world conditions, and scaling it. At all stages, implementation performance should be evaluated, including response time, stability, performance, and security level analysis.

In addition, it is important to develop mechanisms for resilience to new threats, considering the specifics of the industry or organisation. For example, in high-risk environments (finance, critical infrastructure), additional security mechanisms should be integrated, such as transaction-level data encryption or mandatory audit of all accesses. Compliance with these recommendations will create an SDN network that not only provides a high level of security, but is also flexible, adaptive to changes and resistant to modern cyber threats.

Discussion

The integration of the Zero Trust architecture and blockchain platform implemented in this study demonstrates significant potential for improving the security, transparency, and reliability of SDN networks. The proposed approach provides effective protection of network resources through a combination of strict access control, decentralised data storage, and automation of security policies. The use of blockchain enabled the creation of an immutable log of transactions and events, which increases transparency and simplifies auditing. Moreover, the implementation of Zero Trust provided isolation of critical resources and dynamic access control, which significantly reduces the risks of internal and external threats. The analysis showed high efficiency of the proposed approach even in conditions of large networks with high bandwidth and a significant number of users. However, key challenges were considered, in particular, delays in performing operations and high computing costs, which allowed optimising integration using hybrid solutions and modern consensus mechanisms. Thus, the results of this study confirmed the possibility of

scaling the proposed technologies and their practical value for use in complex corporate and government networks.

The architecture that uses blockchain to manage security in SDN for 5G applications is presented in the paper by D. Das *et al*. (2023). The researchers considered methods for protecting communications in 5G environments through the use of blockchain as a means for authentication and storage of records. The results of the study showed that the introduction of blockchain can significantly improve the security of communications by reducing the risks of unauthorised access and data compromise. The researchers proposed the integration of blockchain smart contracts for authentication management, which allowed automating access verification and providing a transparent data storage mechanism. This solution has proven to be particularly effective for high-traffic environments such as 5G, reducing the risk of data loss even in the event of attacks on network nodes.

A comparison with the results obtained in the study shows similarities in the approach to ensuring transparency and decentralisation of access control. In the current study, blockchain was also used to strengthen the security of software-defined networks (SDNs) by recording transactions and automating authentication processes. The same technology using blockchain to improve digital security in the defence sector was proposed by O. Semenenko *et al.* (2024), noting its effectiveness in protecting data from cyber threats.

In this study, the integration of Zero Trust and blockchain to improve the security of SDN networks showed high efficiency in minimising risks and preventing unauthorised access, but this was accompanied by significant computational costs and delays, in particular, due to the constant authentication of each request. As stated by L. Alevizos & V.T. Ta (2024), cybersecurity automation capabilities using AI, blockchain, and smart contracts can be used to solve these problems. The researchers proposed approaches that combine AI for monitoring and predicting threats with blockchain smart contracts for automating access policies. This allows only quickly responding to detected threats, but also provides adaptive resource management in real time. For example, AI can analyse network traffic to identify anomalies, and smart contracts can automatically restrict access in the event of a potential threat. The main challenge is the need to balance blockchain transparency, responsiveness, and computing costs. In the current paper, the main focus is on using a hybrid blockchain to minimise the load, but using AI to predict and automate threat detection, as indicated by L. Alevizos & V.T. Ta, opens up new opportunities for optimising Zero Trust systems. In addition, the use of AI can reduce delays in decision-making, since instead of static security rules, it is possible to implement dynamic models that adapt to new attack scenarios. This makes their approach more promising for scenarios with high threat dynamics, such as in cloud or financial networks.

Features of implementing Zero Trust in cloud networks were discussed by S. Ahmadi (2024). The researcher discussed current challenges and potential areas for developing Zero Trust technology to ensure data security in the cloud environment. Managing Zero Trust policies for a large number of users and devices can be challenging. This problem can be partially overcome by automating access control processes using machine learning and adaptive systems. D. Ajish (2024) examined the role of AI in Zero Trust technologies, in particular, its ability to improve security in the architecture. The researcher analysed how AI helps to detect anomalies, assess access risks, and automate security policy management processes. Due to machine learning algorithms, AI can dynamically adapt to changes in the network environment, quickly respond to new threats, and reduce the human factor in decision-making. AI also helps to automate the access verification process by using behavioural analysis to create user and device profiles. This helps to more accurately determine the level of trust and adapt access rights accordingly. In addition, AI is actively used to monitor network traffic in real time, which allows quickly detecting attempts to compromise or violate access policies.

S. Dhar & I. Bose (2020) explored the possibilities of using blockchain and Zero Trust to protect IoT devices. The researchers proposed an architecture that provides reliable access control and data protection in the IoT. In industrial IoT networks used to monitor and manage production processes, Zero Trust and blockchain can significantly improve system security by preventing unauthorised access to devices and reducing the risk of data manipulation. The decentralised blockchain architecture makes such networks less vulnerable to centralised attacks, and Zero Trust provides control at the level of individual devices.

A. Kulkarni *et al*. (2024) investigated the use of blockchain and physically unclonable functions to protect Field-Programmable Gate Array supply chains. This system uses Zero Trust to ensure the authenticity of components and the security of information exchange between manufacturers and suppliers. The researchers created a model based on the uniqueness of the physical characteristics of each device, which provides identification without the need to store sensitive data on the device. The blockchain in this architecture is used to capture transactions and ensure their immutability, creating a reliable platform for interaction between supply chain participants. The proposed approach provides a high level of security, since the combination of blockchain and physically unclonable functions reduces the risk of component tampering or unauthorised interference in the delivery process. Using Zero Trust adds another layer of protection, because access to components is possible only after multi-level authentication, which minimises the likelihood of compromise even in the presence of insider threats. Compared to conventional supply chain security methods, such as centralised certification systems, this approach is more transparent and reliable, since the blockchain ensures the immutability of data about each transaction, and physically unclonable functions guarantee the authenticity of each physical component.

Compared to the current study, which focused on integrating Zero Trust and blockchain to protect SDN networks, approach of A. Kulkarni *et al*. focused on the physical security of devices and supply chains. Although both approaches demonstrate effectiveness in providing security, their challenges are similar: increasing system load and delays due to decentralisation and multi-level authentication.

A blockchain-based infrastructure for providing Zero Trust models on peripheral computing devices was proposed in the paper by C. Bicer *et al*. (2023). The researchers described the concept of using blockchain to improve device security and data transmission at the edge of the network without having to trust a central administrator. The security benefits of Zero Trust and blockchain may exceed performance limits, but for large-scale networks, there is a need to improve consensus and data processing mechanisms. For example, hybrid blockchain solutions can be used to improve efficiency, where basic data is processed centrally and critical transactions are stored on the blockchain. This allows reducing the load on the network, while maintaining a high level of security.

A blockchain-based authentication scheme for railway networks was proposed by Y. Feng *et al*. (2023). The researchers considered the features of ensuring the security of communications in the railway infrastructure using Zero Trust. Together, they provide reliable protection against a wide range of attacks, especially relevant for modern dynamic network environments. Compared to other approaches, such as network segment encryption and role-based access control, Zero Trust and blockchain offer an additional layer of protection by constantly monitoring and recording all transactions. This makes them indispensable for environments where privacy and trust in data are critical.

Using proxy smart contracts to implement Zero Trust in decentralised oracles networks, as suggested in the paper by A. Gupta *et al*. (2023), is an example of the benefits of integrating these technologies. Such smart contracts allow strictly controlling access to data and transactions in oracles' networks, which is crucial for ensuring security. Integration of Zero Trust with blockchain smart contracts ensures transparency of interaction, since each transaction is recorded in an immutable register. This eliminates the risks associated with data manipulation and reduces the likelihood of compromise even in the event of internal threats. In addition, smart contracts automate access verification processes and the implementation of security policies, which significantly increases the efficiency and speed of the system.

Z. Bassfar *et al*. (2023) proposed a Zero Trust architecture using quantum device identification. The study focused on ensuring the security of network access through quantum technologies. Integration with machine learning tools for adaptive access control will allow faster detection of anomalies, automate the process of managing access policies, and increase the efficiency of the Zero Trust environment. In addition, the use of a hybrid blockchain model, where some information is processed centrally, and critical transactions are stored in a decentralised network, can provide faster data processing while maintaining a high level of security.

Zero Trust architecture automation solutions were investigated by Y. Cao *et al*. (2024). The researchers considered the advantages and problems associated with process automation in Zero Trust. In their research, they focused on using automated access policies that adapt in real time depending on changes in user behaviour or system status. This is achieved through the integration of AI-based analytics, which provides monitoring and detection of threats, and automatic changes to security policies without the participation of an administrator. The results of the study showed that Zero Trust automation significantly increases the effectiveness of threat detection and reduces risks associated with the human factor. For example, in networks with a large volume of traffic or many users, automation can reduce delays that occur due to manual access control. In addition, AI can identify complex attacks that may be invisible to conventional monitoring systems. Automated solutions also allow considering the context (location, device, request time), which is important for building more flexible and secure networks. This is especially true for critical infrastructures or financial SDN networks, where data security is critical. However, as the researchers note, automation also has problems, in particular, related to scalability and the potential complexity of configuring such systems. They point out the need for a combination of automation and transparency that can provide more sustainable and productive Zero Trust networks. Their results support the feasibility of recommendations in the work on AI integration for dynamic access control in environments with high security requirements.

Conclusions

The integration of Zero Trust and blockchain concepts into SDN networks has demonstrated high efficiency in improving security, access transparency, and network resilience to threats. Analysis of the results confirmed that the introduction of multi-level access control, micro-segmentation, and transparent policy management significantly reduces the risk of compromise from both internal and external attacks. The implemented model using the blockchain ensured the immutability of records and reduced the risk of data manipulation, which helped to achieve 96-98% effectiveness of countering threats compared to conventional security methods.

Integration of the Zero Trust architecture and blockchain technology into SDN networks was evaluated as an effective strategy for countering detected threats. Zero Trust provides continuous access verification, reducing the risk of unauthorised actions, especially by internal users. Each element of the network requires authentication and authorisation for access, which minimises the possibility of an attack. Blockchain technology adds a level of decentralisation and transparency, which avoids dependence on a single control centre and prevents data forgery. The blockchain also registers all transactions and

changes in the network, which creates a reliable mechanism for monitoring and auditing.

As a result of the analysis of the main threats to software-defined networks (SDN), it was found that the greatest danger is represented by attacks on the control and management layers. In particular, common threats include unauthorised access to network resources, attacks on a centralised controller, data leaks, and internal threats. These threats significantly affect the security and stability of SDN networks, undermining their integrity and confidentiality. It was found that conventional approaches are not always able to provide an adequate level of protection, since they rely on centralised management, which is vulnerable to external attacks.

The results confirm that the use of Zero Trust and blockchain significantly increases the stability of SDN networks even in high-risk environments, ensuring connection stability and minimising downtime. The use of a hybrid blockchain has reduced the load on the network, while maintaining a high level of security. Despite the slight increase in network latency due to multi-level access verification, the advantages of network transparency and security outweigh these disadvantages. Analysis of the results showed that the use of Zero Trust and blockchain contributes to a significant increase in the level of security in SDN networks, but simultaneously increases the requirements for computing resources and network delays. Further research on methods for optimising blockchain protocols is recommended, in particular, the use of consensus mechanisms with lower resource consumption, such as Proof of Stake, instead of Proof of Work. It is also important to consider integrating machine learning to automate access control and real-time threat detection. This will not only increase efficiency, but also reduce the need for manual intervention.

Promising areas of future research are improving system scalability and reducing delays by optimising blockchain architectures, in particular, the introduction of multi-level models and the development of hybrid models, where the main data is processed centrally, and mission-critical transactions are stored on the blockchain. This approach will help to reduce network load and provide the necessary level of security.

Acknowledgements

None.

Funding

The study received no funding.

Conflict of Interest

None.

References

[1] Ahmadi, S. (2024). Zero trust architecture in cloud networks: Application, challenges and future opportunities. *Journal of Engineering Research and Reports*, 26(2), 215-228. doi: 10.9734/jerr/2024/v26i21083.

[2] Ajish, D. (2024). The significance of artificial intelligence in Zero Trust technologies: A comprehensive review. *Journal of Electrical Systems and Information Technology*, 11(1), article number 30. doi: 10.1186/s43067-024-00155-z.

[3] Alevizos, L., & Ta, V.T. (2024). Automated cybersecurity compliance and threat response using AI, blockchain & smart contracts. *International Journal of Information Technology*, 17, 767-781. doi: 10.1007/s41870-024-02324-9.

[4] Bassfar, Z., Sayeed, A., Bala, P., Alshehri, A., Alanazi, M., & Zubair, S. (2023). Toward secure and resilient networks: A Zero-Trust security framework with quantum fingerprinting for devices accessing network. *Mathematics*, 11(12), article number 2653. doi: 10.3390/math11122653.

[5] Bicer, C., Murturi, L., Donta, P.K., & Dustdar, S. (2023). Blockchain-based Zero Trust on the edge. *ArXiv*. doi: 10.48550/arXiv.2311.16744.

[6] Bykonia, O., & Romanovska, N. (2024). Perspectives of the development of the information and communication technologies sector in Ukraine. *Scientific Bulletin of International Association of Scientists. Series Economy Management Security Technologies*, 3(1). doi.org/10.56197/2786-5827/2024-3-1-8.

[7] Cao, Y., Pokhrel, S.R., Zhu, Y., Doss, R., & Li, G. (2024). Automation and orchestration of Zero Trust architecture: Potential solutions and challenges. *Machine Intelligence Research*, 21(10), 294-317. doi: 10.1007/s11633-023-1456-2.

[8] Das, D., Banerjee, S., Dasgupta, K., Chatterjee, P., Ghosh, U., & Biswas, U. (2023). Blockchain enabled SDN framework for security management in 5G applications. In *ICDCN 23: proceedings of the 24th international conference on distributed computing and networking* (pp. 414-419). New York: Association for Computing Machinery. doi: 10.1145/3571306.3571445.

[9] Dhar, S., & Bose, I. (2020). Securing IoT devices using Zero Trust and blockchain. *Journal of Organizational Computing and Electronic Commerce*, 31(1), 18-34. doi: 10.1080/10919392.2020.1831870.

[10] Dhiman, P., Saini, N., Gulzar, Y., Turaev, S., Kaur, A., Nisa, K.U., & Hamid, Y. (2024). A review and comparative analysis of relevant approaches of Zero Trust network model. *Sensors*, 24(4), article number 1328. doi: 10.3390/s24041328.

[11] Fadhil, J.A., & Zeebaree, S.R. (2024). Blockchain for distributed systems security in cloud computing: A review of applications and challenges. *Indonesian Journal of Computer Science*, 13(2), 1576-1605. doi: 10.33022/ijcs.v13i2.3794.

[12] Feng, Y., Zhong, Z., Sun, X., Wang, L., Lu, Y., & Zhu, Y. (2023). Blockchain enabled Zero Trust based authentication scheme for railway communication networks. *Journal of Cloud Computing*, 12(1), article number 62. doi: 10.1186/s13677-023-00411-z.

[13] Gai, K., She, Y., Zhu, L., Choo, K.W., & Wan, Z. (2022). A blockchain-based access control scheme for Zero Trust cross-organizational data sharing. *ACM Transactions on Internet Technology*, 23(3), article number 38. doi: 10.1145/3511899.

[14] Gai, K., Wu, L., Zhu, L., Zhang, Z., & Qiu, M. (2019). Differential privacy-based blockchain for industrial internet-of-things. *IEEE Transactions on Industrial Informatics*, 16(6), 4156-4165. doi: 10.1109/TII.2019.2948094.

[15] Ghasemshirazi, S., Shirvani, G., & Alipour, M.A. (2023). Zero Trust: Applications, challenges, and opportunities. *ArXiv*. doi: 10.48550/arXiv.2309.03582.

[16] Guo, X., Wang, C., Cao, L., Jiang, Y., & Yan, Y. (2022). A novel security mechanism for software defined network based on blockchain. *Computer Science and Information Systems*, 19(2), 523-545. doi: 10.2298/CSIS210222001G.

[17] Gupta, A., Gupta, R., Jadav, D., Tanwar, S., Kumar, N., & Shabaz, M. (2023). Proxy smart contracts for Zero Trust architecture implementation in Decentralized Oracle Networks based applications. *Computer Communications*, 206, 10-21. doi: 10.1016/j.comcom.2023.04.022.

[18] Kulkarni, A., Hazari, N.A., & Niamat, M.Y. (2024). A Zero Trust-based framework employing blockchain technology and ring oscillator physical unclonable functions for security of field programmable gate array supply chain. *IEEE Access*, 12, 89322-89338. doi: 10.1109/ACCESS.2024.3418572.

[19] Li, J., Lv, H., Lei, B., & Xie, Y. (2022). A consensus approach for SDN controllers based on blockchain. In *CSSE '22: proceedings of the 5th international conference on computer science and software engineering* (pp. 170-174). New York: Association for Computing Machinery. doi: 10.1145/3569966.3570015.

[20] Li, W., Meng, W., Liu, Z., & Au, M.-H. (2020). Towards blockchain-based software-defined networking: Security challenges and solutions. *IEICE Transactions on Information and Systems*, E103.D(2), 196-203. doi: 10.1587/transinf.2019INI0002.

[21] Liu, Y., He, D., Obaidat, M.S., Kumar, N., Khan, M.K., & Choo, K.-K. (2020). Blockchain-based identity management systems: A review. *Journal of Network and Computer Applications*, 166, article number 102731. doi: 10.1016/j.jnca.2020.102731.

[22] Semenenko, O., Kirsanov, S., Movchan, A., Ihnatiev, M., & Dobrovolskyi, U. (2024). Impact of computer-integrated technologies on cybersecurity in the defence sector. *Machinery & Energetics*, 15(2), 118-129. https://doi.org/10.31548/machinery/2.2024.118.

[23] Xu, Y., Ren, J., Wang, G., Zhang, C., Yang, J., & Zhang, Y. (2019). A blockchain-based nonrepudiation network computing service scheme for industrial IoT. *IEEE Transactions on Industrial Informatics*, 15(6), 3632-3641. doi: 10.1109/TII.2019.2897133.

[24] Yan, X., & Wang, H. (2020). Survey on zero-trust network security. In X. Sun, J. Wang & E. Bertino (Eds.), *Artificial intelligence and security* (pp. 50-60). Singapore: Springer. doi: 10.1007/978-981-15-8083-3_5.

[25] Zheng, P., Jiang, Z., Wu, J., & Zheng, Z. (2023). Blockchain-based decentralized application: A survey. *IEEE Open Journal of the Computer Society*, 4, 121-133. doi: 10.1109/OJCS.2023.3251854.

Інтеграція Zero Trust і Blockchain у SDN-мережах: огляд загроз та методів їх усунення

Олександр Підпалий

Аспірант
Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського»
03056, просп. Берестейський, 37, м. Київ, Україна
https://orcid.org/0009-0007-6852-7959

Олександр Романов

Доктор технічних наук, професор
Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського»
03056, просп. Берестейський, 37, м. Київ, Україна
https://orcid.org/0000-0002-8683-3286

Анотація. Дослідження спрямоване на визначення теоретично обґрунтованих методів інтеграції концепцій Zero Trust і Blockchain з метою підвищення загальної безпеки програмно-конфігурованих мереж (SDN). Дослідження базувалося на розробці теоретичної моделі мережі, яка включає в себе SDN-контролер, комутатори, маршрутизатори та хости, для чого було використано інструменти віртуалізації, такі як GNS3, VirtualBox та Docker. Теоретична основа дослідження охоплює аналіз ключових загроз, серед яких DDoS-атаки, маніпуляції з маршрутизацією, інсайдерські загрози, атаки на application programming interface (API), а також специфічні уразливості механізмів консенсусу Blockchain. Імітаційні сценарії були розроблені для демонстрації потенційного впливу цих загроз на безпеку та продуктивність SDN-мереж. Аналіз отриманих результатів теоретично підтверджує, що застосування політик Zero Trust суттєво знижує ризики інсайдерських атак і покращує захист SDN-контролера завдяки принципам постійної перевірки доступу і мікросегментації. Інтеграція технологій Blockchain підвищує надійність маршрутизації та управління трафіком, запобігаючи спробам зловмисного втручання в мережеву інфраструктуру. Теоретичні методи аутентифікації та верифікації запитів з використанням Blockchain значно покращують захист API та інтерфейсів взаємодії. Крім того, гібридні алгоритми консенсусу показали потенціал для підвищення продуктивності мережі та забезпечення її стійкості до атак. Проведене дослідження підкреслює важливість інтеграції Zero Trust і Blockchain як ефективного рішення для усунення широкого спектра загроз у SDN-мережах. Це відкриває нові перспективи для захисту телекомунікаційних систем і закладає теоретичну основу для подальших досліджень і вдосконалення методів безпеки. Практична значимість дослідження полягає у розробці конкретних рекомендацій щодо впровадження комплексної системи захисту SDN на основі технологій блокчейн та принципів Zero Trust. Запропоновані рішення можуть бути використані як у державному секторі для захисту критичної інфраструктури, так і в приватному секторі для забезпечення безпеки корпоративних мереж

Ключові слова: контроль доступу; верифікація даних; зниження ризиків; розподілені системи; стійкість до атак; безпека комунікацій

Journal homepage: https://itce.com.ua/en

Information Technologies and Computer Engineering **Vol. 22, No. 1. 2025**

Article's History: Received: 02.01.2025 Revised: 11.03.2025 Accepted: 24.04.2025

UDC 004.8 DOI: 10.63341/vitce/1.2025.69

Method for constructing a cognitive map of processes in a dynamic system using the cooperation of large language models

Borys Varer*

Postgraduate Student
Vinnytsia National Technical University
21021, 95 Khmelnytske Shose Str., Vinnytsia, Ukraine
https://orcid.org/0000-0002-5860-0100

Vitalii Mokin

Doctor of Technical Sciences, Professor
Vinnytsia National Technical University
21021, 95 Khmelnytske Shose Str., Vinnytsia, Ukraine
https://orcid.org/0000-0003-1946-0202

Abstract. In the context of growing demands for rapid decision-making and in-depth analysis of complex dynamic systems – particularly when available data are limited and the involvement of experienced experts is either impractical or prohibitively expensive – the development of new methods for the construction of the model becomes especially relevant. The use of large language models (LLMs) as expert systems offers significant reductions in resource expenditure and accelerates the modelling of complex technical, environmental, and socio-economic systems. This study aimed to investigate and demonstrate the potential and capabilities of LLMs as expert systems in constructing cognitive maps. The article proposes and substantiates an architecture for the cooperation of LLM ensembles to formally generate vertices-variables and weight coefficients in cognitive maps, thereby enabling the automation of the modelling process without the involvement of human experts. A typical prompt for an LLM was decomposed into structural components: context description (D), model role instruction (R), instruction (I), conditions (C), and response format (F). A method for determining these components through expert-based analysis is proposed. A prompt system was developed to enable structured data processing and the identification of interrelationships among system elements. The practical effectiveness of the approach was demonstrated using a case study on forecasting water quality in the Sabarivske Reservoir near Vinnytsia. For most physicochemical indicators, the modelling showed low error rates (2.09-4.6%), even with a minimal amount of input data. The proposed method is promising for modelling and forecasting tasks in complex systems with limited data availability, particularly in environmental, socio-economic, and engineering contexts, where the speed of obtaining reliable results is critical for informed decision-making

Keywords: LLM; generative artificial intelligence; intelligent technology; systems analysis; modelling; forecasting; dynamic system

Introduction

To address forecasting and decision support tasks, it is crucial to possess a model capable of predicting a system's reaction to certain perturbations. As is known, if a system is well-defined, it is typically modelled using mathematical methods. Conversely, if a system exhibits high uncertainty and a significant volume of data, data science and intelligent models are generally applied. In the case of weakly structured systems with a limited amount of data, an expert approach and intelligent formalisation methods based on cognitive maps are employed. A cognitive map (CM) is a directed graph that connects vertices-variables by arcs with weights, whose values are constant and lie within the range . Cognitive maps are an effective tool for modelling dynamic systems; however, traditional approaches to CM construction rely on expert evaluations and require considerable time. Furthermore, they are susceptible to

Suggested Citation:
Varer, B., & Mokin, V.,(2025). Method for constructing a cognitive map of processes in a dynamic system using the cooperation of large language models. *Information Technologies and Computer Engineering*, 22(1), 69-78. doi: 10.63341/vitce/1.2025.69

*Corresponding author

expert subjectivity, which complicates the development of stable models. Simultaneously, modern large language models (LLMs) already contain a wide spectrum of expert knowledge and can significantly accelerate this process, provided their cooperation is correctly organised and an effective prompting system is in place. This underscores the necessity for developing a method utilising LLMs to automate the construction of cognitive maps.

Over recent years, several studies have been published concerning the use of large language models and cognitive maps for modelling complex systems. However, each of these illuminates somewhat different aspects of this issue and does not offer a comprehensive solution for the full automation of CM construction based on LLMs. For instance, T. Liu *et al*. (2024) demonstrated the capability of LLMs (using the LLaMA 2 model as an example) to perform forecasting of the dynamics of various systems without additional training. Their conclusion posited that LLMs can effectively serve as a basis for predictive models, even for complex dynamic processes. Researchers R. Schuerkamp *et al*. (2025) proposed an original mechanism for integrating multiple expert cognitive maps using LLMs. Their method of automatic "reconciliation" of contradictory statements enables LLMs to analyse causal relationships independently generated by several experts and to propose a single, agreed-upon map without explicit conflicts. While this opens up the prospect of large-scale integration of heterogeneous knowledge in complex tasks, it is based on maps previously formulated by experts rather than on full automation.

In the research by A. Feleki *et al*. (2023), an integrated Deep FCM approach was applied, where a convolutional neural network (CNN) was combined with fuzzy cognitive maps (FCMs) for the diagnosis of heart diseases. However, the LLM (GPT-3.5) was only used for the automatic generation of text explanations in natural language. In practice, this allowed for the analysis of medical images using a convolutional neural network and their combination with clinical data in an FCM classifier with enhanced explanatory capability. In the study by W. Godoy *et al.* (2024), the level of student satisfaction with factors such as teaching quality, infrastructure, and social environment, among others, was compared, with results obtained using an LLM and results modelled using a CM that was constructed traditionally. The results were found to be almost identical: a score of 7.5 was obtained using the CM, and a score of 7.4 was obtained using OpenAI's GPT-4 LLM. However, the cognitive map itself was created separately by expert means, without any attempts at its automatic generation using an LLM.

As for the use of cognitive maps (CMs) themselves in modelling tasks, this approach is demonstrating increasing popularity within the scientific community due to its ability to effectively represent complex causal relationships across various subject domains, whilst providing an intuitively understandable visualisation of system dynamics and supporting decision-making under uncertainty. Specifically, O. Saliieva & Y. Yaremchuk (2020) proved the reliability of modelling the impact of threats on the security level of an information protection system and a critical infrastructure object, conducted based on a cognitive approach. Similarly, S. Shevchenko *et al.* (2024) demonstrated the effectiveness of cognitive maps for modelling information security risk scenarios. CMs in their research enable the identification of key vulnerabilities and optimal enterprise protection strategies through scenario analysis. However, the map itself was constructed manually based on expert experience, and the issue of automated formation of the structure and weighting coefficients using LLMs was not considered.

An analysis of the systemic risks associated with the application of chatbots in education is presented in the study by O. Cherniuk (2023). The author utilised a cognitive map to investigate the interaction between students and generative language models, identifying both positive consequences (such as increased motivation and access to knowledge) and threats (for instance, the temptation to violate academic integrity). In this study, the cognitive map was also constructed manually by experts, and the use of LLMs for automation was not considered.

Another relevant example of implementing cognitive maps is presented in the article by V. Mokin *et al.* (2021), where the authors proposed a mathematically grounded method for synthesising a stable multi-connected CM by sequentially expanding the base map to higher orders. This approach demonstrated its effectiveness through the analysis of ecological processes in the Southern Bug River in the Vinnytsia Region. However, in this case, as well, the identification of vertices and interconnections was based on prior expert input, meaning that full automation without specialist involvement was not envisaged.

Based on the analysis conducted, several aspects can be identified which are either not addressed at all or are insufficiently covered in current scientific literature: the development of a completely automated method for constructing cognitive maps without manual expert involvement; the use of LLMs not merely for explanations or reconciling existing maps, but for generating the entire structure of the CM (vertices-variables and weighting coefficients); and the evaluation of the actual effectiveness of such an approach for complex systems with limited data. Consequently, this article aimed at the development of a method for constructing a cognitive map of processes in a dynamic system using the cooperation of large language models in the role of experts. Specifically, this entailed defining the key elements of the system, their interrelationships and the cognitive map's weighting coefficients, as well as forming a system of prompts to ensure effective LLM cooperation, followed by verifying the adequacy and accuracy of the constructed model using real data.

Materials and Methods

Complex dynamic systems in discrete time $k = 0, 1, 2, \ldots$, can typically be formalised in the form of sets of values for input variables U, state variables X, and an output variable Y. A case was considered where in each modelling scenario, there is only one input and only one output variable. It is

necessary for the values of the variables to be represented on a single scale (e.g., normalised), as the use of different units of measurement can lead to inaccuracies in modelling. Importantly, the graph is undirected and all variables are interchangeable; that is, an output variable or some intermediate state variable can become an input and vice versa, the main point is to always adhere to the requirement that there is only one input and one output variable in each instance. The methodology for constructing cognitive maps posits that all system variables are represented as vertices of a graph, and the interrelationships between them as arcs with weights that characterise the strength of influence (Roberts, 1976; Romanenko & Miliavskyi, 2023). Thus, a cognitive map is a formalised tool for modelling and analysing complex dynamic systems. Each value of the output variable can be calculated using the expression:

$$Y[k] = F(X[k-i], U[k-i]), i = \overline{1,d}, \quad (1)$$

where $F(\cdot)$ is the system model that links input variables to state variables and state variables to output variables, and d is the diameter of the graph, i.e., the length of the longest path between the input and output vertices. A signal propagates from the input vertex to the output vertex. Over one time step k, it moves from one vertex to another. Therefore, from the most distant vertex, it reaches the output in d steps, but from some vertices, the signal may arrive sooner. For a CM to be stable, the absolute values of all eigenvalues of the adjacency matrix must be less than 1 (Mokin *et al*, 2021).

The greater the number of vertices in a CM, the more patterns it accounts for, but the more difficult it is to ensure its stability. In conducting this research, the results of the articles by V. Mokin *et al.* (2020, 2021) were considered, in an attempt was made to synthesise guaranteed stable CMs mathematically. However, such CMs possess a rather simplified (somewhat degenerate) structure, which significantly reduces the set of complex systems for which they would be adequate. Conversely, if a CM has many vertices and a complex structure, it is more adequate for the processes within the system, taking into account features that correspond to external influences upon it. Thus, the main criterion for CM optimisation is finding a compromise between the number of vertices, to ensure maximum adequacy, and ensuring its stability, so that it can be used for tasks such as modelling, scenario-based data forecasting, and supporting optimal decision-making. In this context, LLMs can become an effective tool for solving CM optimisation problems. Specifically, LLMs are capable of automating CM construction by analysing textual and numerical data. Thanks to their ability to work with large volumes of data and account for complex hidden patterns, LLMs can help to achieve a compromise between the model's adequacy and its stability.

The proposed method is based on decomposing the task of forming a CM into a series of subtasks, which are solved using the cooperation of LLMs. This decomposition allows the overall problem to be broken down into individual steps: identification of variables (CM vertices), data preparation and generalisation, construction of CM weighting coefficients, and subsequent integration of the obtained results. Thus, to address the stated problem, an algorithm was developed for the method of constructing a cognitive map of processes in a dynamic system using the cooperation of large language models:

1. Identification of the main vertices-variables of the cognitive map using an ensemble of LLM M_{LLM1}, which most fully characterise the complex system being modelled and, simultaneously, are best supported by data for expert analysis.

2. Transformation of data for a given output value $Y[k]$ (corresponding numerical values from the sets of input variables U, state variables X, and the output variable; a textual description of the general characteristics of the object and each of its components that can be identified within it based on various criteria; a textual description of the current stage of the object's functioning, etc.) using an ensemble of LLM M_{LLM2} into a natural language textual description Ω.

3. Estimation of the upper bound of values for each indicator using an ensemble of LLM M_{LLM3} for subsequent use during the normalisation of indicator values.

4. Generating the weights of the cognitive map using the values of Ω and the LLM ensemble M_{LLM4}, taking into account typical constraints on these values in the range $[-1, 1]$ and ensuring its stability.

5. Checking the cognitive map for stability. If the CM is stable, the problem is solved; otherwise, revert to stages 1, 2, 3, or 4 and repeat them with different parameters (e.g., change the "temperature" parameter, which in LLMs is responsible for the diversity of the output, or similar).

This algorithm ensured the decomposition described above and the utilisation of the advantages of large language model cooperation at each stage of cognitive map formation. Breaking down the task into a sequence of prompts, instead of using a single complex query, significantly reduced the requirements for the context volume and capabilities of the LLMs, allowing for the creation of more detailed and structurally complex cognitive models. Furthermore, the cooperation of LLMs facilitated the combination of the strengths of different neural network architectures to achieve a synergistic effect.

Testing of the proposed method was conducted using real data on the water quality in the Sabarivske Reservoir on the Southern Bug River near Vinnytsia. The data were obtained from the Vinnytsia City Open Data Portal and included average monthly values of water quality indicators (Vinnytsia City Council, 2024). The CM generated using the developed method was evaluated for stability and forecasting accuracy.

Results and Discussion

Designing the architecture for the cooperation of large language models

Common methods for combining multiple LLMs can be divided into two categories. Direct model merging – for instance, parameter merging, which involves combining

several LLMs into one by aggregating their parameters (model weights), for example, through averaging. A key requirement here is that the model architectures must be compatible (Akiba *et al.*, 2025). The second method is the combination of input and output data between models (cooperation of LLMs), which includes ensemble methods (LLM ensemble) or other, more general cooperation techniques. Ensemble methods involve combining texts generated by multiple LLMs to improve the quality of the response and can occur either directly during generation or after the text has been fully generated (Cao *et al.*, 2024; Lu *et al.*, 2024).

The second method is, in general, more promising as it offers flexibility and does not require full compatibility with model architectures. Furthermore, cooperation-based approaches allow for a more effective combination of the strengths of different LLMs, which can significantly enhance the overall response quality without the need to modify the models themselves. Therefore, the use of ensemble methods (LLM ensemble) has been proposed.

The proposed architecture for the cooperation of large language models is based on the principles of ensemble methods in machine learning and collective decision-making. The theoretical foundation of this approach is the concept of "the wisdom of crowds", which suggests that aggregating independent evaluations often leads to more accurate results than individual assessments (Schoenegger *et al.*, 2024). In the context of LLMs, this concept is implemented through the parallel or sequential application of models, followed by the reconciliation of their outputs. To ensure the reliability of results and reduce the risk of "hallucinations" by the models, an architecture with result validation is applied, where the outputs of multiple LLMs are compared and reconciled by a dedicated reconciliation model.

In general, the nature and complexity of each ensemble depend on the specifics of the task. To minimise the typical "hallucinations" of LLMs and improve the reliability of results, it is suggested to use at least several LLMs to generate intermediate results, which are then summarised by another LLM (Fig. 1) (Das & Srihari, 2024). The presence of "hallucinations" in LLMs is a known issue, as the models can generate false or incorrect information, which may nonetheless appear plausible, making it difficult to detect and verify (Huang *et al.*, 2025). One possible reason for this may be that LLMs operate based on statistical patterns and lack a "true understanding" of the text (Bender & Koller, 2020; Bender *et al.*, 2021). In certain cases, some sub-tasks (for example, data conversion to another format) can be performed by individual LLMs without involving an ensemble or even algorithmically – for instance, S.-W. Chen & H.-J. Hsu (2023) demonstrated that integrating an external numerical module significantly reduces numerical hallucinations in Mistral 7B, improving the accuracy of mathematical calculations.

Figure 1 illustrates the structure of a large language model ensemble, where the input prompt P undergoes parallel processing through k independent LLMs (M_1, M_2,..., M_k). Each prompt is initially transformed by pre-processing operators φ_1, φ_2,..., φ_k, which can adapt the query to the specifics of the respective model. Inference parameters ψ_1, ψ_2,..., ψ_k define the configuration of each model (temperature, top_p, etc.). The results from all models are collected by operator T, undergo post-processing φ_{k+1}, and are fed into the coordinating model M_{k+1} with its own parameters φ_{k+1}, which forms the final result Y based on all the data received.

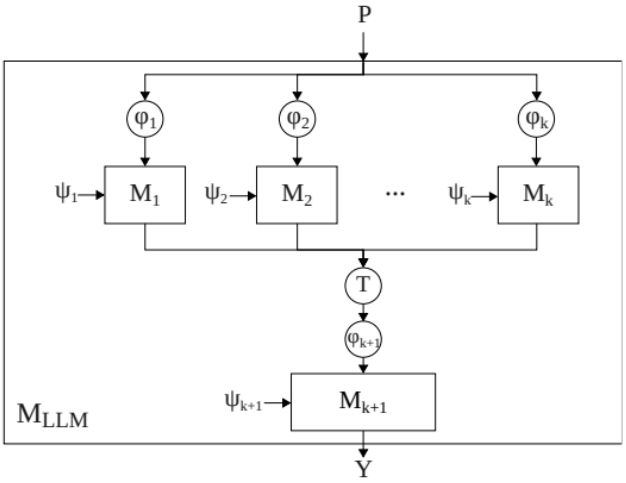

Figure 1. Example of LLM ensemble with parallel prompt processing using k LLM M_1, M_2,..., M_k and subsequent combination of results by a coordinating model M_{k+1}
Source: authors' development

The following set of models was defined: $\Lambda = \{M_1, M_2,..., M_n\}$, where each model M_i maps the prompt P to the response space R_i of model M_i: $M_i : \{P\} \to R_i$. If a single prompt P is simultaneously fed as input to all elements of a subset of models $\{M_1, M_2,..., M_k\} \subseteq \Lambda$, a set of results $Y^{Parallel}$ from the parallel processing of the prompt by the LLM ensemble will be obtained:

$$\left\{Y_1^{\text{Parallel}}, Y_2^{\text{Parallel}}, \ldots, Y_k^{\text{Parallel}}\right\} = \left[M_1^{\psi_1}, M_2^{\psi_2}, \ldots, , M_k^{\psi_k}\right](P), \quad (2)$$

where ψ_i are the inference parameters of model M_i.

When considering an ordered set $\{M_1, M_2,..., M_k\} \subseteq \Lambda$, where the prompt was fed as input to model M_1, and the response of each model was passed to the next in order within the set, the result $Y^{\text{Sequential}}$ of sequential prompt processing by the LLM ensemble was obtained:

$$Y^{Sequential} = M_k^{\psi_k}\left(M_{k-1}^{\psi_{k-1}}\left(\ldots M_1^{\psi_1}(P)\right)\right) = \\ = \left(M_k^{\psi_k} \circ M_{k-1}^{\psi_{k-1}} \circ \ldots \circ M_1^{\psi_1}\right)(P). \quad (3)$$

When combining the parallel (2) and sequential (3) approaches to prompt processing, a method was applied where the prompt was first fed simultaneously to the input of a subset of models $\{M_1, M_2,..., M_k\} \subseteq \Lambda$. Subsequently, the obtained responses from the models, $Y^{Parallel}$, were used as input data for the next sequence of models $\{M_{k+1}, M_{k+2},..., M_m\} \subseteq \Lambda$ for further processing:

$$Y = \left(M_m^{\psi_m} \circ \varphi_m \circ M_{m-1}^{\psi_{m-1}} \circ \varphi_{m-1} \circ \ldots \circ M_{k+1}^{\psi_{k+1}} \circ \varphi_{k+1}\right) \times \times \left(T\left(\left[M_1^{\psi_1} \circ \varphi_1, M_2^{\psi_2} \circ \varphi_2, \ldots, M_k^{\psi_k} \circ \varphi_k\right](P)\right)\right), \quad (4)$$

where $T: R_1 \times R_2 \times \ldots R_k \to P$ is the function for transforming the results of parallel processing into a new prompt, and $\varphi_i: \{P\} \to \{P\}$ is the prompt pre-processing operator before feeding it to the input of the model. For the implementation of the proposed method, four types of LLM ensembles are required: M_{LLM1}, M_{LLM2}, M_{LLM3}, M_{LLM4}, where each of the ensembles can be represented as:

$$M_{LLM_i}(P) = F_i\left(\left(M_m^{\psi_m} \circ \varphi_m \circ M_{m-1}^{\psi_{m-1}} \circ \varphi_{m-1} \circ \ldots \circ M_{k+1}^{\psi_{k+1}} \circ \varphi_{k+1}\right) \times \times \left(T\left(\left[M_1^{\psi_1} \circ \varphi_1, M_2^{\psi_2} \circ \varphi_2, \ldots, M_k^{\psi_k} \circ \varphi_k\right](P)\right)\right)\right). \quad (5)$$

As noted above, the M_{LLM1} ensemble was responsible for analysing the input textual description of the system and identifying the key elements (vertices of the cognitive map). The M_{LLM2} ensemble is intended for generalisation and conversion of heterogeneous data into a textual description , suitable for further processing. The M_{LLM3} ensemble is designated for analysing the available data regarding the system's state and determining the value limits for the CM vertices. This is necessary to normalise the values of the variables and bring them to a single scale. The M_{LLM4} ensemble determined the nature and intensity of influence between system elements based on the available textual descriptions, which allows for the formation of the CM weights.

For all four ensembles $\{M_{LLM1}, M_{LLM2}, M_{LLM3}, M_{LLM4}\}$, a common set of models L was used, and the ensemble architecture involved the parallel generation of results with their subsequent reconciliation by a coordinating model. The set L included: GPT-4o (OpenAI), Claude Sonnet 3.5 (Anthropic) – general multimodal large language models, Gemini 1.5 Flash (Google) – a model with high context processing performance, GPT-1o preview (OpenAI) – a model trained in a particular manner that demonstrates higher efficiency in solving complex tasks. The GPT-1o preview model was used as the coordinating model.

Development of a prompt system for LLM ensembles

To implement the idea proposed above, a system of prompts was developed that allows for the effective utilisation of LLMs in constructing a cognitive map. This process includes the processing of available data regarding the system being modelled, the selection of indicators based on which the CM will be built, the calculation of influence weights between indicators, the calculation of indicator limits for normalising state vectors in the CM, and the interaction between LLMs. The use of a systematic approach to creating these prompts enabled all aspects of the cognitive map construction process to be taken into account.

It is known that the effectiveness of LLM responses increases if the prompt explicitly sets the "role" that the LLM is to "perform" (Kong *et al.*, 2024; Wang *et al.*, 2024). This functions as setting a context that limits the semantic space for generating responses. Therefore, it is proposed that each developed prompt begins with a sentence that explicitly defines the role. Each prompt can be decomposed into constituent parts and represented as a tuple:

$$P = (D, R, I(D), C, F), \quad (6)$$

where D is the task context in terms of the subject domain, R is the role instruction for the LLM, $I(D)$ is instructions regarding data processing, C is additional conditions and constraints, and F is instructions regarding the response format.

D ("Data") – the domain context, which contains data relevant to the task (e.g., a textual description of the system state, observation JSON data, etc.). The domain context should provide the LLM with sufficient information for decision-making and may consist of several independent parts, for instance: D_E – a list of elements of the system being modelled, D_1, D_2, D_3 – evaluations from three independent experts that need to be generalised, and so forth. The structured presentation of data is justified by the need for a clear distribution of information blocks to avoid confusion and ensure flexibility in adaptation to different scenarios.

R ("Role") – the "role" instruction for the LLM (e.g., "You are an expert in assessing water quality in river basins"). Such a role assignment sets a specific context for the prompt and helps to obtain more specialised responses. The use of role instructions allows the LLM to adapt the style, terminology, and depth of the generated text according to the chosen area of expertise.

$I(D)$ ("Instructions") – instructions on precisely what needs to be done with the data. The instructions should be precise, understandable, and detailed. Correctly formulated instructions ensure a more accurate generation result. The more detailed the expected actions are described, the lower the probability of obtaining an incorrect interpretation or a result that does not meet expectations.

C ("Constraints") – additional conditions, formal limitations, and clarifications. Additional conditions are introduced to prevent the LLM from deviating beyond the defined semantic and contextual space. Constraints help to reduce the risk of obtaining incorrect results and also ensure compliance with practical requirements.

F ("Format") – instructions regarding the response format. Clear requirements for the response format are justified by the necessity for automated subsequent processing, integration with other tools, or verification of results. The absence of a formalised format would complicate the application of the generated data in real-world scenarios, reducing the overall effectiveness of the method.

This formalisation ensured a structured approach to prompt construction, which, in turn, enhanced the manageability and transparency of the result generation

process. Without such a clear structure, prompts could be inconsistent, overly complex, or insufficiently formalised, which would lower the quality of the generated CMs and their practical applicability. Furthermore, the formalised approach simplifies the process of debugging and optimising prompts, as it allows for a systematic analysis of the influence of individual components on the final generation result and enables targeted changes to improve the quality of the output.

Example application of the method for forecasting surface water quality

The proposed method was applied in practice using real data. The Vinnytsia City Open Data Portal provides average monthly values for water quality indicators in the Sabarivske Reservoir on the Southern Bug River, upstream of the drinking water intake for Vinnytsia Vodokanal (Vinnytsia City Council, 2024). Significantly more data are available on the water intake itself, but these are not published. Public data are primarily needed by the city authorities and population for using the Sabarivske Reservoir for recreational purposes, fishing, and so forth. However, for these purposes, knowledge of future values is more valuable than retrospective ones. To implement the described approach, the ensembles for solving this problem were defined as follows:

The set L is common to all four ensembles: L = (GPT-4o, Claude Sonnet 3.5, Gemini 1.5 Flash, GPT-1o preview). The architecture of the ensembles is analogous to that shown in Figure 1. A query to the ensemble indicated that for solving the problem of forecasting the ecological state and water quality indicators in the river, significantly more indicators measured at different points upstream and downstream with greater regularity are needed, along with information about water discharge, hydrological parameters of the river (flow velocity, sinuosity, roughness of the channel bed, turbulent diffusion coefficient), meteorological conditions (precipitation, temperature, atmospheric pressure), and so forth. Such data are absent in this dataset, and the available data are very limited (one data point, averaging interval – one month, information on hydrological and meteorological parameters is absent, information on hydrobiological indicators is absent). Therefore, it is impossible to identify either a mathematical or an intelligent model. However, building a cognitive map is, theoretically, possible. Taking expert knowledge of the research objects into account suggests that the best period for modelling with the available data is the winter season, as during this time, the impact of biotic indicators on the ecological condition of the water is minimal, and physical-chemical characteristics (temperature, oxygen concentration) are predominantly determined by abiotic processes. The LLM GPT-4o was used to formulate this assumption, followed by verification by a human expert. According to the algorithm of the proposed method, a textual description of the water state in January 2024 was generated using the ensemble. Subsequently, this description was used in prompts as the structural element D from formula (6). An example prompt is shown in Figure 2.

You are an expert in assessing water quality in river basins.	R
From the provided textual description of the water state in the river section and the list of indicators, determine the degree of influence of the indicator "dissolved oxygen concentration (mg/L)" on the rest of the indicators	I
The degree of influence of an indicator on each other indicator must be expressed numerically within the range of −1 to 1, and the absolute value of the sum of all influences must be less than 1	C_1
The sign of the influence should reflect how the indicator changes under the influence of "dissolved oxygen level (mg/L)": if with an increase in "dissolved oxygen level (mg/L)", the indicator increases, then the sign of the influence should be positive. If with an increase in "dissolved oxygen level (mg/L)", the indicator decreases, then the sign of the influence should be negative	C_2
D1 <description> D1 </description>	D_1
D2 <list of indicators> D2 </list of indicators>	D_2

Provide the final answer at the end in JSON format
{
"Dissolved oxygen level (mg/L)": {
"Indicator": influence
}
}
	F

Figure 2. Prompt $P_{weights} = (D_1, D_2, R, I, C_1, C_2, F)$ for generating relationships between indicators

Source: authors' development

According to the developed method, a cognitive map was generated, depicted in Figure 3. It reflects five key water quality indicators: biological oxygen demand, temperature, ammonium salt, dissolved oxygen, and total iron, as well as their interrelationships. These indicators were selected for modelling using the LLM ensemble. They are minimally sufficient for identifying certain patterns regarding water status, and it is for these indicators that the most data are available on the Vinnytsia City Council portal (Vinnytsia City Council, 2024). The weighting coefficients of the relationships between them were generated according to the proposed method.

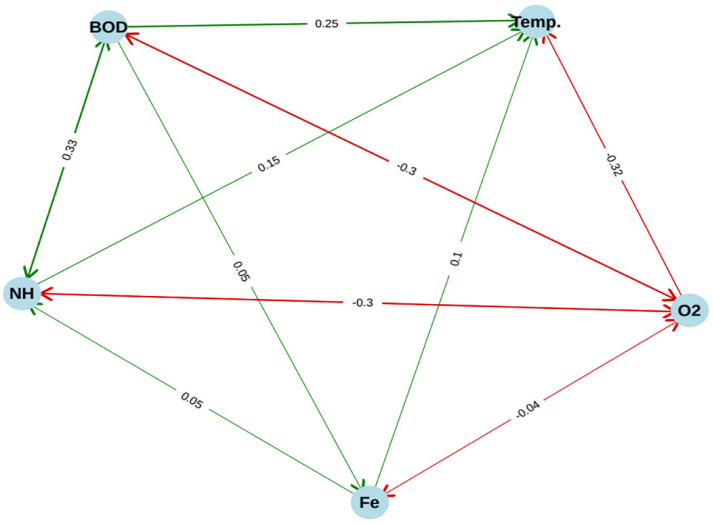

Figure 3. Generated cognitive map of water quality indicators in the Southern Bug River near Vinnytsia for the winter period

Note: BOD – biochemical oxygen demand; Temp – temperature; NH – ammonium salt (NH4+); O2 – dissolved oxygen; Fe – total iron. Green lines indicate a positive influence, where an increase in one indicator leads to an increase in another; red lines indicate a negative influence, where an increase in one indicator causes a decrease in another. Numerical values on the arcs reflect the strength of the mutual influence of the indicators in the range from –1 to 1, where larger absolute values indicate a stronger relationship between the indicators

Source: authors' development

Based on the obtained CM, cognitive modelling of a temperature change scenario for the following month, February 2024, was performed. To verify the reliability of the forecast, the modelling results were compared with the actual temperature and water quality indicators for February 2024. The obtained results are presented in Table 1.

Table 1. Comparison of modelling results with actual water quality indicators

Indicator	Actual value	Model prediction	Relative modelling error (%)
Dissolved oxygen concentration	9.5 mgO_2/dm^3	9.125 mgO_2/dm^3	3.95%
Ammonium salt (NH4+)	0.43 mg/dm^3	0.421 mg/dm^3	2.09%
BOD (biochemical oxygen demand)	6.3 mgO_2/dm^3	6.01 mgO_2/dm^3	4.6%
Total iron	0.35 mg/dm^3	0.6 mg/dm^3	71.4%

Source: authors' development

As can be seen from the table, three out of the four indicators were modelled with high accuracy. This indicates the effectiveness of the applied method in forecasting the main physicochemical parameters of water. However, for the "Total iron" indicator, the relative error was 71.4%. Such a significant discrepancy may be due to additional factors that were not taken into account in the modelling. This suggests that for this indicator, the aforementioned assumption regarding the greater adequacy of the model for the winter period is not significant. This is quite expected, as the concentration of iron in water is considerably less related to the activity of aquatic organisms than other indicators of the ecological state of the water.

Based on the generated cognitive maps, which were produced by the LLM ensemble, there is partial overlap with the results of R. Schuerkamp *et al.* (2025). In their experiment, ChatGPT successfully merged several expert maps into one; in the present study, the LLM ensemble managed to integrate disparate fragments of knowledge from text into a coherent map. This integration process proved particularly effective for complex systems where the interrelationships between components are not always immediately obvious. The use of a model ensemble allowed for a more balanced representation of knowledge and helped to avoid the potential biases of individual models. Thus, the findings support the idea that LLMs can act as knowledge integrators, forming a cognitive model of a system from potentially conflicting statements derived from expert cognitive maps. Simultaneously, this research goes further by employing LLMs to construct the map from scratch, rather than merely merging existing maps.

Compared to the approach of A. Feleki *et al.* (2023), where GPT-3.5 generated textual explanations for an already existing FCM, the method described in this study effectively does the opposite – it generates the FCM itself.

Instead of requiring the black box to explain its decisions in human language, the black box is "compelled" to explain the problem in the language of causal relationship graphs. This "inverted" approach has made it possible to obtain a model that is more interpretable from the outset. This is particularly important in the context of increasing demands for the transparency of algorithmic solutions and the necessity to explain not just individual conclusions, but also the general logic of the model's reasoning. Furthermore, such an approach potentially reduces the time required for creating cognitive maps and makes this tool more accessible to researchers without deep expertise in modelling complex systems.

In the context of dynamic systems, it is worth comparing the obtained results with those of T. Liu *et al*. (2024). That research demonstrated the ability of LLMs to forecast time series, effectively imitating the system dynamics through a sequence. The proposed method, however, aims for an explicit representation of dynamics via a graph of influences, which makes it more transparent and interpretable for end-users. In this research, attention was focused on the structure of interrelationships between the system components, rather than solely on their behaviour over time. It can be noted that the model constructed by the LLM in the conducted study successfully reproduced the qualitative structure of the system (the set of influences between variables). This allowed not only for forecasting changes in the system but also for understanding the causes of these changes and potential levers of influence. Thus, this complemented the results of T. Liu *et al*. (2024), as LLMs can not only predict behaviour but also reveal the structure of interrelationships, which significantly expands the analytical toolkit for working with complex systems in various subject domains.

Compared with Ukrainian studies, it can be stated that the solution presented in this research fills an important gap. In the article by S. Shevchenko *et al*. (2024), experts manually modelled cybersecurity risks via FCMs. Similarly, in the study by O. Cherniuk (2023), a map of the impact of chatbots on education was constructed based on expert analysis. The proposed method, in contrast, allows for the automation of such steps: instead of manual map construction, it is sufficient to provide the LLM with a full description of the problem. This significantly accelerates the modelling process and makes it less dependent on the availability of specific experts. Moreover, the automated approach may prove less susceptible to individual expert biases, especially if an ensemble of different LLMs is utilised. At the same time, the proposed method does not exclude expert contribution but rather shifts it to the level of validation and correction of automatically generated models, which optimises the use of valuable expert time and knowledge.

Conclusions

The article addresses the issue of using LLMs as experts for constructing cognitive maps of complex dynamic systems. The research goal, which was to develop a method for constructing a cognitive map of processes in a dynamic system using the cooperation of large language models in the role of experts, has been successfully achieved. The proposed method allows the automation of the cognitive map construction process without the involvement of human experts and requires only a minimal set of input data.

In the course of the study, the architecture for the cooperation of LLM ensembles was proposed and substantiated for the formal generation of vertices-variables and weight coefficients in cognitive maps. The decomposition of a typical prompt into structural components was carried out, and approaches for their definition were suggested. A system of prompts was developed to ensure structured data processing and the identification of relationships between system elements. The practical effectiveness of the approach was demonstrated through the example of predicting water quality in the Sabarivske Reservoir, where modelling for three of the four physicochemical indicators (dissolved oxygen concentration, ammonium salt, and biological oxygen demand) showed a small error (2%-5%), even with a minimal amount of input data.

The proposed method holds significant importance for the applied modelling of complex systems, as it allows for the rapid creation of formalised models in situations where traditional approaches face limitations due to a lack of data or experts. The developed formal system of prompts helps to improve the accuracy of LLM responses, thanks to its structured nature and the use of known techniques for enhancing the quality of the generated text. Utilising ensembles of different LLMs instead of a single model helps to minimise potential biases and "hallucinations", which is critically important when modelling real-world systems. For one of the four physicochemical indicators ("Total iron"), the method showed a significant error, indicating the necessity of considering additional factors for specific system variables. This confirms that cognitive maps constructed with the assistance of LLMs require validation and potential adjustment for specific variables or relationships.

Promising directions for future research include expanding the LLM cooperation architecture to work with multimodal data, improving methods for ensuring the stability of generated cognitive maps and developing approaches for the automatic correction of cognitive maps based on feedback from real data. Separate attention is needed to investigate the effectiveness of the proposed method in other subject domains, particularly for complex technical, ecological, and socio-economic systems.

Acknowledgements

None.

Funding

The study received no funding.

Conflict of Interest

None.

References

[1] Akiba, T., Shing, M., Tang, Y. & Ha, D. (2025). Evolutionary optimization of model merging recipes. *Nature Machine Intelligence*, 7(2), 195-204. doi: 10.1038/s42256-024-00975-8.

[2] Bender, E.M., & Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 5185-5198). Stroudsburg: Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.463.

[3] Bender, E.M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency (FAccT'21)* (pp. 610-623). Stroudsburg: Association for Computing Machinery. doi: 10.1145/3442188.3445922.

[4] Cao, H., Ma, R., Zhai, Y., & Shen, J. (2024). LLM-Collab: A framework for enhancing task planning via chain-of-thought and multi-agent collaboration. *Applied Computing and Intelligence*, 4(2), 328-348. doi: 10.3934/aci.2024019.

[5] Chen, S.-W., & Hsu, H.-J. (2023). MisCaltral: Reducing numeric hallucinations of mistral with precision numeric calculation. *Research Square*. doi: 10.21203/rs.3.rs-3789011/v1.

[6] Cherniuk, O. (2023). Modeling the impact of AI-based chatbots on the quality of higher education using system analysis methods. *Information Technologies and Society*, 3(9), 80-90. doi: 10.32689/maup.it.2023.3.11.

[7] Das, S., & Srihari, R. (2024). Compos mentis at SemEval2024 Task6: A multi-faceted role-based large language model ensemble to detect hallucination. In *Proceedings of the 18th international workshop on semantic evaluation (SemEval-2024)* (pp. 1449-1454). Mexico City: Association for Computational Linguistics. doi: 10.18653/v1/2024.semeval-1.208.

[8] Feleki, A., Apostolopoulos, I.D., Moustakidis, S., Papageorgiou, E.I., Papathanasiou, N., Apostolopoulos, D., & Papandrianos, N. (2023). Explainable deep fuzzy cognitive map diagnosis of coronary artery disease: Integrating myocardial perfusion imaging, clinical data, and natural language insights. *Applied Sciences*, 13(21), article number 11953. doi: 10.3390/app132111953.

[9] Godoy, W.F., Fabri, J.A., Palácios, R.H.C., Mendonça, M., Gonçalves, J.F.S., & Moraes, L.O.M. (2024). Using fuzzy cognitive maps and chatbots to evaluate student satisfaction in a university: A comparison between strong and weak AI. In *Proceedings of the eighteenth international conference on mobile ubiquitous computing, systems, services and technologies* (pp. 16-20). Wilmington: IARIA Press.

[10] Huang, L., *et al.* (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2), article number 42. doi: 10.1145/3703155.

[11] Kong, A., Zhao, S., Chen, H., Li, Q., Qin, Y., Sun, R., Zhou, X., Wang, E., & Dong, X. (2024). Better zero-shot reasoning with role-play prompting. In *Proceedings of the 2024 conference of the North American chapter of the association for computational linguistics: Human language technologies* (Vol. 1, pp. 4099-4113). Mexico City: Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.228.

[12] Liu, T.J., Boulle, N., Sarfati, R., & Earls, C. (2024). LLMs learn governing principles of dynamical systems, revealing an in-context neural scaling law. In *Proceedings of the 2024 conference on empirical methods in natural language processing* (pp. 15097-15117). doi: 10.18653/v1/2024.emnlp-main.842.

[13] Lu, J., Pang, Z., Xiao, M., Zhu, Y., Xia, R., & Zhang, J. (2024). Merge, ensemble, and cooperate! A survey on collaborative strategies in the era of large language models. *ArXiv*. doi: 10.48550/arXiv.2407.06089.

[14] Mokin, V.B., Burdeina, O.V., & Varchuk, I.V. (2020). On the optimization of topologically observable cognitive maps while preserving their robustness. *Visnyk of Vinnytsia Polytechnic Institute*, (6), 84-92. doi: 10.31649/1997-9266-2020-153-6-84-92.

[15] Mokin, V.B., Dratovanyi, M.V., Kozachko, O.M., & Zhukov, S.O. (2021). Method for synthesizing a robust multi-connected cognitive map of a complex system. *Visnyk of Vinnytsia Polytechnic Institute*, (6), 114-122. doi: 10.31649/1997-9266-2021-159-6-114-122.

[16] Roberts, F. (1976). *Discrete mathematical models with applications to social, biological, and environmental problems*. Englewood Cliffs: Prentice-Hall.

[17] Romanenko, V., & Miliavskyi, Y. (2022). Coordinating control of a cognitive map impulse process in stochastic environment. *Problems of Control and Informatics*, 67(4), 49-58. doi: 10.34229/2786-6505-2022-4-4.

[18] Saliieva, O., & Yaremchuk, Yu. (2020). Study of the reliability of the impact of threats on the level of security of the information protection system and the object of critical infrastructure based on the results of cognitive modeling . *Bulletin of Cherkasy State Technological University*, 25(3), 85-93. doi: 10.24025/2306-4412.3.2020.216251.

[19] Schoenegger, P., Tuminauskaite, I., Park, P.S., Bastos, R.V.S., & Tetlock P.E. (2024). Wisdom of the silicon crowd: LLM ensemble prediction capabilities rival human crowd accuracy. *Science Advances*, 10(45), article number eadp1528. doi: 10.1126/sciadv.adp1528.

[20] Schuerkamp, R., Ahlstrom, H., & Giabbanelli, P.J. (2025). Automatically resolving conflicts between expert systems: An experimental approach using large language models and fuzzy cognitive maps from participatory modeling studies. *Knowledge-Based Systems*, 313, article number 113151. doi: 10.1016/j.knosys.2025.113151.

[21] Shevchenko, S., Zhdanova, Y.A., Kryvytska, O., Shevchenko, H., & Spasiteleva, S. (2024). *Fuzzy cognitive mapping as a scenario approach for information security risk analysis (short paper)*. In *Proceedings of the 2024 cybersecurity providing in information and telecommunication systems II* (pp. 356-362). Kyiv: Borys Grinchenko Kyiv Metropolitan University.
[22] Vinnytsia City Council. (2024). *Ecology and natural resources*. Retrieved from https://www.vmr.gov.ua/ecology.
[23] Wang, Z.M., *et al*. (2024). RoleLLM: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. In *Findings of the Association for computational linguistics (ACL 2024)* (pp. 14743-14777). Bangkok: Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.878.

Метод побудови когнітивної карти процесів у динамічній системі із використанням кооперації великих мовних моделей

Борис Варер
Аспірант
Вінницький національний технічний університет
21021, вул. Хмельницьке шосе, 95, м. Вінниця, Україна
https://orcid.org/0000-0002-5860-0100

Віталій Мокін
Доктор технічних наук, професор,
Вінницький національний технічний університет
21021, вул. Хмельницьке шосе, 95, м. Вінниця, Україна
https://orcid.org/0000-0003-1946-0202

Анотація. В умовах постійного зростання вимог до швидкого прийняття рішень і глибокого аналізу складних динамічних систем, коли доступні дані обмежені, а залучення досвідчених експертів часто є неможливим або занадто витратним, розроблення нових методів побудови моделей набуває особливої актуальності. Використання великих мовних моделей (LLM) як експертних систем дозволяє суттєво знизити ресурсні витрати та прискорити процес моделювання складних технічних, екологічних та соціально-економічних систем. Метою даної роботи було дослідження та практична демонстрація потенціалу та можливостей LLM, як експертних систем, у процесі побудови когнітивних карт. У даній роботі запропоновано та обґрунтовано архітектуру кооперації ансамблів LLM для формалізованого генерування вершин-змінних та вагових коефіцієнтів когнітивних карт, що дозволяє автоматизувати процес моделювання без залучення експертів-людей. Здійснено декомпозицію типового промпта (вказівки) до LLM на структурні складові: опис контексту (D), рольову настанову моделі (R), інструкцію (I), умови (C) та формат відповіді (F) та запропоновано підхід їх визначення експертним шляхом. Розроблено систему таких промптів, яка забезпечує структуроване оброблення даних та ідентифікацію взаємозв'язків між елементами системи. Практичну ефективність підходу продемонстровано на прикладі прогнозування стану води у Сабарівському водосховищі біля м. Вінниця, де для більшості фізико-хімічних показників моделювання продемонструвало малу похибку (2.09–4.60 %) навіть за мінімального обсягу вхідних даних. Запропонований метод є перспективним для задач моделювання та прогнозування у складних системах з обмеженим обсягом даних, зокрема в екологічних, соціально-економічних та інженерних сферах, де швидкість отримання надійних результатів має критичне значення для прийняття обґрунтованих рішень

Ключові слова: ВММ; генеративний штучний інтелект; інтелектуальна технологія; системний аналіз; моделювання; прогнозування; динамічна система

Journal homepage: https://itce.com.ua/en

Information Technologies and Computer Engineering **Vol. 22, No. 1. 2025**

Article's History: Received: 06.01.2025 Revised: 14.03.2025 Accepted: 24.04.2025

UDC 004.42:004.77 DOI: 10.63341/vitce/1.2025.79

Adaptive performance monitoring in cloud environments via recurrent neural networks

Pavlo Kudrynskyi*

Postgraduate Student
State University of Information and Communication Technologies
03110, 7 Solomianska Str., Kyiv, Ukraine
https://orcid.org/0009-0008-6314-6150

Oleksandr Zvenihorodskyi

PhD in Technical Sciences, Associate Professor
State University of Information and Communication Technologies
03110, 7 Solomianska Str., Kyiv, Ukraine
https://orcid.org/0009-0008-6235-1638

Abstract. The study aimed to develop an adaptive methodology for analysing the performance of cloud computing infrastructures to improve the efficiency of resource management and reduce maintenance costs. The research addressed the implementation of the latest approaches to automate monitoring and analysis processes. The research methodology included the integration of data from monitoring platforms (Amazon Web Services CloudWatch, Google Cloud Monitoring, Prometheus) to collect key performance indicators. Data processing was conducted using Python libraries (NumPy, pandas, scikit-learn) to detect anomalies and generate time series. Recurrent neural networks and long-short-term memories based on TensorFlow and PyTorch were used to model performance. The implementation of continuous learning was used to adapt the models to the changing conditions of cloud systems in real-time. The main results of the study include the creation of an innovative system for predicting key performance metrics of cloud infrastructures with high accuracy. This was confirmed using the mean absolute error and root mean square error metrics. Real-time data integration was provided through the Amazon Kinesis platform, and visualisation and management were performed using Amazon CloudWatch and Grafana dashboards. Virtual machines and containers interacted with Nova, Glance, Cinder, and Neutron modules, and the Keystone module provided security through authentication and authorisation. Automatic resource scaling based on neural networks optimised the use of computing, network and storage resources. The developed methodology can be used to automate the management of cloud resources, reducing the need for manual intervention and cutting costs. The proposed method provided high speed due to interaction via REST and HTTPS and collected data in a time series format for primary processing. The integration of OpenStack with Apache Spark and the use of a high-speed data channel has increased the efficiency of the infrastructure. The findings demonstrated that the implementation of this methodology significantly increases the efficiency of cloud infrastructure management

Keywords: cloud computing systems; resource performance; time series; continuous learning; cloud infrastructure optimisation; load forecasting

Introduction

The research relevance is determined by the rapid development of cloud computing systems, which have become the basis of modern information technologies. The main problems are the lack of universal evaluation methods, the complexity of performance forecasting, and the high cost of experimental testing. Effective estimation methods will help optimise resource management, reduce costs and ensure the stability of services. The use of neural network models and optimisation methods will help create adaptive algorithms for load forecasting. This will increase the reliability, scalability and efficiency of cloud solutions. Cloud system performance was studied in several areas: load

Suggested Citation:
Kudrynskyi, P., & Zvenihorodskyi, O. (2025). Adaptive performance monitoring in cloud environments via recurrent neural networks. *Information Technologies and Computer Engineering*, 22(1), 79-92. doi: 10.63341/vitce/1.2025.79

*Corresponding author

modelling, performance forecasting, resource balancing, energy efficiency, and architectural analysis.

The scientific novelty of the proposed topic was the development and implementation of an adaptive approach to assessing the performance of cloud infrastructures, which used neural network technologies for dynamic analysis and management of resources in real time. The integration of recurrent neural networks and long-term memory models with monitoring systems enabled real-time data acquisition, increasing the efficiency and accuracy of response to changes in the cloud environment. The proposed methodology of continuous learning and adaptation of neural networks improved predictions and adaptation to new conditions.

The need for effective monitoring of complex computing systems to prevent failures and detect anomalies was studied by R.Y. Zahvoyskyi & I.Y. Kazymyra (2024), proposing various approaches to anomaly detection using mathematical modelling and machine learning. As a result, models were obtained to detect changes in system parameters and create early warnings. Threats to the cybersecurity of cloud services, including DDoS attacks and data leaks, were analysed by A. Vavilenkova (2024), who studied the vulnerabilities of cloud environments and ways to protect them. The problem of cybersecurity of cloud services is related to their vulnerability to attacks and unauthorised access, which can lead to data leaks and disruption of critical systems. DDoS attacks cause infrastructure overload, reduce productivity, and can completely disrupt services. Data leaks caused by insufficiently secure configurations or insider threats put users' confidential information at risk. Methods for optimising firewall configurations and analysing logs using AI have been developed.

Methods of investigating cybercrime in cloud environments were analysed by I. Opirskyy *et al.* (2021). The authors analysed the use of traps to collect information about potential attacks in cloud infrastructures. The authors identified the benefits of traps that can reduce infrastructure costs and increase the efficiency of investigations. Approaches to optimising the use of cloud system resources were developed by M. Dorosh *et al.* (2020). The authors analysed methods for optimising cloud infrastructures and resource management models. As a result, algorithms for dynamic resource allocation that ensure greater productivity and uninterrupted operation of systems were developed. Methods of risk assessment in cloud systems were considered by L. Nikitina *et al.* (2024). The authors proposed threat assessment models and methods for improving the cyber resilience of cloud systems. The authors developed an expert system for risk assessment that identifies potential threats and creates recommendations for their minimisation.

The mechanisms of access rights management were studied by A. Kuprienko & L. Galchynskyi (2023). The authors proposed approaches to optimising access control through role models and attribute-based methods. As a result, an agent-based model of access rights mining that adapts to changes in the cloud environment was developed. The architecture of adaptive content management systems in cloud environments was analysed by Y. Baytelman & V. Potsepaiev (2024). The authors developed a content management system that uses PHP as a server-side language to create dynamic web pages, MySQL to store and manage data in a database, and Amazon Web Services to host, scale, and manage infrastructure in the cloud. This ensured the high availability, flexibility and reliability of the system with dynamic loads, which is typical for cloud environments.

The problem of automating infrastructure deployment in the cloud environment was studied by S. Behlitsov (2024). The author analysed the automation of cloud resources using Terraform and Amazon Web Services to improve performance. Methods were developed to automate the deployment of Amazon Web Services resources through Terraform, including Elastic Compute Cloud, Simple Storage Service, and Virtual Private Cloud, with state management for easy collaboration.

An analysis of scientific sources demonstrated the urgency of the problem of assessing and improving the efficiency of cloud environments, which determined the need to develop new approaches to their research. The study aimed to develop models and methods for analysing the performance of cloud infrastructures based on optimising the processes of collecting key performance indicators. To achieve this goal, it was necessary to study the available methods and tools for analysing cloud computing performance and determine the most effective ones in different conditions and for different types of loads. Theoretical foundations and practical approaches used to monitor and evaluate the efficiency of cloud infrastructures were emphasised.

Materials and Methods

Existing methods for monitoring and evaluating the efficiency of cloud infrastructures were identified and analysed based on a detailed analysis of scientific publications. This included a review of existing approaches to collecting and processing cloud system performance data, such as the use of CPU, memory, network activity, and storage metrics. Performance assessment included methods based on time-series analysis, load forecasting and anomaly detection, including the use of statistical methods and machine learning algorithms. In addition, approaches to optimising resources by automatically adjusting system parameters that reduce maintenance costs and improve infrastructure efficiency were reviewed.

The virtual data centre infrastructure was modelled using a simulation model of a multi-channel queuing system. The following formula (1) was used to describe the dynamics of changes in the system state over time:

$$x(t+\Delta t) = \\ = x(t) - \Sigma K \Sigma N s_{i,j}(t)\, u_{j,k}(t) + \Sigma N s_{i,l}(t)\, u_{j,k}(t) + y_{i,j}(t), \quad (1)$$

where N – number of nodes included in the infrastructure; K – sum of the types of services and applications; $s_{i,j}(t)$ is

the channel width between the *i*-th node and the *j*-th storage; $y_{i,j}(t) = \lambda_{i,j}(t)\Delta t$ – traffic per time; $\lambda_{i,j}(t)$ – intensity of client requests; $u_{j,k}(t)$ – share of network channel bandwidth (i, l) of the flow of user requests to an application of type k that works with information in storage j.

The basis for developing an appropriate monitoring architecture, which included tools and technologies for collecting, storing and analysing performance data, such as OpenStack, and Apache Spark, was the management performance analysis methodology. An important component of this architecture was agentless monitoring, which was used to collect performance metrics such as CPU, memory, network bandwidth and storage usage across all components of the cloud infrastructure. This ensured that accurate and relevant data on system health could be obtained in real-time.

Predictive models based on neural networks were used to predict future workloads for in-advance planning and pre-processing of resources, increasing the efficiency of cloud infrastructure management. Detection and analysis of performance anomalies using deep learning provided quick response to unexpected changes and maintained stable system operation. The original model of collecting and processing metrics in the OpenStack cloud infrastructure with the integration of Apache Spark was used to analyse the collected metrics. The Gnocchi module collected metrics from the main components of the infrastructure, which provided real-time data on resource consumption. This data was stored in a time series format for further analysis in Apache Spark.

The initial model for collecting and processing metrics in the cloud was developed based on OpenStack, which included the main components of this infrastructure. The model provided for integration with the Apache Spark big data processing system to analyse the collected metrics, which evaluated the effectiveness of the proposed approach. The following main components were used in the initial architecture of the OpenStack cloud infrastructure:

1. Nova was used to virtualise computing resources for the instances.

2. Glance was used to store images of virtual machines and containers.

3. Cinder was used to manage block data stores for instances.

4. Neutron was used to manage network resources for virtual machines.

5. Keystone was used for authentication and authorisation of users and services.

6. Gnocchi was used to collect metrics and monitor the state of the instances.

To verify the system's performance, a test architecture was deployed, which included six servers for the main OpenStack components and seven servers for the Apache Spark deployment. The virtual machines and containers interacted directly with Nova, Glance, Cinder, and Neutron modules, ensuring the system's correct operation and collecting the necessary metrics. The infrastructure provided an effective metrics collection model that provides an initial basis for detecting abnormal patterns of instances' behaviour and optimising resource usage. The main idea of the model was to integrate OpenStack and Apache Spark components for collecting and processing metrics, which reduces the load on the system and ensures quick identification of potential problems in the infrastructure.

The collected data was thoroughly analysed to identify possible problems and bottlenecks in the system. This was done using time series analysis, statistical analysis and machine learning techniques. Time series analysis identified patterns in resource usage, which predicted future loads and adapted resources accordingly. Statistical analysis identified deviations from normal operating conditions and detected anomalies that may indicate problems in the system. Machine learning techniques were used to build prediction models and identify complex relationships between different performance metrics.

Critical performance indicators were defined, such as response time, throughput and reliability. An important aspect was an integrated approach to system lifecycle management, including planning, development, testing and operation. Analysis of changes, including software or hardware updates, was used to assess the potential impact on performance. In distributed cloud environments, factors such as network latency and bandwidth were incorporated. System scalability and data security were also important considerations, as high performance had to be balanced with data protection requirements. An important part of the performance analysis was the validation and verification of the data collected. This included checking the accuracy and completeness of the data, which ensured the reliability of the analysis results. Hence, methods of comparison with the original data, as well as testing on specially designed test environments, were used.

The performance analysis methodology also involved ongoing monitoring and support of the cloud system. This included regular updates to monitoring tools and methods, scheduled inspections and audits, and staff training to ensure a high level of professional competence. Ongoing monitoring provided timely detection and elimination of problems, which maintained a stable and high-performance cloud system.

Results

Modern methods for analysing cloud infrastructure performance

Modern performance analysis methods use a variety of tools and techniques to monitor, measure and analyse the performance of cloud applications and infrastructure. One of the main methods of performance analysis is monitoring, which involves continuously observing the state and performance of the cloud infrastructure. Monitoring can collect data on CPU utilisation, memory usage, network bandwidth, I/O latency, and other key metrics. Tools such as Amazon CloudWatch, Azure Monitor, and Google Cloud Monitoring provide detailed information about the status of cloud services and identify potential problems.

Another method is application profiling, which can be used to analyse the performance of individual application components in detail. Tools for this purpose include Amazon Web Services X-Ray, Google Cloud Profiler, and Microsoft Application Insights. Profiling identifies bottlenecks in code execution, such as slow functions or algorithms that consume excessive amounts of resources. This describes the application behaviour in real-time and optimises their performance (Malallah *et al.*, 2023). Stress testing can determine how the cloud infrastructure can withstand high loads and identify its performance limits. Stress testing using Apache JMeter, Locust, and Gatling includes modelling peak loads on the system, checking its resilience to failures, and determining the maximum number of requests that the system can process without degrading performance (Apeh *et al.*, 2023; Chinamanagonda, 2023).

Automation of performance management is also an important aspect of modern performance analysis techniques. The use of orchestration and automated scaling mechanisms can dynamically adjust resources according to the current load on the system, ensuring optimal resource utilisation and reducing costs. For this purpose, applications such as Kubernetes, Docker Swarm, and Amazon Web Services Auto Scaling are used to automatically manage containers and virtual machines based on defined policies and rules (Krishnaveni *et al.*, 2021; Ileana *et al.*, 2024).

Machine learning and artificial intelligence are also used in cloud performance analysis. The use of machine learning algorithms can predict workloads, detect anomalies, and make automatic adjustments in real-time. This is made possible by Google Cloud AI, Amazon Web Services SageMaker, and Azure Machine Learning, which can integrate machine learning into performance analysis and management processes (Bagai, 2024; Shaffi, 2025). Service Level Monitoring is another important aspect of performance analysis. Tools such as New Relic, Dynatrace, and AppDynamics can assess whether the performance of cloud services meets the requirements defined in service level agreements. This includes monitoring metrics such as response times, service availability and failure rates, and ensures that cloud services meet established requirements and standards.

Correlation and causality analysis are also important aspects of modern performance analysis. These methods can identify relationships between different performance indicators and events occurring in the system. Correlation analysis identifies how changes in one component of the system affect other components, which identifies the root causes of performance problems. For this purpose, Splunk and ELK Stack are used. They perform sophisticated analysis of log and event data to help identify correlations and cause-and-effect relationships (Duan *et al.*, 2025). Transaction tracing can be used to track the full path of individual transactions through all system components, from the initial request to the result. Tools such as Amazon Web Services X-Ray, Dynatrace, and AppDynamics help identify delays and bottlenecks in transaction execution, as well as determine which system components need to be optimised.

Distributed monitoring systems are used to collect performance data in large and complex cloud environments. These systems collect metrics from a variety of sources, such as virtual machines, containers, databases, and network devices, and combine them into a single dashboard. This provides a centralised view of system health and quickly identifies and resolves issues. Prometheus, Grafana, and Zabbix are popular solutions for distributed monitoring (Krishnan *et al.*, 2023; Rahman *et al.*, 2024). Synthetic monitoring involves the use of special scripts and tools to simulate user activity and measure application performance. This method can be used to verify the availability and performance of applications even in the absence of real users, which identifies issues before they affect end users. Tools such as Pingdom, Uptrends, and New Relic Synthetics were widely used for synthetic monitoring. Real user monitoring collects performance data directly from real users of applications. This method provides accurate information about how users interact with applications, what delays they experience, and where problems occur. Real user monitoring helps to ensure a high quality of user experience and improve the overall app experience. Tools such as Google Analytics, Dynatrace Real User Monitoring, and New Relic Browser are used for real-user monitoring.

Virtualisation and containerisation are the main technologies that use resources in cloud environments efficiently. The use of virtual machines and containers provides application isolation, flexibility in deployment and scaling, and optimal use of hardware resources (Anbalagan, 2024; Li *et al.*, 2024). Containerisation technologies, such as Docker and Kubernetes, simplify the deployment, scaling, and management of applications in cloud environments. Automated performance management using orchestration platforms such as Kubernetes can dynamically adjust resources according to the current system load. Automated scaling mechanisms can increase or decrease the number of resources used in real-time, which ensures optimal resource utilisation and reduces costs. Resource and service quality management policies define rules and criteria to ensure the required level of performance and reliability of cloud services. They prioritise business-critical applications and resources, ensuring smooth and continuous operations even during peak loads. QoS policies ensure high quality of service and meet user requirements.

Infrastructure performance characteristics and organisation of computing in cloud environments

Cloud infrastructure performance is a multidimensional concept that includes various aspects of system operation, such as data processing speed, resource availability, reliability, scalability, and resource efficiency (Table 1). Performance analysis identifies and measurement of key metrics that reflect the ability of a system to perform its functions at a high level. Evaluation of these metrics identifies bottlenecks, optimise resource utilisation, and improve the overall performance of the cloud infrastructure.

Table 1. Key criteria and performance indicators for cloud infrastructure

Criteria	Description	Indicators
Bandwidth	Defines the amount of data that the system can process in a certain period	Number of processed requests per second (TPS), network bandwidth (Mbps), amount of processed data (GB/sec)
Response time	Speed of system response to user requests	Average response time (ms), highest recorded response time (ms), 90th percentile of response times (ms).
Reliability	The ability of the system to operate continuously and without errors	The average interval between failures, the average time to restore the system after a failure, and the number of failures during a given period.
Availability	The proportion of time the system is available to users	Percentage of availability (%), average downtime (ms), number of unavailability incidents
Scalability	The ability of the system to adapt to load changes by adding or removing resources	Scaling time (sec), scaling efficiency (%), number of resources added/removed
Resource usage efficiency	Optimal use of computing, network and storage resources	CPU utilisation (%), memory utilisation (%), network bandwidth (%), storage utilisation (% utilisation)
Energy efficiency	The ability of the system to perform its functions with minimal energy consumption	Energy consumption (W), energy efficiency, cost of energy for data processing (UAH/GB)
Data security and confidentiality	Protecting data from unauthorised access and loss	Number of security incidents, average time to detect and resolve incidents, level of compliance with security standards

Source: compiled by the authors

The definition and measurement of key performance indicators of cloud infrastructures can be used to assess the efficiency and stability of the system. For instance, throughput determines the ability of a system to process large amounts of data, while response time affects the speed of response to user requests. Reliability and availability assess the system's resilience to failure, which is important for maintaining continuous operation. Scalability determines the system's ability to adapt to changes in load, and energy efficiency helps reduce energy costs. Data security and confidentiality protect unauthorised access and data integrity.

Regular performance monitoring and analysis identifies potential problems and bottlenecks in the system before they become critical. Modern monitoring tools can continuously monitor key performance indicators, such as CPU, memory, network bandwidth and disk space usage. Machine learning created capable prediction and optimisation methods in cloud environments. Deep learning models are used to predict the load on computing resources based on historical data, which preallocates resources in advance and avoids peak overloads.

The process of organising cloud computing involves the deployment of interconnected data centres. The user accesses the resources without concern for the technical details, which are handled by the cloud service operator. The user agrees with the operator (e.g., Amazon Amazon Web Services, Microsoft Azure) and receives a remote server on which the necessary software is installed. Server setup can be automated using configurations such as:

1. SaaS – software rental.
2. PaaS – a platform for application development.
3. DaaS – workplace rental.
4. IaaS – infrastructure rental.

The most common model is software as a service (SaaS), which provides access to software products hosted on remote servers. The efficiency of cloud infrastructure is assessed based on the technical reliability of the system. The distribution of data between infrastructure devices not only balances the load but also significantly reduces the probability of information loss in the event of technical failures or system failures. To ensure fault tolerance, "snapshots" or "checkpoints" are used. These are special mechanisms for saving the current state of data at a certain point in time, which restores the system to this state after failures. Snapshots are created based on established technical reliability criteria, such as the average interval between failures and the average recovery time after a failure. These methods can be used to quickly restore the infrastructure and reduce downtime, as after a failure, the system can be restored to the nearest checkpoint without significant data loss. The application of such approaches is an important element in ensuring the high availability and reliability of cloud systems, guaranteeing their continuity even in the event of technical problems.

Creation of a cloud infrastructure architecture model

The virtual data centre infrastructure simulation model describes a multi-channel queuing system for optimising data processing in cloud computing (Fig. 1). It includes components such as client request sources, queues, schedulers, applications, compute nodes, and storage. The model maximises system performance while maintaining security and reliability. The basic principles of modelling are based on equations for traffic distribution, channel capacity, and queue constraints. Optimisation of the placement of applications can improve the efficiency of request processing and reduce the load on the infrastructure.

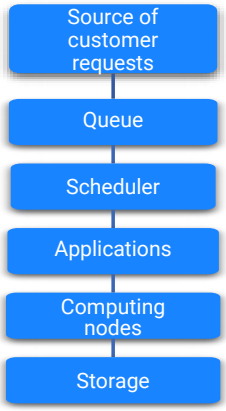

Figure 1. Simulation model of the virtual data centre infrastructure
Source: compiled by the authors

The proposed method is based on the architectural features of the cloud infrastructure. Data collection was conducted in an agentless manner, in which devices that provided computing resources to instances transmitted information about the allocated resources directly to virtual machines and containers. The architectural specificity of cloud infrastructures was the clustering of computing nodes, which distributed resources of one instance among several physical hypervisors. The agentless approach meant that there were no software agents installed on virtual machines or containers deployed within the cloud infrastructure.

The method description was based on the definition of the research object, which was the cloud infrastructure. Virtual machines and containers deployed within this infrastructure were considered as instances, and the numerical indicators of the utilisation of their main computing resources served as metrics. The metrics of computing resource utilisation included the utilisation of the central processor, RAM, storage, and the network interface of the virtual machine or container.

The method of collecting and processing metrics of cloud infrastructure instances was based on a survey of services with subsequent data on the resource utilisation of individual virtual machines and containers operating in the cloud environment. The obtained metrics were transferred to the big data processing system for further primary processing. The proposed method consisted of two main stages:

- the preparatory stage, which involved setting up the infrastructure and defining key data collection parameters;
- the data collection stage, which included the formation of a model of the normal behaviour of cloud infrastructure instances and ongoing monitoring of their performance.

At the first stage of the method, information about the cloud infrastructure was collected, including the number of modules that provided resources to the instances, the number and characteristics of the instances deployed in the system, and information about the internal network capacity of the infrastructure (Table 2). Determination of the network bandwidth and average traffic volume was necessary to calculate the coefficients of the infrastructure modules' polling functions to minimise the impact on the cloud infrastructure as a whole. The key parameters for metric collection were also defined: target resources for monitoring (CPU, memory, disk, network), polling frequency, and so on.

Table 2. System metrics for the model of the normal behaviour of instances

Timestamp	Instance_ID	CPU_Usage (%)	Memory_Usage (MB)	Disk_Usage (GB)	Network_In (KB/s)	Network_Out (KB/s)
10/03/2024 00:00:00	inst-001	25	2,048	50	100	50
10/03/2024 00:05:00	inst-001	30	2,048	52	110	55
10/03/2024 00:10:00	inst-001	28	2,048	51	105	52
10/03/2024 00:00:00	inst-002	22	4,096	100	150	75
10/03/2024 00:05:00	inst-002	26	4,096	102	160	80
10/03/2024 00:10:00	inst-002	23	4,096	101	155	78
10/03/2024 00:00:00	inst-003	20	1,024	25	80	40
10/03/2024 00:05:00	inst-003	22	1,024	26	85	42
10/03/2024 00:10:00	inst-003	21	1,024	26	82	41

Source: compiled by the authors

The initialisation of the second stage of the method of collecting and processing metrics of cloud infrastructure instances began with the receipt of a notification by the cloud infrastructure module about its readiness to receive metrics. After setting the target parameters, the data collection procedure was initiated. At this stage, the cloud

infrastructure modules were polled about the level of consumption of computing resources by each instance, which formed a set of values of M of the following form (2):

$$M(nod1, value1_{cpu}, value1_{storage}, value1_{ram}, value1_{network}). \quad (2)$$

The collected metric data are presented in Table 3. Based on these data, a reference model of instance behaviour was constructed to enable the detection of anomalies. This allowed for sustained performance analysis of virtual machines and containers in real time.

Table 3. Metrics for detecting anomalies (for standard monitoring)

Timestamp	Instance_ID	CPU_Usage (%)	Memory_Usage (MB)	Disk_Usage (GB)	Network_In (KB/s)	Network_Out (KB/s)
10/03/2024 01:00:00	inst-001	85	2,048	55	500	250
10/03/2024 01:05:00	inst-001	90	2,048	58	520	260
10/03/2024 01:00:00	inst-002	95	4,096	110	700	350
10/03/2024 01:05:00	inst-002	99	4,096	112	750	375
10/03/2024 01:10:00	inst-002	100	4,096	120	800	400
10/03/2024 01:00:00	inst-003	15	1,024	20	30	15

Source: compiled by the authors

The obtained data had to be converted into a time series format for further processing. The process of processing the MM metrics that characterised the cloud infrastructure and its instances was based on the transformation of a set of values into time series corresponding to individual indicators of the use of computing resources, including the central processor, RAM, storage system, and network interface (3):

$$M(nod1, value1_{cpu}, value1_{storage}, value1_{ram}, value1_{network}, nod2,...),$$
$$M \rightarrow (M1_{cpu}, M2_{cpu},...),$$
$$M_{cpu} \rightarrow M1_{cpu}\,(time1, value1). \quad (3)$$

After that, it was necessary to convert the data to a time series format (4):

$$M1_{cpu}\,(time1, value1) \rightarrow R_{cpu} = (R_t : t \in T). \quad (4)$$

The method incorporated the peculiarities of the operation of individual system modules, and batch processing of information by the Apache Spark big data system by the concept of Resilient Distributed Dataset, which involved the creation of an unchanged distributed collection of – Dataset objects. The computing servers used Dataset to convert the resource utilisation metrics of cloud infrastructure instances into a time series format (5):

$$R_{cpu} = (R_t : t \in T) \rightarrow Dataset_{cpu}(R_t : t), \quad (5)$$

where R_t – value of the metric at time t; T – set of time points; $Dataset_{cpu}(R_t:t)$ – representation of the obtained time series in the form of Dataset for further processing in Apache Spark.

The obtained data in time series format were stored in the big data processing system as immutable units of information following the concept of Resilient Distributed Dataset. The proposed method ensured the efficient collection and processing of metrics of cloud infrastructure instances, which provided detailed information on resource usage in real-time. The use of the time series approach and data processing within Apache Spark contributed to the improvement of analysis performance and system scalability (Table 4).

Table 4. Metrics for anomaly analysis with Apache Spark

Timestamp	Instance_ID	CPU_Usage (%)	Memory_Usage (MB)	Disk_Usage (GB)	Network_In (KB/s)	Network_Out (KB/s)	Anomaly_Score
10/03/2024 00:00:00	inst-001	25	2,048	50	100	50	0.2
10/03/2024 00:05:00	inst-001	30	2,048	52	110	55	0.3
10/03/2024 00:10:00	inst-001	28	2,048	51	105	52	0.25
10/03/2024 01:00:00	inst-001	85	2,048	55	500	250	0.95
10/03/2024 01:05:00	inst-001	90	2,048	58	520	260	0.97

Table 4. Continued

Timestamp	Instance_ID	CPU_Usage (%)	Memory_Usage (MB)	Disk_Usage (GB)	Network_In (KB/s)	Network_Out (KB/s)	Anomaly_Score
10/03/2024 01:00:00	inst-002	95	4,096	110	700	350	0.9
10/03/2024 01:05:00	inst-002	99	4,096	112	750	375	0.92
10/03/2024 01:10:00	inst-002	100	4,096	120	800	400	0.98

Source: compiled by the authors

Thus, the proposed method involved transforming the collected metrics into time series, enabling the analysis of resource usage dynamics by cloud infrastructure instances. Data processing was performed using Apache Spark based on the Resilient Distributed Dataset concept, ensuring high scalability and analysis efficiency. This approach established a reliable foundation for building load forecasting models and optimising resource allocation within the cloud environment.

Validation of the developed method for creating a cloud infrastructure architecture model

Based on the description of the method, a typical architecture was designed for the subsequent creation of an experimental testbed, which verified the effectiveness of the proposed approach. The configuration of the cloud infrastructure was presented in a basic version, where only the servers necessary for the correct operation of virtual machines and containers were used. The metrics collection module interacted with all the main components of the cloud infrastructure, which ensured efficient data collection to identify abnormal behavioural patterns of instances, as well as to monitor abnormal resource consumption throughout the cloud infrastructure. The basic set of cloud infrastructure modules included:

1. Nova is a cloud computing management component that forms an abstraction layer for virtualising computing resources of massive servers and supports the corresponding functions to improve performance.

2. Glance is a component that provides support for instance images, including system disks, which were intended to be used when launching virtual machines.

3. Cinder is a component that manages the block data stores used by Nova instances.

4. Neutron is a component that provided local network management, supporting virtual networks, DHCP and IPv6 protocol.

5. Keystone is a module that manages the list of clients and the list of services that clients can access. The module was designed to support a centralised authentication mechanism for all OpenStack components.

6. Gnocchi is a Telemetry project module, a data collection module used to provide customer billing, resource tracking, and failover capabilities across all major OpenStack components.

The block diagram of the interaction of the basic modules of the OpenStack cloud infrastructure for the implementation of the proposed method of collecting and processing metrics is shown in Figure 2.

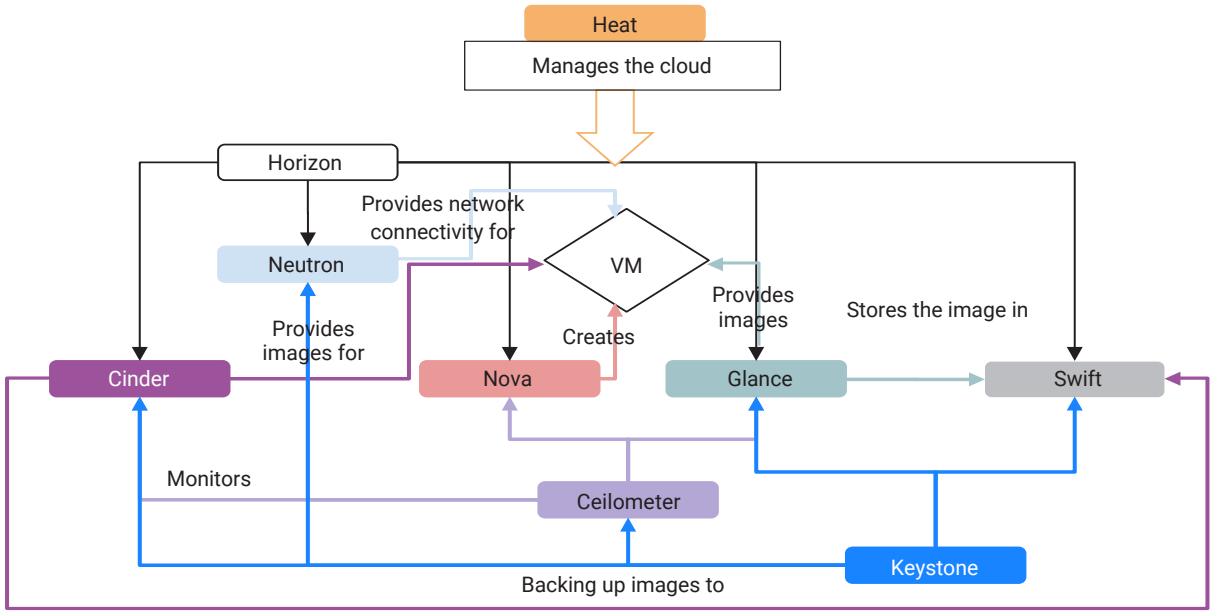

Figure 2. Structure diagram

Source: compiled by the authors

To evaluate the effectiveness of the method of collecting and processing metrics of cloud infrastructure instances, a test basic architecture was developed. The architectural solution in the proposed method of interaction between the OpenStack cloud infrastructure and the Apache Spark big data processing system was conducted following one of two tasks: the initial construction of a model of normal behaviour of the cloud infrastructure and individual instances; and standard system monitoring. The initial construction of a model of the normal behaviour of the cloud infrastructure meant collecting and storing metrics of the workload of instances under the condition of their normal operation, excluding all external influences on virtual machines or containers. Standard monitoring was defined as a collection of data on the workload of virtual machines and containers running in the cloud infrastructure, which was necessary to detect external influences. When considering the task of building a model of the normal behaviour of the cloud infrastructure, including instances, a special model tag was set in the headers of the relevant files, which identified this information as indicators of the metrics used to build the model. This information was stored in a separate repository for further processing.

To deploy the OpenStack cloud infrastructure, six separate computers were used, each of which deployed one of the infrastructure components. Ubuntu Server was chosen as the operating system, as the OpenStack developers' documentation described in detail the installation and configuration process on this particular operating system. After the operating system was installed, network connections were set up so that all components of the cloud infrastructure could interact with each other.

The first step was to configure Keystone, as this component served as the authorisation centre for the entire cloud infrastructure. This included the creation of the Keystone and the Endpoint services, which were required to ensure that the Keystone service interacted with other components. The next step was to install and configure the Glance component, which was used to store templates of virtual machines and containers that were subsequently deployed. The Glance service was installed by configuring a configuration file that specified the address or name of the server on which the Keystone service was running. Then, the Glance service was launched, completing the configuration.

Next, the Nova module was deployed, which was installed from the OpenStack repository and implemented as a Nova service. The computer on which this module was deployed provided computing resources to virtual machines and containers. During the module setup, the possibility of horizontal scaling of the cloud infrastructure was included, providing a certain pool of addresses for future expansion. After that, the Neutron module was installed and configured to provide network access for all virtual machines and containers in the cloud infrastructure. The module was implemented as a service that was configured manually.

The last stage of deploying the main components of OpenStack was the installation of the Cinder module, which ensured the interaction of cloud infrastructure instances with data storage. This module provided access to disc space for virtual machines and containers. Cinder was installed and configured in the same way as the previous components. At the final stage of preparing the cloud infrastructure for testing the proposed method, the Gnocchi module was installed. This component, such as others, was implemented as a Gnocchi service. After its installation and configuration, metrics from cloud infrastructure instances were collected by accessing this module.

To implement the second stage, seven separate computers were used, divided into two functional groups: ApacheSpark and Knowledgebase. The ApacheSpark functional group included the computers on which the big data processing system was deployed. It consisted of the Controlnode control node, as well as compute nodes and storage nodes. Since this system was intended to be used as an experiment to confirm the effectiveness of the proposed method of collecting and primary processing of cloud infrastructure instance metrics, each computer and storage node performed mixed functions: compute nodes also functioned as information storage and storage nodes processed data.

Ubuntu Server was selected as the operating system for deployment of the big data processing system due to the availability of detailed documentation. The first component to be deployed was the Controlnode, as this module was responsible for distributing computing tasks among all nodes in the system, as well as for authorising nodes. The installation of the ApacheSpark module included the installation of several management services that ensured the operation of the big data processing system. Client services were installed on the compute and storage nodes of the system. Because the compute and storage nodes were combined, the configuration files of the compute nodes specified the network addresses of the data warehouses as localhost, and the addresses of the computing devices on the data warehouses were also specified as localhost. After completing this configuration, the big data processing system was ready to use.

The final step was to set up the network between the cloud infrastructure and the big data analysis system. For this purpose, a high-speed data transmission channel with a speed of 10 Gbps was organised between RouterAS and RouterOS routers. Two virtual machines and two containers were deployed to test the performance of the testbed. The software was also launched to collect metrics from the instances. The collected metrics in a time series format were fed into the big data processing system, which confirmed the correct operation of the testbed and its readiness for experimental testing.

The proposed method significantly improved all the main metrics of cloud infrastructure monitoring. It demonstrated a significant reduction in metrics processing time, increased anomaly detection accuracy, increased metrics

collection speed, and a decrease in the number of false positives and missed anomalies (Table 5). This indicates the high efficiency of the proposed approach compared to traditional monitoring methods.

Table 5. Comparison of the results of the method of collecting and processing metrics of cloud infrastructure instances

Metric	Before methodology	After methodology	Other monitoring methods
Metric processing time (s)	150	60	100
Anomaly detection accuracy (%)	70	95	80
Metrics collection speed (KB/s)	200	1,000	400
Average latency (ms)	500	200	300
Errors of false positives	20	5	12
Errors of false negatives	15	3	8

Source: compiled by the authors

Thus, the study introduced automatic resource scaling based on neural network predictions, which ensured the optimal use of computing, network and storage resources. In addition, methods for analysing resource utilisation were developed to identify inefficiencies and suggest optimisation, which improved the overall performance of the infrastructure. The proposed method demonstrated high speed compared to similar solutions, as it used the principle of organising REST interaction via the HTTPS protocol. Furthermore, it collected information about the load directly from the cloud infrastructure modules that provided these resources. In contrast to traditional solutions, this implementation provided data in a time series format, which enabled primary data processing before creating a Dataset. This, in turn, reduced the amount of stored information in the big data processing system and did not require additional iterations to convert the information into the required format. The main idea behind the proposed method was to use a high-speed data transmission channel and integrate the OpenStack cloud infrastructure with the Apache Spark big data processing system.

Recommendations for cloud system performance optimisation

To improve the automatic scaling mechanism, advanced neural network architectures should be used to improve the accuracy of load forecasting. Optimisation of model training algorithms will facilitate faster adaptation to changes in workload, and the use of a hybrid scaling approach will help balance resource usage. Concerning resource utilisation analysis, it is advisable to implement adaptive monitoring methods that enable effective detection of inefficiencies by analysing anomalies in system behaviour. The use of cluster analysis and machine learning methods will automatically identify typical load patterns. Integration of analytical modules directly into the resource management system can be used to automatically adjust configuration parameters.

To improve the performance of data exchange, it is worth addressing the use of alternative communication protocols, such as gRPC instead of REST, which will help reduce data transfer overheads. Reducing the amount of transmitted data can be achieved through compression and aggregation on the source side while optimising caching and load balancing mechanisms at the API gateway level will increase the speed of request processing. The amount of stored information can be reduced by using efficient formats such as Parquet or ORC that support data compression. Implementing deduplication and smart archiving mechanisms based on the frequency of access to information will help optimise storage usage. The use of distributed file systems with improved access will help speed up the processing of large data sets.

The integration of OpenStack and Apache Spark can be improved by optimising data transfer channels between modules using direct data access. Distributed stream processing in Spark Streaming can be used to analyse load changes more quickly, and effective in-memory caching helps reduce query processing delays. High availability and fault tolerance can be achieved through automatic load balancing between cloud infrastructure nodes. The implementation of data replication and backup mechanisms minimises the risk of information loss, and the use of Docker or Kubernetes-based containerisation ensures scalability and rapid deployment of new service instances. The application of these recommendations can increase the efficiency of cloud infrastructure resources, improve its performance and ensure stable operation in highly loaded conditions.

Discussion

The study aimed to improve approaches to collection, processing and analysis of cloud infrastructure instance metrics, which ensures prompt processing and analysis of information in a time series format. This increases the speed and accuracy of monitoring the status of virtual resources. In addition, the study was aimed at developing methods for efficient management of cloud infrastructure resources, as the use of Apache Spark modules for primary metrics processing reduced the amount of stored data and reduced the cost of information processing. This is especially important in the context of exponential growth in the amount of data generated by cloud infrastructures.

The integration of high-speed data transmission channels between the cloud infrastructure and data analytics systems has reduced delays and improved the speed of information transfer, which ensures the timely detection of problems in the systems. As a result, the stability

and high availability of services even under high loads were ensured. An important advantage of the proposed system is its scalability, which efficiently processes huge amounts of data in real-time while maintaining high accuracy and efficiency of decision-making.

This research analyses the integration of virtual machines and containers with OpenStack modules to automatically scale resources through neural networks, which optimises the use of compute, network, and storage resources. The Keystone module provides security by authenticating and authorising all infrastructure components. Using the HTTPS protocol for data exchange reduces the amount of stored information and processes data in a time series format. At the same time, M. Aslanpour *et al.* (2020) analysed the use of appropriate metrics to evaluate the performance of cloud, fog, and edge computing, as well as to analyse their effectiveness. H. Li *et al.* (2024) integrated large-scale language models to improve resource management, increase forecasting accuracy, and optimise processes in cloud environments. The authors also considered mobile computing for real-time data processing, particularly in the healthcare sector. Security was enhanced through intrusion detection systems that protect cloud infrastructures from cyberattacks. As a result, the study combined approaches to optimise resources, increase efficiency, and ensure the security and scalability of cloud computing, while modern methods such as neural networks and large language models ensured high accuracy and adaptability of the infrastructure to changing conditions.

The current study examined the problem of optimising resource utilisation in cloud environments by integrating OpenStack with Apache Spark and automatically scaling based on neural network predictions. This approach can be used for efficient allocation of computing, network, and storage resources, and reduction of the amount of stored data by collecting information in a time series format. This differs from the approaches considered in a study by S. Tang *et al.* (2020) on Spark optimisation and methods to increase the versatility of the platform, or in a study by M. Cluci *et al.* (2023) on big data processing problems solutions based on Apache Spark and Hadoop with variable data volume. S. Gumaste *et al.* (2020) and D. Varanitskyi *et al.* (2024), on the other hand, addressed security in cloud environments, proposing a system for detecting DDoS attacks. M. Islam *et al.* (2020) addressed the optimisation of task scheduling in cloud clusters, particularly in Apache Spark, to reduce resource costs. All of these studies aimed to increase the efficiency of big data processing, reduce costs, and improve security, but the approaches to solving this problem ranged from technical optimisations to the integration of neural networks to predict resource requirements.

This study addressed the integration of OpenStack with Apache Spark to optimise resource utilisation through automatic scaling based on neural networks, which has improved performance and reduced the amount of data stored. This solution is focused on efficient resource management and reducing storage overheads. Compared to other works, such as the study by S. Namasudra *et al.* (2021), which analysed data classification and distributed databases to ensure data security, this study addressed dynamic scaling and neural networks to optimise resources. This solution is also notable in the context of the study. Thus, the comparison demonstrated that all the studies use different technologies to solve the problems of storage, data processing improved access and security in cloud environments, but the approaches based on the integration of scaling and neural networks are the most focused on resource optimisation and reducing overheads.

The current study analysed the integration of OpenStack with Apache Spark for optimisation of resources using neural networks, which increases performance and reduces the amount of data stored. This approach analysed efficient real-time data management, improving both performance and scalability. Compared to the study by A. Al-Jumaili *et al.* (2023), which analysed the use of cloud and parallel computing to improve performance in the big data and energy sectors, the approach of this study emphasises the integration of a high-speed data link to improve real-time performance. The study by O. Akindote *et al.* (2023) analysed the evolution of storage technologies, in particular cloud and edge computing, to adapt to the new demands of big data, IoT, and machine learning. While the former study proposes new technological solutions for complex infrastructures, the latter applies more traditional storage strategies to improve data security and processing efficiency.

Under conditions of high loads and large volumes of data, the ability of the system to adapt to new conditions in real time can significantly reduce downtime or overloads. The proposed method not only confirms its effectiveness in practical application but also demonstrates the importance of combining advanced data processing and cloud computing technologies to ensure reliable and efficient operation of modern cloud services. In addition, this study supports current trends in the development of cloud infrastructures, in the context of adaptive and self-learning systems that can self-optimise in constant change and increasing loads.

To improve automatic scaling, advanced neural networks should be implemented to improve the accuracy of load forecasting, and a hybrid approach to scaling will balance the use of resources. Adaptive monitoring methods and cluster analysis will help to effectively detect anomalies and automatically adjust configuration parameters. To improve data exchange performance, gRPC should be used instead of REST, data compression and aggregation should be applied, and caching and load balancing should be optimised at the API gateway level.

Reduction in the amount of stored data can be achieved through efficient formats such as Parquet or ORC, as well as deduplication and archiving. Integration of OpenStack and Apache Spark with optimised data channels and the use of Spark Streaming to handle load changes will help reduce latency. Automatic load balancing, data replication,

and containerisation using Docker or Kubernetes will ensure scalability and high availability of the infrastructure, which will increase its stability and efficiency.

Conclusions

The adoption of cloud technologies contributes to a significant increase in the productivity, flexibility and scalability of business processes, as well as support for innovative research and development in various industries. This requires consideration of several challenges related to the security, reliability and efficiency of cloud infrastructures. The study aimed to create an adaptive methodology for assessing the performance of cloud computing infrastructures, which helps to optimise resource management and reduce maintenance costs. The study analysed existing methods for assessing the performance of cloud infrastructures and identified their key characteristics, including scalability, request processing delays, computing resource utilisation, network bandwidth and load tolerance.

The developed cloud infrastructure model provided flexible resource management based on automatic scaling and integration with big data processing systems. Important features of this model include the use of a high-speed data transmission channel via HTTPS, automatic scaling of resources using neural network predictions, and resource utilisation analysis to identify inefficiencies. These innovations have resulted in increased efficiency and reduced maintenance costs compared to traditional approaches.

The proposed methodology combines modern neural network technologies and optimisation methods such as OpenStack and Apache Spark, which has resulted in high forecasting accuracy and resource management efficiency in cloud infrastructures. Its implementation has significantly improved performance and reduced maintenance costs of cloud systems, which has helped to reduce infrastructure costs due to accurate forecasts and timely scaling. The flexibility of the methodology implementation ensured its integration into various cloud computing platforms, such as Amazon Web Services, and Azure. The results of the study demonstrated that the proposed methodology not only improved performance but also reduced the load on the system by reducing the amount of stored data. Thanks to integration with cloud platforms such as OpenStack and big data processing via Apache Spark, a significant increase in the efficiency of cloud infrastructure management has been achieved.

Machine learning methods can be used to create adaptive models that can predict system load and optimise resource allocation in real-time. Deep learning and reinforcement learning provide powerful tools for analysing complex data and automating optimisation processes. Future research is needed to improve load forecasting methods using more sophisticated machine learning algorithms, as well as to explore the possibility of integrating with other cloud platforms to improve scalability and security.

Acknowledgements

None.

Funding

The study received no funding.

Conflict of Interest

None.

References

[1] Akindote, O.J., Adegbite, A.O., Dawodu, S.O., Omotosho, A., & Anyanwu, A. (2023). Innovation in data storage technologies: From cloud computing to edge computing. *Computer Science & IT Research Journal*, 4(3), 273-299. doi: 10.51594/csitrj.v4i3.661.

[2] Al-Jumaili, A.H., Muniyandi, R.C., Hasan, M.K., Paw, J.K., & Singh, M.J. (2023). Big data analytics using cloud computing based frameworks for power management systems: Status, constraints, and future recommendations. *Sensors*, 23(6), article number 2952. doi: 10.3390/s23062952.

[3] Anbalagan, K. (2024). AI in cloud computing: Enhancing services and performance. *International Journal of Computer Engineering and Technology*, 15(4), 622-635. doi: 10.5281/zenodo.13353681.

[4] Apeh, A.J., Hassan, A.O., Oyewole, O.O., Fakeyede, O.G., Okeleke, P.A., & Adaramodu, O.R. (2023). GRC strategies in modern cloud infrastructures: A review of compliance challenges. *Computer Science & IT Research Journal*, 4(2), 111-125. doi: 10.51594/csitrj.v4i2.609.

[5] Aslanpour, M.S., Gill, S.S., & Toosi, A.N. (2020). Performance evaluation metrics for cloud, fog and edge computing: A review, taxonomy, benchmarks and standards for future research. *Internet of Things*, 12, article number 100273. doi: 10.1016/j.iot.2020.100273.

[6] Bagai, R. (2024). Comparative analysis of AWS model deployment services. *International Journal of Computer Trends and Technology*, 72(5), 102-110. doi: 10.14445/22312803/IJCTT-V72I5P113.

[7] Baytelman, Y., & Potsepaiev, V. (2024). Development of a content management system and its cloud deployment. *Scientific Papers of Donetsk National Technical University. Series: "Computer Engineering and Automation"*, 2(34), 14-31. doi: 10.31474/2786-9024/v2i2(34).313761.

[8] Behlitsov, S. (2024). Software migration to a cloud architecture automation using an infrastructure as code tool Terraform AWS environment. *Computer-Integrated Technologies: Education, Science, Production*, 56, 99-106. doi: 10.36910/6775-2524-0560-2024-56-12.

[9] Chinamanagonda, S. (2023). Focus on resilience engineering in cloud services. *Academia Nexus Journal*, 2(1).
[10] Cluci, M.I., Pinzaru, C., Fotache, M., Rusu, O., & Gasner, P. (2023). OpenStack in higher education and academic research: A case study on benchmarking big data processing tools. In *Proceedings of the international conference on advanced scientific computing* (pp. 1-6). Cluj-Napoca: IEEE. doi: 10.1109/ICASC58845.2023.10328025.
[11] Dorosh, M., Hrek, I., & Buhai, Yu. (2020). Development of a model of automated personnel selection system using artificial intelligence methods. *Technical Sciences and Technologies*, 20(2), 158-166.
[12] Duan, T., Chen, R., Wang, P., Zhao, J., Liu, J., Han, S., Liu, Y., & Xu, F. (2025). BSODiag: A global diagnosis framework for batch servers outage in large-scale cloud infrastructure systems. *ArXiv*. doi: 10.48550/arXiv.2502.15728.
[13] Gumaste, S., Narayan, D.G., Shinde, S., & Amit, K. (2020). Detection of DDoS attacks in OpenStack-based private cloud using Apache Spark. *Journal of Telecommunications and Information Technology*, 82(4), 62-71. doi: 10.26636/jtit.2020.146120.
[14] Ileana, M., Oproiu, M.I., & Marian, C.V. (2024). Using docker swarm to improve performance in distributed web systems. In *Proceedings of the international conference on development and application systems* (pp. 1-6). Suceava: IEEE. doi: 10.1109/DAS61944.2024.10541234.
[15] Islam, M.T., Srirama, S.N., Karunasekera, S., & Buyya, R. (2020). Cost-efficient dynamic scheduling of big data applications in Apache Spark on cloud. *Journal of Systems and Software*, 162, article number 110515. doi: 10.1016/j.jss.2019.110515.
[16] Krishnan, P., Jain, K., Aldweesh, A., Prabu, P., & Buyya, R. (2023). OpenStackDP: A scalable network security framework for SDN-based OpenStack cloud infrastructure. *Journal of Cloud Computing*, 12(1), article number 26. doi: 10.1186/s13677-023-00406-w.
[17] Krishnaveni, S., Sivamohan, S., Sridhar, S.S., & Prabakaran, S. (2021). Efficient feature selection and classification through ensemble method for network intrusion detection on cloud computing. *Cluster Computing*, 24(3), 1761-1779. doi: 10.1007/s10586-020-03222-y.
[18] Kuprienko, A., & Galchynskyi, L. (2023). Agent-based model of access rights mining in cloud environments. In *Proceedings of the 3rd international scientific and practical conference "Science and education in progress"* (pp. 482-490). Dublin: InterConf.
[19] Li, H., Wang, S.X., Shang, F., Niu, K., & Song, R. (2024). Applications of large language models in cloud computing: An empirical study using real-world data. *International Journal of Innovative Research in Computer Science & Technology*, 12(4), 59-69. doi: 10.55524/ijircst.2024.12.4.10.
[20] Li, L., Ke, X., Wang, G., & Shi, J. (2024). AI-enhanced security for large-scale Kubernetes clusters: Advanced defense and authentication for national cloud infrastructure. *Journal of Theory and Practice of Engineering Science*, 4(12), 25-38. doi: 10.5281/zenodo.14195743.
[21] Malallah, H.S., Qashi, R., Abdulrahman, L.M., Omer, M.A., & Yazdeen, A.A. (2023). Performance analysis of enterprise cloud computing: A review. *Journal of Applied Science and Technology Trends*, 4(1), 1-12. doi: 10.38094/jastt401139.
[22] Namasudra, S., Chakraborty, R., Kadry, S., Manogaran, G., & Rawal, B.S. (2021). FAST: Fast accessing scheme for data transmission in cloud computing. *Peer-to-Peer Networking and Applications*, 14, 2430-2442. doi: 10.1007/s12083-020-00959-6.
[23] Nikitina, L., Dzheniuk, N., & Borysova, L. (2024). An expert system for cloud service risk assessment. *Control, Navigation and Communication Systems. Academic Journal*, 1(75), 146-151. doi: 10.26906/SUNZ.2024.1.146.
[24] Opirskyy, I., Vasylyshyn, S., & Susukailo, V. (2021). Investigating cybercrime with honeypots in the cloud. *Ukrainian Scientific Journal of Information Security*, 27(1), 20-26. doi: 10.18372/2225-5036.26.15574.
[25] Rahman, A., Ashrafuzzaman, M., Jim, M., & Sultana, R. (2024). Cloud security posture management automating risk identification and response in cloud infrastructures. *Academic Journal on Science, Technology, Engineering & Mathematics Education*, 4(3), 151-162. doi: 10.69593/ajsteme.v4i03.103.
[26] Shaffi, S.M. (2025). Transforming healthcare with real-time big data analytics: Opportunities, challenges, and future directions. *International Journal for Multidisciplinary Research*, 7(1). doi: 10.36948/ijfmr.2025.v07i01.36459.
[27] Tang, S., He, B., Yu, C., Li, Y., & Li, K. (2020). A survey on spark ecosystem: Big data processing infrastructure, machine learning, and applications. *IEEE Transactions on Knowledge and Data Engineering*, 34(1), 71-91. doi: 10.1109/TKDE.2020.2975652.
[28] Varanitskyi, D., Rozkolodko, O., Liuta, M., Zakharova, M., & Hotunov, V. (2024). Analysis of data protection mechanisms in cloud environments. *Technologies and Engineering*, 25(1), 9-16. doi: 10.30857/2786-5371.2024.1.1.
[29] Vavilenkova, A. (2024). The threats from using cloud services in the field of cyber security. *Electronic Professional Scientific Journal "Cybersecurity: Education, Science, Technique"*, 2(26), 409-416. doi: 10.28925/2663-4023.2024.26.704.
[30] Zahvoyskyi, R.Y., & Kazymyra, I.Y. (2024). Monitoring complex computing systems using artificial intelligence tools. In *International science-practical conference "Forestry education and science: Current challenges and development prospects"*. doi: 10.36930/conf150.5.12.

Адаптивний моніторинг продуктивності у хмарних середовищах з використанням рекурентних нейронних мереж

Павло Кудринський

Аспірант
Державний університет інформаційно-комунікаційних технологій
03110, вул. Солом'янська, 7, м. Київ, Україна
https://orcid.org/0009-0008-6314-6150

Олександр Звенигородський

Кандидат технічних наук, доцент
Державний університет інформаційно-комунікаційних технологій
03110, вул. Солом'янська, 7, м. Київ, Україна
https://orcid.org/0009-0008-6235-1638

Анотація. Метою роботи була розробка адаптивної методології для аналізу продуктивності хмарних обчислювальних інфраструктур, яка дозволяє підвищити ефективність управління ресурсами та зменшити витрати на обслуговування. Дослідження зосереджене на впровадженні новітніх підходів для автоматизації процесів моніторингу і аналізу. Методологія дослідження включала інтеграцію даних із платформ моніторингу (Amazon Web Services CloudWatch, Google Cloud Monitoring, Prometheus) для збору ключових показників продуктивності. Обробка даних здійснювалася за допомогою Python-бібліотек (NumPy, pandas, scikit-learn) для виявлення аномалій і формування часових рядів. Для моделювання продуктивності застосовувалися рекурентні нейромережі та довго-короткочасних пам'ятей на базі TensorFlow і PyTorch. Реалізація безперервного навчання дозволила адаптувати моделі до змінних умов хмарних систем у реальному часі. Основні результати дослідження включають створення новаторської системи для прогнозування ключових метрик продуктивності хмарних інфраструктур з високою точністю. Це було підтверджено за допомогою метрик середньої абсолютної помилки та корінної середньоквадратичної помилки. Інтеграція даних у реальному часі була забезпечена через платформу Amazon Kinesis, а візуалізація і управління виконувались за допомогою панелей моніторингу Amazon CloudWatch і Grafana. Віртуальні машини та контейнери взаємодіяли з модулями Nova, Glance, Cinder та Neutron, а модуль Keystone забезпечував безпеку через автентифікацію та авторизацію. Автоматичне масштабування ресурсів на основі нейронних мереж оптимізувало використання обчислювальних, мережевих та сховищних ресурсів. Розроблена методологія дозволяє автоматизувати управління хмарними ресурсами, знижуючи потребу в ручному втручанні та зменшуючи витрати. Запропонований метод забезпечував високу швидкість завдяки взаємодії через REST і HTTPS, а також збирав дані у форматі тимчасового ряду для первинної обробки. Інтеграція OpenStack з Apache Spark та використання високошвидкісного каналу передачі даних підвищили ефективність роботи інфраструктури. Висновки показали, що впровадження цієї методології значно підвищує ефективність управління хмарними інфраструктурами

Ключові слова: хмарні обчислювальні системи; продуктивність ресурсів; часові ряди; безперервне навчання; оптимізація хмарної інфраструктури; прогнозування навантаження

Journal homepage: https://itce.com.ua/en

Information Technologies and Computer Engineering **Vol. 22, No. 1. 2025**

Article's History: Received: 28.12.2024 Revised: 10.03.2025 Accepted: 24.04.2025

UDC 004.4 DOI: 10.63341/vitce/1.2025.93

Module for integrating parking hubs with the parking lot occupancy forecasting system

Vadym Kopytsia

Postgraduate Student
Vinnytsia National Technical University
21021, 95 Khmelnitske Shosse Str., Vinnytsia, Ukraine
https://orcid.org/0009-0009-6246-7793

Roman Kvyetnyy

Doctor of Technical Sciences, Professor
Vinnytsia National Technical University
9521021, 95 Khmelnitske Shosse Str., Vinnytsia, Ukraine
https://orcid.org/0000-0002-9192-9258

Abstract. The growing number of vehicles in cities creates complex challenges for parking management systems that require effective tools for predicting parking congestion. The purpose of this study was to develop and implement an integrated module for predicting parking space congestion in real time. To achieve this goal, a hybrid approach to data processing was applied, combining machine learning methods with time series analysis and spatial dynamics, and integration with modern software technologies. The results of experimental testing showed an increase in the accuracy of forecasts of parking space congestion by 20-25% compared to conventional methods, which significantly contributed to the rapid response to dynamic changes in the urban environment. By automating real-time data collection, cleaning, and aggregation, information update delays have been reduced by 10-12%, providing a more up-to-date and reliable analytical framework for management decisions. However, the increased accuracy of forecasts and prompt access to updated data helped to increase the efficiency of using parking spaces by 15-20%, optimising the distribution of traffic flows, and reducing congestion. The implementation included the use of Java and Spring Boot 3 for backend logic, AWS S3 for cloud storage, PostgreSQL as the main database, and Python algorithms using NumPy, Pandas, and python-dateutil for machine learning. Statistics, trends, and forecasts were visualised using React, which allowed users to get interactive access to results and make informed decisions. In addition, the module is easily scalable, adapts to different types of infrastructure, and can be successfully integrated into existing parking management systems. The practical significance of the development is to improve the quality of urban life by reducing congestion, reducing the environmental burden and rationalising the use of urban transport infrastructure

Keywords: intelligent parking management; machine learning in forecasting; integration of parking systems; time series analysis; traffic load forecasting; spatial data in parking; optimisation of urban infrastructure

Introduction

In modern cities, the rapid growth in the number of cars and limited parking infrastructure create significant challenges for effective management of the transport environment. The lack of free parking spaces leads to an increase in the time spent searching for parking, the appearance of traffic jams, and an increase in the environmental burden. Conventional approaches to parking resource management based on static data and manual control are not flexible and operational enough for dynamic urban environments. With this in mind, there is a need to implement innovative solutions for predicting parking space congestion in real time, able to adapt to rapid changes in traffic flows, optimise the use of infrastructure resources and reduce the negative impact of transport on the environment and the quality of life of residents.

Various aspects of the problem of parking resource management have been investigated by many researchers,

Suggested Citation:
Kopytsia, V., & Kvyetnyy, R. (2025). Module for integrating parking hubs with the parking lot occupancy forecasting system. *Information Technologies and Computer Engineering*, 22(1), 93-102. doi: 10.63341/vitce/1.2025.93

*Corresponding author

which emphasises its relevance in modern conditions. The study by A. Gonzalez-Vidal *et al*. (2022) considered the use of machine learning methods for short-term forecasting of parking space congestion. The researchers stressed that the introduction of such methods can significantly improve the accuracy of forecasts by processing data in real time. Y. Huang *et al*. (2024) proposed a transformer-based model for integrating multi-source data, which allows adapting predictive systems to changes in traffic flow. The study demonstrated the high effectiveness of this approach in difficult urban environments. Special attention was paid to the integration of parking systems with server platforms in the study by U. Yahya *et al*. (2022), who examined the use of RFID technologies and cloud environments for parking space management. Such solutions provide fast data access and system scalability. The importance of implementing IoT for monitoring parking space congestion was highlighted in the study by A.A. Elsonbaty & M. Shams (2020). The researchers noted that the use of sensor networks significantly reduces the time spent on data collection and ensures their relevance. M. Schneble & G. Kauermann (2021) proposed statistical modelling of parking space occupancy based on space-time analysis, which allows considering the dynamics of traffic flow. This approach helps to improve the accuracy of forecasting in the context of intelligent transport systems.

The development of cloud technologies and their application for parking resource management is covered in the paper by K. Nakamura *et al*. (2020). The researchers reviewed the integration of LoRaWAN – a low-power long-range radio network protocol designed to connect Internet of Things (IoT) devices to large-scale data collection and analysis systems with blockchain technologies, which ensures transparency and security of data exchange between parking hubs. J. Li *et al*. (2023) developed an integrated approach aimed at predicting parking space occupancy in a mode close to real time. Due to the use of sensor data and machine learning algorithms, it was possible to achieve high accuracy of forecasts, which is especially important for dynamic urban conditions. The researchers noted that their method not only reduces delays in forecasting, but also improves the efficiency of parking management. A. Sebatli-Saglam & F. Cavdur (2023) focused on comparing statistical and machine approaches for predicting parking space availability. They used ARIMA models and neural networks, demonstrating the advantages of combining these methods for short-term forecasting. In particular, the effectiveness of the method was confirmed based on tests in various conditions of urban infrastructure. Prediction method proposed by C. Zeng *et al*. (2022) was based on considering multiple factors using GRU-LSTM models. This approach is characterised by the ability to integrate both historical data and current conditions, providing more accurate forecasts in complex urban environments. The researchers emphasised that the proposed model is superior to conventional methods in many aspects.

A review of the latest literature has shown that most modern approaches focus on improving the accuracy of forecasts and scalability of systems, but require improvement in terms of dynamic adaptation to changes in parking hub configurations. The innovative approaches proposed in this paper are aimed at overcoming these difficulties by developing an integrated module that allows automating the collection and processing of data in real time, improving the accuracy of forecasting using machine learning, and providing convenient visualisation of results for making informed decisions on parking resource management.

The purpose of this study was to develop an integrated system that will significantly improve the accuracy of predicting parking space congestion in urban environments. This is achieved through the introduction of the latest technologies for automating data collection and processing, improving machine learning algorithms for forecasting, and developing convenient tools for data visualisation and decision-making. The objective of this study was to develop a module for integrating parking hubs with a system for predicting parking space congestion, which will provide real-time data collection, processing, and analysis. Particular attention is paid to automating the process of collecting data from various parking hubs, cleaning it, and adapting it to changing configurations. The task included developing machine learning algorithms for analysing time series and spatial data, and creating an interface for visualising prediction results.

Materials and Methods

For the development of an integrated module for predicting parking space congestion, a comprehensive approach was chosen that combines the use of real data on parking infrastructure, machine learning methods, and tools for operational processing of information in real time. The first stage of the study was the development of basic requirements for the system and the identification of key indicators that will be used to evaluate performance: the accuracy of congestion forecasts, the delay in updating data, and the efficiency coefficient for using parking spaces. To collect primary data, touch devices (parking sensors, entry/exit controllers) and external information sources (records of parking payment transactions, GPS data from mobile applications for parking search) were used. To build the architecture of the module for integrating parking hubs with the system for predicting parking congestion, the principles of microservice architecture were used, which allowed creating a flexible and scalable system for managing and coordinating various aspects of the parking process.

Data transmission to the central database was carried out via the LoRaWAN network, which provided economical and reliable information exchange over long distances. The resulting data sets were systematised in the PostgreSQL environment, and their backup storage and archived data processing were performed on AWS S3. A platform based on Java and Spring Boot 3 was chosen to provide scalability and dynamically manage data requests. Modelling, preprocessing, and data cleaning were performed using Python programming languages and libraries (NumPy,

Pandas, and python-dateutil), which allowed time series generation, standardisation of input data, and comparison of different prediction scenarios. Machine learning methods (in particular, time series regression models, ensemble methods, and spatial analysis algorithms) were selected by experimentally comparing their performance with historical data from a real parking network.

An important stage was the pilot implementation and testing of the developed module in the central business district of the city of Aarhus. The selected area was characterised by high traffic density and a shortage of parking spaces, which allowed assessing the flexibility and accuracy of the forecast system in difficult conditions. Real-time forecasts were tested based on actual parking congestion data obtained from the existing urban transport management infrastructure, and the results were compared with previous approaches without dynamic forecasting. Thus, the research materials and methods included a comprehensive combination of modern data collection and processing technologies, the application of machine learning methods to the analysis of temporal and spatial indicators of parking resource congestion, and testing the system's performance and accuracy in real urban conditions.

Results and Discussion

In order to solve this problem, it was first necessary to build the system architecture. The architecture of the parking hub integration module with the parking space congestion prediction system was built using a microservice architecture to ensure efficient data collection, processing and transmission, and adaptation to dynamic changes in the environment. The system consists of several key components, each of which performs specific tasks. The goal of the architecture is to ensure the reliability, scalability, and accuracy of real-time information processing. The system component diagram is shown in Figure 1.

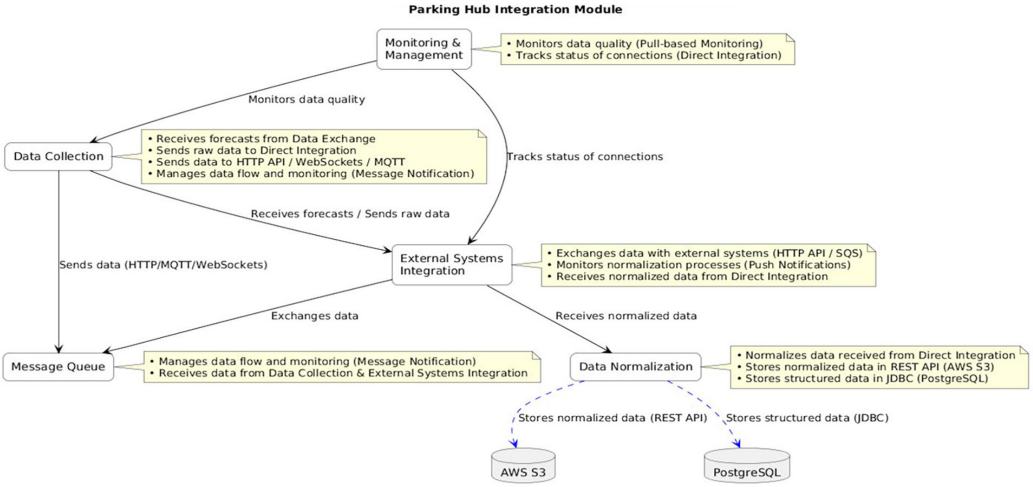

Figure 1. System component diagram
Note: diagram of components of the parking hub integration system
Source: compiled by the authors

The diagram shows the corresponding stages:

1. Connection initialisation: the component initialises connections to parking hubs and configures the communication protocol.

2. Data availability check: the component checks for input data.

- If data is available, it is extracted and parsed.
- If the data is valid, it is converted to an internal format and sent to the message queue.
- If the data is incorrect, an error in the data format is recorded.

3. Processing missing data: if data is not available, the system waits for it to arrive.

4. Connection error handling: handling connection errors and reconnection attempts.

5. Closing connections: shutting down and safely closing connections.

The first component of the system is the Data Collection component. This component is responsible for automatically collecting data from various parking hubs. The system works with several different communication protocols, such as the HTTP API, WebSockets, or gRPC, which allows connecting hubs from different manufacturers with different standards. Data comes in the form of events or requests sent by parking devices. The Java programming language and Spring Boot are used to create a service that converts this data to an internal format that will be used in subsequent processing steps. To ensure fault tolerance, the data acquisition system is integrated with message queues such as AWS SQS or Apache Kafka, which allows storing received messages and processing them asynchronously. The module regularly requests and receives information about the availability of parking spaces, transactions, technical condition of hubs, and other data. The main task of this component is to ensure uninterrupted communication with various hubs that can have different interfaces and update configurations. The activity diagram is shown in Figure 2.

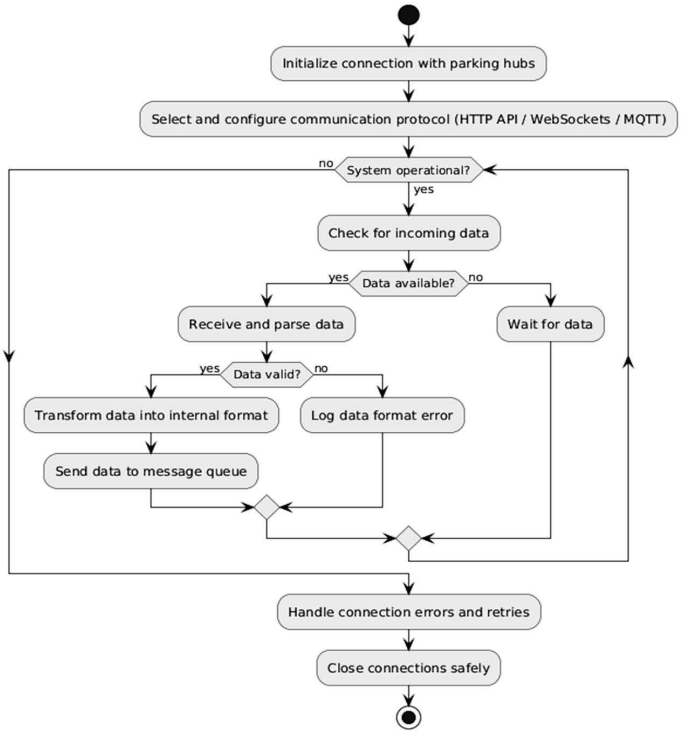

Figure 2. Flowchart of activities of the data collection component of the parking hub integration system
Source: compiled by the authors

Data transfer begins with the raw data collection component sending input data to the normalisation service, which then passes it to the validation module to verify format, completeness, and compliance with specifications. The validation module returns the validation result, and if the data turns out to be valid, the normalisation service passes it to the transformation module for conversion to a standardised format. The normalisation service then stores the converted data in the AWS S3 cloud environment and in a structured form in PostgreSQL. If the data turns out to be incorrect, the normalisation service notifies the data collection component of the error. When processing is complete, the normalisation service sends a confirmation to the data collection component that this process has been successfully completed.

The data is then transmitted to the Data Normalisation service, where it is cleaned up, validated, normalised, and standardised before being stored in an AWS S3 cloud environment or PostgreSQL database. Validation includes checking for correctness of the data format, completeness, and compliance with specifications. Since data can be received in different formats, this component is designed to convert it to a standardised format for further processing. The main task is to combine data into a single model that allows integrating information from different hubs without losing accuracy. This also includes filtering and verifying data, removing incomplete or incorrect records to avoid affecting the accuracy of forecasts. The sequence diagram of the data normalisation component is shown in Figure 3.

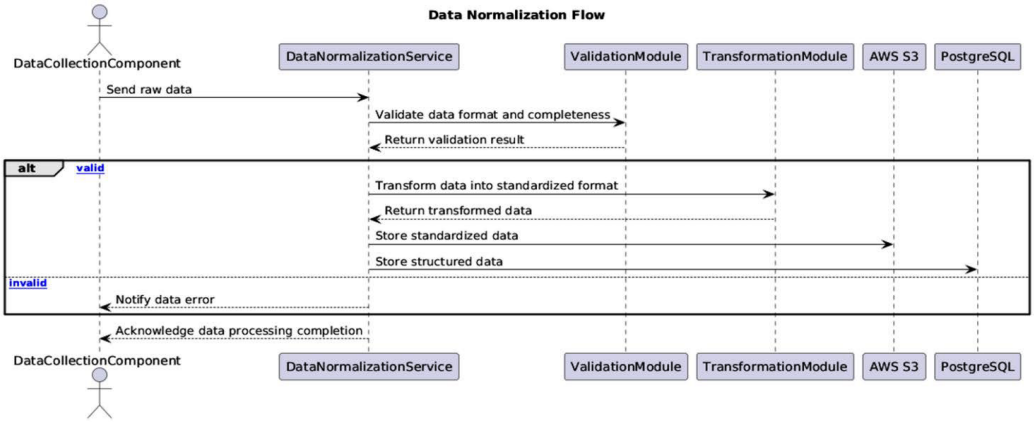

Figure 3. Sequence diagram of the data normalisation component of the parking hub integration system
Source: compiled by the authors

The data collection component first sends the raw data to the normalisation service, which then passes it to the validation module to verify the format, completeness, and compliance with specifications. After receiving the validation results, if they are correct, the normalisation service sends the data to the transformation module, which returns the data converted to a standardised format. The normalisation service then stores this standardised data in the AWS S3 cloud environment and PostgreSQL database. If the data turns out to be incorrect, the normalisation service notifies the data collection component of the error. The final action is to send the normalisation service a confirmation of successful completion of data processing back to the data collection component.

For communication with external systems, the External Systems Integration component is used. This component provides communication with municipalities or parking service operators, allowing them to use forecasts or provide their own data. The idea is to set up two-way communication to get additional data (for example, weather conditions or events in the city) that may affect the congestion of parking spaces, or to send up-to-date forecasts to operators for resource management. The sequence diagram of the integration component is shown in Figure 4.

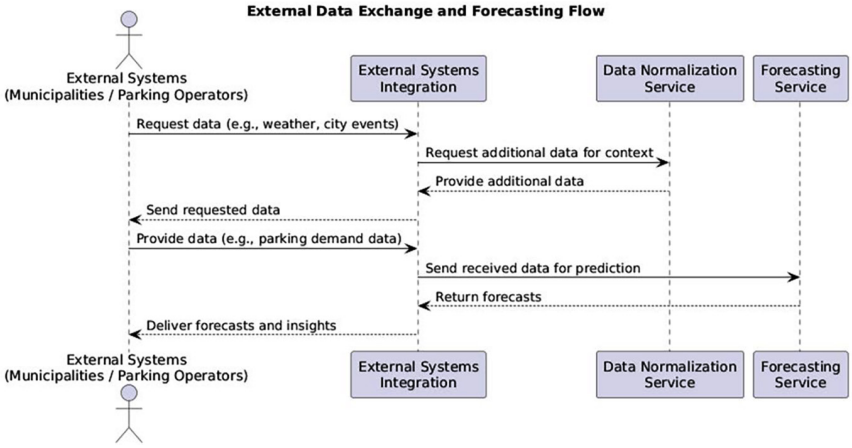

Figure 4. Sequence diagram of the parking hub integration component
Source: compiled by the authors

The diagram shows the process of data exchange between external systems, the integration component, the data normalisation service, and the forecasting service. External systems, such as municipalities or parking service operators, send data requests that are processed by the integration component. To provide context for this data, the integration component accesses the normalisation service, which provides the necessary additional information. Next, the integration component returns the finished data to external systems. External systems also provide the integration component with its own data, such as demand for parking spaces. The received data is transmitted to the forecasting service, where analysis is performed and forecasts are generated. These forecasts, together with analytical information, are returned to the integration component and delivered to external systems. The sequence of interaction between system components shown in the sequence diagram forms the basis for implementing the functionality of the monitoring and management component. The class diagram in Figure 5 details the structure of this component, which ensures the performance of its key tasks.

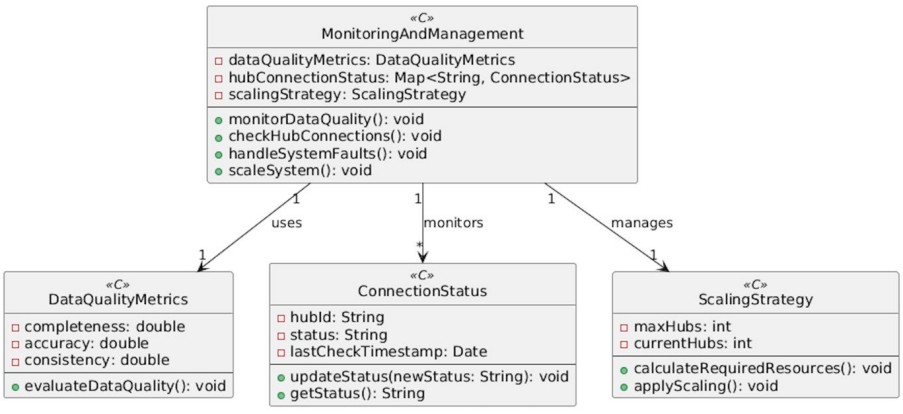

Figure 5. Class diagram of the monitoring component of the integration system with parking hubs
Source: compiled by the authors

This diagram describes the classes: MonitoringAndManagement is the main class of the monitoring and management component that monitors data quality using DataQualityMetrics, monitors connections to hubs via ConnectionStatus, manages system scaling using ScalingStrategy, and contains methods for monitoring data quality, checking connection status, managing failures, and zooming in or out of the system. DataQualityMetrics is a class that stores and evaluates data quality metrics such as completeness, accuracy, and consistency, and has methods for evaluating their level. ConnectionStatus is used to represent the connection status of each individual hub, has attributes for storing the hub ID, status, and last checked time, and methods for updating and retrieving this status. ScalingStrategy defines a system scaling strategy, contains attributes for the maximum number of hubs and the current number of hubs, and provides methods for calculating the required resources and applying scaling measures.

Integration of parking hubs with the parking space congestion prediction system has significant potential to improve both the forecasting system itself and the overall efficiency of parking space management. The substantiation for improving the efficiency of using parking spaces was obtained based on the results of pilot implementation and testing of the system in real conditions of urban infrastructure. Testing was conducted over a two-week period in the central business district of Aarhus, which is characterised by a high intensity of traffic flows and a shortage of parking spaces. The study involved the existing infrastructure of parking hubs with connected sensors that recorded the actual workload and time characteristics of parking space rotation. Comparison of the results with the period before the implementation of the system showed that the average time to search for a free parking space decreased by 12-15%, and with it, delays associated with irrational use of parking space decreased accordingly. The involved forecasting system, which promptly updated and provided data on available parking spaces, contributed to a more even distribution of cars between available places and avoided excessive congestion of cars in certain sectors. These results are consistent with previous studies, such as by H. Qu *et al*. (2022), which states that integrating data from multiple sources in real time increases the accuracy of forecasts by 20-30%, which directly affects the efficiency of using parking spaces. Figure 6 shows a comparison of forecast errors before and after the implementation of the integration module.

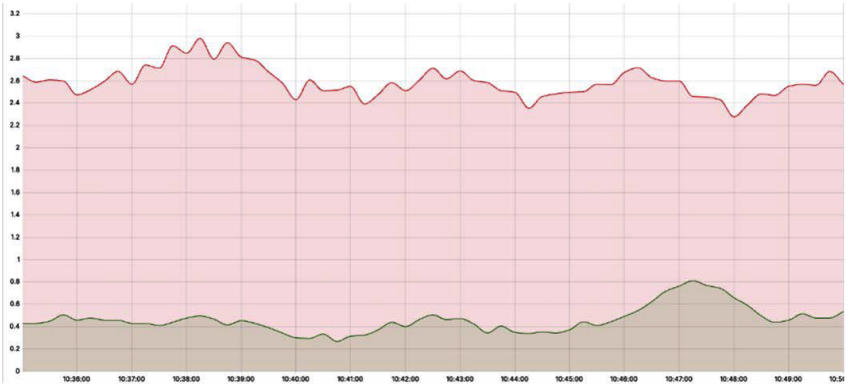

Figure 6. Graph for comparing forecast errors before and after the implementation of the integration module with parking hubs

Note: red line – forecast error before the implementation of the integration module with parking hubs; green line – forecast error after its implementation; X-axis – time scale; Y-axis – error value, where higher values indicate a larger error, and lower values indicate a more accurate forecast

Source: compiled by the authors

The results obtained, which indicate a reduction in the forecast error by 10-15% and an increase in the efficiency of using parking spaces by 15-20% after the introduction of integration with parking hubs, should be considered through the prism of previous research in this area. The current study uses an approach based on simultaneous use of data from various sources (sensor networks, mobile applications, and cloud services) to get a more complete picture of the urban transport situation. A similar multi-source strategy for collecting information was considered in the study by X. Yang *et al*. (2020), which emphasises the importance of integrating different data channels to improve the accuracy of traffic flow forecasting. The researchers noted that the use of multi-source data helps to more effectively account for dynamic changes in the urban environment. However, their approach, although it showed a significant reduction in prediction error, still had limitations in the form of less adaptive algorithms compared to those used in the current study.

The study by O. Abdulkader *et al*. (2018) proposed a smart parking management system (SPMS) that uses IoT, WSN, and RFID to improve parking efficiency and safety. The system provides real monitoring of parking availability and optimises management, reducing the search time for parking spaces and improving data accuracy. M. Barraco *et al*. (2021) analysed the forecast of parking availability based

on real data from the implemented solution. The researchers investigated how different algorithms and data sources can affect the accuracy of forecasts, and also evaluated the system's performance in real time. The results showed that the integration of sensor data and historical information can improve the accuracy of forecasting, contributing to more efficient parking management. The study by F. Terroso-Sáenz *et al*. (2016) focused on social media analysis to predict human mobility. The innovative approach included the use of algorithms for processing complex events, which helped to create accurate forecasts. The results of this study demonstrated a significant potential for analysing social data to solve problems related to traffic flows, and the tools proposed in these works significantly accelerate mobility and dynamics of predicting parking congestion. The current study also has a dynamic component – the receipt of operational data not only from stationary sensors, but also from mobile users, which allows quickly responding to changes in traffic flows, respectively increasing the relevance and reliability of forecasts.

H. Tavafoghi *et al*. (2019) proposed an approach to modelling parking dynamics based on queue theory, which involves the use of probabilistic models to predict parking space occupancy in real time. The researchers presented a model with heterogeneous vehicle arrival rates and variable service time distributions that allow considering the variability in the use of parking spaces. Using data from 29 truck parking locations over a 16-month period, the researchers confirmed the statistical assumptions of the model and demonstrated its effectiveness for forecasting. Although their approach provided probabilistic estimates of occupancy, the results show a limited increase in prediction accuracy, since the methodology does not involve integrating data from dynamic sources such as sensor networks or mobile applications. In the current study, these limitations are overcome by using adaptive models that take operational data into account, providing more accurate forecasts and the ability to scale the system in real time. For example, the study by T. Schuster & R. Volz (2019), dedicated to predicting the demand for parking spaces in German cities, used open data to simulate parking congestion. The study suggests an approach that combines historical data from open sources for demand analysis. However, the results showed that the accuracy of forecasting remained limited due to the lack of adaptive algorithms capable of accounting for dynamic changes in real time. The approach was effective for cities with stable parking infrastructure, but was not flexible enough for large metropolitan areas with high variability in traffic flows. Instead, in the current study, the integration of data from dynamic sources such as mobile applications and sensor networks allowed for significantly higher prediction accuracy and faster adaptation to changing transport situations.

In addition, R.K. Kasera & T. Acharjee (2022b) proposed the use of algorithms based on the LSTM model to predict parking space congestion. The algorithms worked with pre-structured input data, which limited their flexibility in real-world conditions. Although the researchers achieved some improvement in forecasting, the accuracy was only 8-10%, and the speed of updating data remained insufficient for dynamic urban environments. In contrast, the solution presented in the current study uses adaptive models and heterogeneous data sources, such as mobile applications and sensor networks. This allowed reducing the delay in updates by 20-30% and providing more accurate forecasting, considering fluctuations in supply and demand in real time. The study by Y. Feng *et al*. (2022) proposed the ST-GBGRU model, which combines graph convolutional networks (GCN) and gated recurrent units (GRU) to predict parking space availability based on space-time relationships. The study demonstrated the high effectiveness of the approach in short - and long-term forecasting, based on real data from public parking lots in Santa Monica. This method allows considering the dynamics of traffic flows and dependencies between different parking zones. T. Kreshchenko & Y. Yushchenko (2023) proposed a method for classifying parking space occupancy based on in-depth learning. The researchers developed a model that analyses images of parking areas obtained from surveillance cameras to determine the state of the space (free or occupied). The use of deep neural networks has made it possible to achieve high accuracy in classification even in difficult conditions, such as different lighting or variable weather. The results of the study highlight the effectiveness of in-depth learning for creating automated parking resource management systems. J. Fan *et al*. (2018) proposed an approach to predicting parking space availability based on support vector regression (SVR) optimised by the fruit fly optimisation algorithm (FOA). Their method demonstrates high accuracy and stability even in difficult conditions, in particular, for large urban parking lots. Experimental results show that FOA-SVR outperforms conventional approaches, including neural networks, by effectively accounting for multi-factor impacts on parking space occupancy. The researchers also emphasised the potential of integrating this method into smart urban transport management systems.

Thus, compared to other studies, the results show a higher efficiency of the integrated approach, which provides a more complete picture of the urban parking environment, dynamic scaling, and continuous adaptation to real-time conditions. The use of hybrid models that combine machine learning with space-time analysis has facilitated a more accurate estimation of parking space congestion, especially during peak hours.

Conclusions

In this paper, the possibilities of improving the efficiency of urban parking resource management by integrating parking hubs with the system for predicting parking space congestion were investigated. In the course of the study, an integration module was designed and implemented, which is based on automated collection, processing and analysis of large amounts of data in real time. Approaches were applied that considered changes in parking hub

configurations, ensuring stable operation and scalability of the system under high load conditions, and methods for optimising data processing and storage, which contributed to more efficient use of technical resources. The main server component is implemented in Java using Spring Boot 3, which allows real-time data processing. Data is stored in PostgreSQL, and AWS S3 is used for archiving and working with large volumes. Workload forecasting is performed by machine learning algorithms implemented in Python using the NumPy, Pandas, and python-dateutil libraries. React is used to visualise the results, which provides an interactive interface for users and easy access to analytical information. In the process of testing in real conditions, it was confirmed to reduce delays in updating data, improve the accuracy of traffic forecasting, and improve parking space efficiency indicators. The results showed that the integration of the forecasting system with parking hubs helped to significantly improve the efficiency of responding to changes in traffic flows, reduce the search time for free places, and optimise the process of managing urban parking infrastructure. The results showed an increase in the accuracy of forecasts of parking space occupancy by 20-25%, a reduction in delays in updating information by 10-12%, and an increase in the efficiency of using parking spaces by 15-20%. The implementation of the integration module allowed achieving a more comprehensive and dynamic view of the congestion of parking resources, considering unpredictable changes in the urban environment. A significant reduction in data update delays and the correction of forecasts based on the current situation were key factors that contributed to more efficient use of parking spaces, and thus reduced congestion and increased overall mobility of public transport. It was confirmed that due to the integration of various data sources and the introduction of a flexible architecture, the system can quickly scale and adapt to new workloads, ensuring smooth operation even during peak periods. The ability to quickly implement updates without stopping services had a positive impact on the overall functionality of the system, and optimising data processing and storage processes helped reduce resource costs.

The study was conducted within a limited sample and in a specific urban context, which may affect the scalability of the results in different settings. Another limitation was the lack of publicly available statistics and standards that could have improved the accuracy of forecasts and the flexibility of the system. Further research may be aimed at using an expanded range of information sources, developing algorithms for adapting to seasonal and weather factors, and integrating the system with other elements of urban infrastructure to create more integrated solutions in the field of intelligent transport management.

Acknowledgements
None.

Funding
The study received no funding.

Conflict of Interest
None.

References

[1] Abdulkader, O., Bamhdi, A. M., Thayananthan, V., Jambi, K. J., & Alrasheedi, M. (2018). A novel and secure smart parking management system (SPMS) based on integration of WSN, RFID, and IoT. In *2018 15th learning and technology conference (L&T)* (pp. 102-106). Jeddah: IEEE. doi: 10.1109/LT.2018.8368500.

[2] Barraco, M., Bicocchi, N., Mamei, M., & Zambonelli, F. (2021). Forecasting parking lots availability: Analysis from a real-world deployment. In *2021 IEEE international conference on pervasive computing and communications workshops and other affiliated events (PerCom Workshops)* (pp. 299-304). Kassel: IEEE. doi: 10.1109/PerComWorkshops51409.2021.9430942.

[3] Elsonbaty, A.A., & Shams, M. (2020). The smart parking management system. *International Journal of Computer Science & Information Technology (IJCSIT),* 12(4), 55-66. doi: 10.5121/ijcsit.2020.12405.

[4] Fan, J., Hu, Q., & Tang, Z. (2018). Predicting vacant parking space availability: A support vector regression method with fruit fly optimisation. *IET Intelligent Transport Systems*, 12(10), 1414-1420. doi: 10.1049/iet-its.2018.5031.

[5] Feng, Y., Hu, Q., & Tang, Z. (2022). Predicting vacant parking space availability zone-wisely: A graph-based spatio-temporal prediction approach. *ArXiv.* doi: 10.48550/arXiv.2205.02113.

[6] Gonzalez-Vidal, A., Terroso-Sáenz, F., & Skarmeta, A. (2022). Parking availability prediction with coarse-grained human mobility data. *Computers, Materials & Continua*, 71(3), 4355-4356. doi: 10.32604/cmc.2022.021492.

[7] Huang, Y., Dong, Y., Tang, Y., & Li, L. (2024). Leverage multi-source traffic demand data fusion with transformer model for urban parking prediction. *ArXiv*. doi: 10.48550/arXiv.2405.01055.

[8] Kasera, R.K., & Acharjee, T. (2022). Parking slot occupancy prediction using LSTM. *Innovations in Systems and Software Engineering*. doi: 10.1007/s11334-022-00481-3.

[9] Kreshchenko, T., & Yushchenko, Y. (2023). Parking spot occupancy classification using deep learning. *NRPCOMP*, 5, 72-78. doi: 10.18523/2617-3808.2022.5.72-78.

[10] Li, J., Qu, H., & You, L. (2023). An integrated approach for the near real-time parking occupancy prediction. *IEEE Transactions on Intelligent Transportation Systems,* 24(4), 3769-3778. doi: 10.1109/TITS.2022.3230199.

[11] Nakamura, K., Manzoni, P., Zennaro, M., Cano, J.-C., Calafate, C. T., & Cecilia, J.M. (2020). FUDGE: A frugal edge node for advanced IoT solutions in contexts with limited resources. In *Proceedings of the 1st workshop on experiences with the design and implementation of frugal smart objects* (pp. 30-35). New York: ACM. doi: 10.1145/3410670.3410857.

[12] Qu, H., Liu, S., Guo, Z., You, L., & Li, J. (2022). Improving parking occupancy prediction in poor data conditions through customization and learning to learn. In G. Memmi, G. Yang, B. Kong, L. Zhang, T. Qiu & M. Qiu (Eds.), *Knowledge science, engineering and management. KSEM 2022. Lecture notes in computer science* (Vol. 13368, pp. 175-189). Cham: Springer. doi:10.1007/978-3-031-10983-6_13.

[13] Schneble, M., & Kauermann, G. (2021). Statistical modeling of on-street parking lot occupancy in smart cities. *ArXiv*. doi: 10.48550/arXiv.2106.06197.

[14] Schuster, T., & Volz, R. (2019). Predicting parking demand with open data. In I.O. Pappas, P. Mikalef, Y.K. Dwivedi, L. Jaccheri, J. Krogstie & M. Mäntymäki (Eds.), *Digital transformation for a sustainable society in the 21st century. I3E 2019. Lecture notes in computer science* (Vol. 11701). Cham: Springer. doi: 10.1007/978-3-030-29374-1_18.

[15] Sebatli-Saglam, A., & Cavdur, F. (2023). Parking occupancy prediction using machine learning algorithms. *Endüstri Mühendisliği,* 34(1), 86-108. doi: 10.46465/endustrimuhendisligi.1241453.

[16] Tavafoghi, H., Poolla, K., & Varaiya, P. (2019). A queuing approach to parking: Modeling, verification, and prediction. *ArXiv*. doi: 10.48550/arXiv.1908.11479.

[17] Terroso-Sáenz, F., Cuenca-Jara, J., González-Vidal, A., & Skarmeta, A. F. (2016). Human mobility prediction based on social media with complex event processing. *International Journal of Distributed Sensor Networks*, 12(9). doi: 10.1177/1550147716668060.

[18] Yahya, U., Noah, N., Hanifah, A., Faham, L., Kasule, A., & Mubarak, H. R. (2022). RFID-cloud integration for smart management of public car parking spaces. *ArXiv*. doi: 10.48550/arXiv.2212.14684.

[19] Yang, X., Yuan, Y., & Liu, Z. (2020). Short-term traffic speed prediction of urban road with multi-source data. *IEEE Access*, 8, 87541-87551. doi: 10.1109/ACCESS.2020.2992507.

[20] Zeng, C., Ma, C., Wang, K., & Cui, Z. (2022). Parking occupancy prediction method based on multi-factors and stacked GRU-LSTM. *IEEE Access,* 10(10), 47361-47370. doi: 10.1109/ACCESS.2022.3171330.

Модуль інтеграції паркувальних хабів з системою прогнозування завантаженості паркомісць

Вадим Копиця

Аспірант
Вінницький національний технічний університет
21021, вул. Хмельницьке шосе, 95, м. Вінниця, Україна
https://orcid.org/0009-0009-6246-7793

Роман Квєтний

Доктор технічних наук, професор
Вінницький національний технічний університет
21021, вул. Хмельницьке шосе, 95, м. Вінниця, Україна
https://orcid.org/0000-0002-9192-9258

Анотація. Зростання кількості автотранспортних засобів у містах створює складні виклики для систем управління паркуванням, які потребують ефективних інструментів прогнозування завантаженості паркомісць. Метою даного дослідження є розробка та впровадження інтегрованого модуля прогнозування завантаженості паркомісць у реальному часі. Для досягнення цієї мети застосовано гібридний підхід до оброблення даних, що поєднує методи машинного навчання з аналізом часових рядів і просторової динаміки, а також інтеграцію з сучасними програмними технологіями. Результати експериментальної апробації показали підвищення точності прогнозів завантаженості паркомісць на 20–25 % порівняно з традиційними методами, що істотно сприяло оперативному реагуванню на динамічні зміни у міському середовищі. Завдяки автоматизації збору, очищення та агрегування даних у реальному часі затримки оновлення інформації скоротилися на 10–12 %, забезпечуючи більш актуальні та надійні аналітичні основи для управлінських рішень. Водночас підвищена точність прогнозів та оперативність доступу до оновлених даних сприяли збільшенню ефективності використання паркомісць на 15–20 %, оптимізуючи розподіл транспортних потоків та зменшуючи затори. Реалізація передбачала використання Java і Spring Boot 3 для бекенд-логіки, AWS S3 для хмарного зберігання даних, PostgreSQL як основної бази даних, а також алгоритмів на Python із застосуванням NumPy, Pandas і python-dateutil для машинного навчання. Візуалізація статистики, трендів та прогнозів здійснена за допомогою React, що дало змогу користувачам отримувати інтерактивний доступ до результатів і приймати зважені рішення. Крім цього, модуль легко масштабується, адаптується до різних типів інфраструктури та може бути успішно інтегрований у наявні системи управління паркуванням. Практична цінність розробки полягає у підвищенні якості міського життя завдяки скороченню заторів, зменшенню екологічного навантаження та раціоналізації використання міської транспортної інфраструктури

Ключові слова: інтелектуальне управління паркуванням; машинне навчання в прогнозуванні; інтеграція паркувальних систем; аналіз часових рядів; прогнозування транспортного завантаження; просторові дані у паркуванні; оптимізація міської інфраструктури

Journal homepage: https://itce.com.ua/en

Information Technologies and Computer Engineering **Vol. 22, No. 1. 2025**

Article's History: Received: 08.01.2025 Revised: 19.03.2025 Accepted: 24.04.2025

UDC 004.89:519.816 DOI: 10.63341/vitce/1.2025.103

Adjustment of the analytic hierarchy process indicators using AI tools

Mykhailo Klymenko

Postraduate Student, Assistant
Uzhhorod National University
88000, 3 Narodna Sq., Uzhhorod, Ukraine
https://orcid.org/0000-0002-6938-4941

Pavlo Fedorka*

PhD, Associate Professor
Uzhhorod National University
88000, 3 Narodna Sq., Uzhhorod, Ukraine
https://orcid.org/0000-0002-9242-5588

Abstract. This study aimed to enhance the analytic hierarchy process (AHP) by integrating artificial intelligence (AI) algorithms for the automatic adjustment of its indicators, thereby improving the method's accuracy, consistency, and adaptability. A conceptual analysis of both the traditional and AI-oriented approaches was conducted. The research methodology included a systematic literature review, identification of the key limitations of the classical method, and testing of AI capabilities to improve the consistency and precision of weighting coefficients. The findings demonstrate that the integration of AI into AHP significantly reduces the subjectivity of expert evaluations, lowers the need for manual adjustment of pairwise comparison matrices, and enhances the consistency of decision-making. Specifically, optimisation algorithms automatically identify conflicting judgements and correct them without human intervention, thus reducing decision-making time. The use of clustering methods facilitates the automatic grouping of criteria and alternatives based on similar characteristics, thereby reducing the number of required pairwise comparisons. The application of machine learning-based algorithms for predicting weighting coefficients enables the AHP to adapt to dynamic changes in data, enhancing the stability and reproducibility of results. Furthermore, the incorporation of Explainable AI methods improves the transparency of the decision-making process by allowing the influence of each criterion on the final outcome to be clearly explained. The analysis also demonstrated that the application of AI in multi-criteria analysis significantly reduces the cognitive load on experts, minimises the impact of human factors, and increases the accuracy of calculations. However, despite these substantial advantages, the integration of AI into AHP requires careful model configuration, as the effectiveness of such systems depends on the quality of the input data and the explainability of the outcomes. The practical significance of these findings lies in the potential to apply the proposed approaches to optimise decision-making processes in business, public administration, and the technical sciences, thereby contributing to the improved efficiency of analytical systems

Keywords: decision support system; recommender system; information models; artificial intelligence; data analysis; information technology

Introduction

Current challenges in the field of multi-criteria analysis necessitate the refinement of the analytic hierarchy process (AHP), which remains one of the most widely used approaches to decision-making. Specifically, traditional AHP faces issues regarding the subjectivity of expert judgements, the complexity of processing large volumes of data, and insufficient flexibility in the face of changing conditions. Using artificial intelligence (AI) tools presents

Suggested Citation:
Klymenko, V., & Fedorka, D. (2025). Adjustment of the analytic hierarchy process indicators using AI tools. *Information Technologies and Computer Engineering*, 22(1), 103-114. doi: 10.63341/vitce/1.2025.103

*Corresponding author

opportunities for automating the adjustment of pairwise comparison matrices, enhancing decision consistency, and significantly improving the accuracy of multi-criteria analysis. The integration of machine learning algorithms and optimisation methods into AHP facilitates its more effective application across various domains, including business, public administration, and technical sciences, thereby expanding the capabilities of analytical systems for solving complex problems.

Despite the widespread popularity of AHP, several unresolved issues persist in the scientific literature concerning the accuracy and stability of the obtained results. Research by M.M. Potomkin *et al*. (2024) demonstrated that using different variations of AHP can lead to significant discrepancies in the ranking of alternatives. This issue is particularly relevant for tasks requiring high assessment accuracy, as even minor changes in the construction of pairwise comparison matrices can substantially affect the final outcomes. Similar difficulties were corroborated by the study of O. Tymchenko *et al*. (2022), who compared various methods for prioritising factors in complex systems. The authors noted that the problem of consistency across different methods remains unresolved, and existing algorithmic approaches do not always prevent contradictions in selecting the best alternative. This indicates the necessity for further research aimed at enhancing the method's stability through the application of adaptive algorithms and mechanisms for automatic decision adjustment.

An additional aspect requiring consideration was the application of AHP in quantitative research. H. Dźwigoł (2023) highlighted that multi-criteria analysis is frequently employed in management and technical studies; however, classical methods do not always account for complex interrelationships between criteria. The author underscored that traditional AHP can yield inaccurate results when dealing with a large number of alternatives, as it is not adapted to the analysis of dynamic changes. This indicates the necessity of modernising the method through the application of technologies capable of providing adaptive adjustment of weighting coefficients in real time.

A separate direction for improving AHP involves investigating the influence of various factors on the evaluation process. Specifically, O. Andriichuk *et al*. (2024) experimentally demonstrated that the order in which pairwise comparisons are conducted can significantly impact the accuracy of weighting coefficient calculations. They emphasised that altering the sequence of evaluations can lead to result variability, which underscores the risk of subjectivity in the classical approach. Concurrently, research by T. Krenicky *et al*. (2022) analysed the conceptual foundations of AHP and confirms that the human factor remains the primary source of error in the criterion evaluation process. Thus, there is a need to develop methods that could minimise the influence of the evaluation order and enhance the reproducibility of results in multi-criteria analysis tasks.

One promising avenue for addressing these issues is the automation of evaluation processes using AI. I. Svoboda & D. Lande (2024) proposed the use of large language models to automate the assessment of pairwise comparisons, significantly reducing the influence of the human factor and enhancing the objectivity of results. They demonstrated that the application of AI helps to avoid contradictions in pairwise comparison matrices, which is a key limitation of the classical approach. Concurrently, S. Pidchenko *et al*. (2024) investigated an alternative approach to multi-criteria analysis using the fuzzy TOPSIS method. They indicated that fuzzy logic allows for reducing the impact of subjective expert evaluations, yet the mechanism for ranking alternatives itself remains dependent on the initial data. This confirms the importance of further developing automated methods for adjusting weighting coefficients, particularly through the application of AI.

Overall, while various aspects of AHP improvement are actively researched in the scientific literature, several unresolved issues still remain. Studies indicate significant variability in results depending on the evaluation method, the order of pairwise comparisons, and the subjectivity of expert decisions. Automating the process of determining weighting coefficients using AI is one of the most promising directions for enhancing AHP, yet existing approaches require further optimisation to ensure the stability and consistency of results. Therefore, this research is aimed at evaluating the impact of integrating AI algorithms on the effectiveness and accuracy of decision-making and overcoming the identified problems by developing an approach to adjusting AHP indicators using AI algorithms, which will allow for increased objectivity, accuracy, and adaptability of the method in conditions of dynamic changes.

Materials and Methods

The methodological basis of the research was built upon the principles of conceptual analysis of multicriteria decision-making methods, aimed at identifying the key distinctions between classical AHP and its modern interpretation incorporating AI algorithms. The primary focus was placed on studying the theoretical foundations of both approaches, determining their strengths and weaknesses, and evaluating the effectiveness of the proposed AI solutions for improving the decision-making process. The mechanisms of classical AHP functioning were analysed, and contemporary algorithmic optimisation methods capable of overcoming limitations related to the subjectivity of evaluations, the complexity of processing large data volumes, and insufficient decision consistency were identified.

The analysis of traditional AHP was conducted through a critical evaluation of its key structural components: constructing the hierarchical structure of criteria, completing pairwise comparison matrices, calculating weighting coefficients using the eigenvector method, and assessing the consistency of the input data. Particular attention was paid to analysing the influence of the human factor on the results of classical AHP, specifically addressing the problems of experts' cognitive load, contradictory evaluations, and the necessity for repeated adjustments of pairwise comparisons.

The effectiveness of classical AHP was compared with AI-modified approaches based on several predefined criteria, including computational accuracy, level of consistency, robustness to changes in input data, computational efficiency, and interpretability of results. These criteria were determined based on an analysis of scientific sources that outline the requirements for modern decision support systems in various fields. The following formula was used to estimate the number of comparisons required for a certain number of criteria (n):

$$\frac{n\times(n-1)}{2}. \quad (1)$$

To evaluate the impact of AI on improving the decision-making process, methods allowing for the automation of individual stages of AHP were analysed. The application of machine learning algorithms for optimising weighting coefficients and identifying contradictory evaluations in pairwise comparison matrices was investigated. The use of neural network approaches for predicting criterion weights based on historical data and clustering algorithms to group similar alternatives and reduce the volume of necessary comparisons was considered. Particular attention was paid to optimisation methods applied for correcting inconsistencies in comparison matrices. The effectiveness of using genetic algorithms, the gradient descent method, and other adaptive mechanisms to ensure the stability and consistency of adopted decisions was investigated. The possibilities of applying these methods to reduce evaluation subjectivity, enhance matrix consistency, improve the accuracy of weighting coefficients, and for the automatic identification and elimination of contradictions were considered. A conceptual comparison of classical and novel theoretical approaches allowed for determining precisely how AI algorithms can optimise the decision-making process and ensure computational accuracy in multi-criteria analysis.

Theoretical approaches to using Explainable AI to ensure greater transparency of adopted decisions were analysed. The focus was placed on theoretical models that demonstrate the potential for adaptive optimisation of the decision-making process. Evaluating the capacity of AI algorithms to generate not only accurate but also interpretable results allowed for the formation of additional methodological principles for comparing classical and AI-modified AHP models. This approach ensured the interdisciplinary validity of the research and deepened the understanding of the dynamics of cognitive and computational interaction in the context of modern decision analytics.

Results

Limitations of classical AHP and the application of AI to overcome them

AHP is widely applied for multi-criteria decision-making in various fields; however, its classical implementation is accompanied by several significant limitations that affect the accuracy, consistency, and effectiveness of the evaluation process. One of the most crucial problems is the high subjectivity of expert evaluations, which directly influences the structure and results of the analysis. In traditional AHP, experts conduct pairwise comparisons between criteria or alternatives based on their own experience and knowledge; however, this process is largely dependent on their level of competence, biases, and cognitive characteristics. Different experts may have varying perceptions of the relative importance of criteria, often leading to contradictory evaluations. This, in turn, creates difficulties in constructing a consistent matrix of pairwise comparisons, particularly if a large number of individuals are involved in the process or when complex multi-factorial decisions are being evaluated. Attempts to average such evaluations do not always yield an objective result, as even minor differences in judgements can significantly impact the final distribution of weighting coefficients.

Another critical limitation is the difficulty in correcting and checking consistency when dealing with a large number of criteria and alternatives. The classical approach involves using a consistency index to assess the reliability of pairwise comparisons; however, this indicator does not always effectively resolve identified contradictions (Wang *et al.*, 2023). The more criteria and alternatives included in the matrix, the more challenging it becomes to maintain a logical sequence of evaluations. For instance, if one group of experts rates a particular criterion as significantly important, while another considers it secondary, a problem arises that cannot be resolved without additional agreement or data adjustment. In large systems, where the number of criteria can reach dozens and alternatives hundreds, such a situation significantly complicates the analysis process. Furthermore, when discrepancies are substantial, the question remains open as to whether all evaluations should be revised, or only specific contradictory elements. The absence of a flexible adjustment mechanism forces analysts either to accept suboptimal decisions or to conduct additional evaluation iterations, which substantially complicates the process (Goepel, 2018).

A separate issue is the time cost associated with processing large pairwise comparison matrices. The AHP method becomes significantly less efficient in situations where a large number of objects or criteria need to be evaluated, as the number of necessary comparisons grows exponentially. For n criteria, the number of pairwise comparisons is determined by the formula (1). For example:

$$\frac{10\times(10-1)}{2}=45.$$

This means that for 10 criteria, 45 evaluations are needed, for 20 criteria – 190, and for 50 criteria –over 1,200. This not only creates an excessive burden on experts but also increases the risk of mechanical errors and fatigue, which further exacerbates the consistency problem (Ali *et al.*, 2023). Moreover, conducting such a volume of evaluations requires significant time resources, making the method less suitable for operational decision-making in dynamic conditions. Even with the involvement of automated tools for collecting and analysing evaluations, the process can remain

excessively lengthy, particularly if each stage requires additional consistency checking and adjustment of results.

Another significant problem is the variability of evaluations, which leads to instability in the analysis results. Different experts may evaluate the same set of criteria differently, depending on their professional experience, the context of the task, or even personal preferences. While discrepancies may be minimal within a small group of specialists, in larger systems, variability increases to such an extent that results can differ significantly depending on the composition of the expert group (Moslem, 2024). Even minor changes in the initial evaluations can lead to substantial changes in the final ranking of alternatives, which reduces the predictability and reliability of the method. Furthermore, variability complicates the reproducibility of results: in different instances of analysis, the same system of criteria can yield different outcomes, creating additional difficulties for standardising the decision-making process.

The combination of these problems limits the application of classical AHP in complex multifactorial tasks, particularly when it is necessary to obtain well-founded and stable results quickly. Therefore, the issue of automation and the use of AI to improve this method becomes particularly relevant, as it allows for minimising the impact of subjectivity, speeding up data processing, and improving the consistency of pairwise comparisons. AI opens up new possibilities for solving the main problems of the classical approach by automating key analysis stages, which include adjusting contradictory evaluations, structuring criteria and their weights, and optimising the decision-making process (Kuraś *et al*., 2024). The main AI methods that enhance the effectiveness of AHP are as follows:

1. Classification, is used to identify the most significant criteria based on the analysis of historical data and their influence on decision-making.

2. Clustering, which enables grouping criteria or alternatives by similar characteristics, simplifying the subsequent evaluation process.

3. Optimisation algorithms, are used to identify and correct contradictions in the pairwise comparison matrix, improving the consistency of evaluations.

4. Natural language processing, which allows for analysing textual information sources and automatically forming evaluation criteria.

5. Machine learning models, used for forecasting weighting coefficients and adaptively adjusting the pairwise comparison matrix in dynamic conditions.

The use of classification in AHP can significantly reduce the burden on experts and improve the accuracy of evaluation. In the traditional approach, criterion selection is performed manually, which can lead to the inclusion of secondary factors or, conversely, the omission of important analysis aspects. Classification algorithms, such as logistic regression or gradient boosting, are capable of automatically determining which criteria have the greatest impact on the decision outcome, based on historical data or previously conducted analyses (Kumar, 2025). For example, in the process of selecting an enterprise resource planning supplier, classification can help determine that cost, integration with existing infrastructure, and the level of technical support have the greatest influence on the decision, while less significant criteria such as the interface or brand popularity can be excluded from primary consideration.

Clustering, in turn, serves as an important tool for the automatic structuring of criteria and alternatives. In the traditional approach, experts manually group similar criteria, which can lead to subjective decisions that are not always consistent across different specialists. The use of algorithms such as k-means or hierarchical clustering allows for automatically identifying the relationships between criteria and assigning them to appropriate groups (Ren *et al*., 2019; Fedorov & Utkina, 2022). For instance, when evaluating car brands based on various parameters, clustering can combine criteria such as “fuel efficiency” and “environmental friendliness”, thereby reducing the number of necessary comparisons in the matrix. This not only lowers the burden on the analytical process but also enhances the consistency of evaluations, as all criteria within a single group are assessed based on similar characteristics.

Optimisation algorithms are indispensable for correcting contradictory pairwise comparisons, which constitutes one of the biggest problems in classical AHP. Since experts may provide incompatible evaluations, this necessitates reviewing a large number of pairwise comparisons manually. The use of genetic algorithms or the gradient descent method enables the automatic detection of such contradictions and the suggestion of optimal adjustments (Nazim *et al*., 2022). For example, if in one part of the matrix experts indicate that criterion A is significantly more important than B, and in another that B has a higher weight than A, the algorithm can determine an average value that best aligns with other evaluations. This allows for a significant reduction in the number of manual checks and improves the stability of the analysis results.

Natural language processing opens up new possibilities for the automatic definition of evaluation criteria, which is particularly relevant in cases where decisions are based on a large number of textual sources, such as reports, customer feedback, or expert comments. In the traditional approach, experts manually form criteria, which can lead to the omission of important analytical aspects. Natural language processing algorithms, including topic modelling and text vectorisation, enable the automatic identification of key factors that appear most frequently in texts (Dos Santos *et al*., 2023). For example, when analysing feedback on medical services, an algorithm can determine that the main criteria are “quality of service”, “waiting time”, and “doctor competence”, allowing for the formation of a criterion hierarchy based on real data rather than solely on expert assumptions.

Using machine learning models to predict weighting coefficients allows the decision-making process to become more adaptive to changing conditions. In the classical approach, criterion weights remain fixed after the initial

evaluation, which can lead to outdated or irrelevant conclusions in the long term. The use of neural networks and deep learning algorithms enables weights to be adjusted based on changes in external conditions or updates to historical data. For instance, in financial analysis, an algorithm can track currency fluctuations or changes in market trends and automatically update the significance of relevant criteria in real time.

The integration of AI algorithms into AHP significantly enhances its effectiveness. Thanks to classification, clustering, optimisation algorithms, natural language processing, and machine learning, most key processes can be automated, making the method more reliable and flexible in use. However, despite the advantages, the application of AI in AHP requires careful model tuning and adaptation to specific tasks, which remains an open area for further research.

Comparison of traditional and AI-oriented approaches in AHP

The traditional approach to the analytic hierarchy process faces several problems that complicate the process of evaluation and decision-making. In particular, expert evaluations often turn out to be contradictory, leading to the need for their correction, while an increase in the number of criteria and alternatives significantly complicates calculations. Filling the pairwise comparison matrix is a laborious process that requires significant time expenditure and increases the risk of errors. The use of AI allows for automating this process, improving the consistency of evaluations, and speeding up computations. However, despite the obvious advantages, automated methods cannot always fully replace expert analysis (Kim & Kim, 2022).

One of the key aspects compared between the traditional and AI-oriented approaches was the consistency of pairwise comparisons. Within the classical procedure, each expert independently formed the matrix, which often led to discrepancies in values caused by differing perceptions of criteria and limited cognitive resources. As a result, logical contradictions arose that required manual review and correction. In the context of the AI-oriented approach, the possibility of applying algorithms capable of automatically detecting and correcting such contradictions was implemented, reducing the level of inconsistency to permissible values without significant expert involvement. This, in turn, contributed to increased result stability and the reliability of adopted decisions.

With an increasing number of criteria, the traditional approach demonstrated a tendency towards decreasing consistency, which negatively affected the accuracy of determining weighting coefficients. The problem was further complicated by the fact that reviewing a large number of pairwise evaluations required significant time and involved multiple rounds of expert assessment. In contrast, AI approaches allowed for the application of heuristic and optimisation algorithms that automatically detected conflicts and corrected the input data. This approach not only reduced the need for repeated expert participation but also enabled working with larger systems of criteria.

Another aspect that gained significance in the context of AI application was the transparency of the decision-making procedure. The traditional approach involved full expert participation at all stages, which ensured the interpretability of the obtained weights and the final choice. In the case of AI-oriented solutions, this process was less obvious, raising doubts about the validity of the results. To partially address this problem, methods of Explainable AI began to be applied, allowing the contribution of individual criteria to the final outcome to be traced. However, such methods predominantly functioned in a post-hoc analysis mode, meaning explanations appeared only after the decision had been formed, which limited the possibility of modifying it at early stages.

It is also worth noting that the traditional approach was accompanied by a high cognitive load on experts. The necessity of performing dozens or even hundreds of pairwise comparisons in multicriteria models significantly complicated the process and increased the probability of errors. Methods based on fuzzy logic (Fuzzy AHP) offered a way to partially reduce this pressure through the application of linguistic variables and fuzzy numbers, which better corresponded to the nature of human judgements. In turn, AI-oriented approaches proposed even greater automation –by constructing models capable of evaluating alternatives –taking into account previously trained information, they significantly reduced the number of necessary expert actions.

Furthermore, the application of hybrid approaches that combine classical methods with AI models opens up new prospects for the adaptive tuning of AHP models. For example, neural networks can be used for the preliminary classification of alternatives or for detecting hidden dependencies between criteria that are not always obvious to experts. In turn, traditional methods remain important at the stage of validation and interpretation of results. Such interaction allows for achieving a better balance between the accuracy, transparency, and adaptability of decision-making models. In the future, this could contribute to the creation of interactive decision support systems capable of not only automatically calculating weights but also adapting to changes in the external environment or user preferences.

Thus, the application of AI in AHP contributes to increased accuracy of calculations, improved consistency of evaluations, and reduced labour intensity of the process. The automation of pairwise comparisons, interpretability through Explainable AI, reduced subjectivity, and accelerated computations make the AI-oriented approach an effective tool for complex tasks with a large number of criteria. Table 1 presents a generalised comparison of the key aspects of three approaches to the analytic hierarchy process: traditional, Fuzzy, and AI-oriented.

At the same time, the use of AI changes the nature of the evaluation procedure, reducing the direct role of

experts in forming weighting coefficients. This increases the efficiency of the analysis but can affect the transparency of the process, as the results of automatic adjustment are not always easily interpretable. Consequently, although the modernised approach eliminates a significant proportion of the limitations of classical AHP, its implementation requires consideration of the specifics of the particular task and the potential consequences of automation.

Table 1. Comparison of technical components of traditional and AI-oriented approaches in AHP

Parameter	Traditional AHP	Fuzzy AHP	AHP with AI
Average consistency ratio (CR)	10%-15%	8%-10%	3%-5%
Need for adjustment	High	Medium	Minimal
Interpretability of decisions	Low	Medium	High
Explanation method	Absent	Partially intuitive	SHAP, LIME
Impact of subjectivity	High	Reduced	Minimal
Cognitive load	High	Medium	Low
Decision time	Lengthy	Medium	Fastest

Source: created by the authors based on D. Lande *et al*. (2023), W. Wongvilaisakul *et al*. (2023), M.I. Merhi & A. Harfouche (2024)

Challenges, limitations, and risks of AI integration into AHP

The use of AI in AHP opens up significant opportunities for process automation, reducing labour intensity, and increasing the accuracy of calculations. However, like any tool, AI has its limitations and challenges that must be considered depending on the specific nature of the task. High requirements for input data quality, the computational complexity of algorithms, the problem of decision explainability, and potential ethical risks are the main factors that can influence the effectiveness of AI application in AHP.

One of the key issues lies in the high requirements for the quality and volume of input data, upon which the accuracy and stability of the obtained results directly depend. AI algorithms operate with large datasets, including historical data, expert evaluations, textual reports, or statistical indicators. If these data are incomplete, outdated, or contain biases, this can significantly influence the final results (Ding *et al.*, 2020). For example, if an AI model is trained on a limited data sample lacking certain categories of alternatives or evaluation criteria, it may underestimate their significance or not consider them at all. This creates a risk of distorted conclusions and incorrect ranking of alternatives. A similar problem arises in cases where the input data have a significant level of noise, i.e., contains contradictory or inaccurate information. If AI uses unstructured data, such as user feedback or analytical reports, it may identify patterns that do not always correspond to real decisionmaking criteria (Gupta *et al.*, 2022). In the classical approach, experts have the opportunity to manually check and correct the obtained results, whereas, in the case of automated processing, this may require additional control mechanisms. This is particularly relevant in tasks where the stability of priorities over time is of critical importance.

Another challenge is the issue of AI's adaptability to the unique aspects of a specific task. Machine learning algorithms work effectively with large volumes of structured data, but they may not account for contextual specificities that are obvious to human experts (Abdel-Basset *et al.*, 2024). For example, when evaluating the effectiveness of strategic planning within an organisation, there may be hidden factors, such as corporate culture or organisational constraints, which are difficult to formalise as specific criteria. While experts are capable of adapting the evaluation process, taking into account the specific conditions of the task, AI may ignore such details if they are not represented in the input data. Furthermore, automated models often operate based on standard sets of functions and do not always have the flexibility to adjust their parameters in the event of environmental changes or specific constraints (Soori *et al.*, 2024). This limitation becomes even more critical in interdisciplinary or rapidly changing conditions where adaptability is a key requirement.

Limited transparency in algorithmic decision-making is another important aspect that complicates the integration of AI into the analytic hierarchy process. Classical AHP has a clear and understandable logic, as all pairwise comparisons are directly accessible for expert checking and correction. In the case of applying optimisation algorithms, such as genetic algorithms or gradient descent methods, the adjustment of pairwise comparisons occurs automatically, and users may not have a full understanding of why certain changes were made (Mai, 2024). This can lead to a reduction in trust in the analysis results, particularly if the system makes decisions that contradict the intuitive understanding of experts. The lack of explainability in AI decisions is a serious challenge for critically important tasks where it is necessary to justify every step of the decision-making process (Araujo *et al.*, 2020). Even with the availability of explanation tools, such as SHAP or LIME, the interpretation of their results requires specialised knowledge and is not always obvious to end-users.

No less important factor to consider is the significant technical and financial resources required for implementing AI in AHP. The traditional approach can be realised without the use of complex computational systems, whereas the integration of machine learning algorithms necessitates appropriate technical infrastructure, sufficient computational power, and the availability of qualified specialists for configuring and managing the models (Zhou *et al.*, 2024). The use of AI also demands continuous updating and checking

of models, as changes in the input data can affect the accuracy of predictions. For large companies, such expenses may be justifiable, but for smaller organisations, the classical approach may remain more appropriate due to its accessibility and simplicity of implementation.

A separate challenge is the risk of excessive automation of the decision-making process, where the role of human experts is minimised. While AI is capable of significantly accelerating the analysis of large volumes of data, it cannot always account for strategic or contextual aspects of the task that require deep understanding and critical thinking (Prasetyaningrum *et al.*, 2020). If one relies entirely on the algorithmic determination of criterion weights and the consistency of pairwise comparisons, there is a threat of losing control over the decision-making process. This can be particularly dangerous in areas requiring high responsibility, such as the financial sector or public administration, where even a minor error in evaluating alternatives can have serious consequences. In such cases, human control and the ability to question the obtained results become critically important.

Also, no less important risk is the cybersecurity threat. Integrating AI into the decision-making process involves processing significant volumes of data, which can make such systems vulnerable to cyberattacks. For instance, deliberate manipulation of input data or attacks on machine learning algorithms can lead to the distortion of analysis results and the incorrect determination of weighting coefficients. Furthermore, centralised AI systems can become targets for unauthorised access, creating a risk of confidential information leakage and compromising the decision-making process.

For better visualisation of the key differences and challenges associated with using AI in AHP, Table 2 is provided below, which compares the main characteristics of the traditional and AI-oriented approaches.

Table 2. Key characteristics and limitations of traditional and AI-oriented approaches in AHP

Criterion	Traditional AHP	AHP with AI
Quality of input data	Experts adapt manually	High sensitivity to data quality
Transparency of the process	Full: all steps are logically justified	Limited explainability
Flexibility to task specifics	High: experts consider context	Low, without retraining models
Labour intensity	High: requires manual input	Low: has a configuration
Volume of processed data	Limited	Large, thanks to automation
Need for technical infrastructure	Minimal	High due to computational importance
Adaptation to new task conditions	Via expert re-evaluation	Via model retraining
Processing speed	Slow	Fast
Risk of errors	Human factor	Errors due to data distortion or incompleteness, or algorithm errors
	Local risk	Cyber threats, centralised attacks

Source: created by the authors

While the use of AI in AHP opens up new possibilities for automating and increasing the effectiveness of the evaluation process, it is also accompanied by several limitations related to the quality of input data, algorithmic transparency, the need for significant resources, and the risk of excessive automation. The most effective strategy may be to combine the classical approach with AI methods, where algorithms are used for processing large volumes of data and eliminating contradictions, while experts provide strategic oversight of the process and evaluate the unique aspects of tasks. This approach will allow for maximising the advantages of AI while simultaneously minimising its drawbacks and maintaining the quality of decisions made.

Discussion

The development of AI technologies has contributed to significant changes in multi-criteria analysis, offering new approaches to solving complex tasks. The integration of AI has enabled the automation of processes, reduced the influence of the human factor, and increased the effectiveness of calculations. Research focusing on these aspects has provided a better understanding of the potential of such technologies in improving traditional analysis methods. A review of other authors' results has allowed for analysing the advantages and challenges of implementing AI in AHP, as well as outlining the prospects of its application in various fields.

To enhance the efficiency of multi-criteria analysis, T.M. Nguyen *et al*. (2024) proposed a hybrid approach that integrated AI with Pythagorean Fuzzy AHP and COCOSO. In their study, the authors demonstrated that AI integration significantly reduced the influence of the human factor, increasing the consistency and accuracy of calculations. Particular attention was paid to tasks with fuzzy criteria, which complicated evaluation using classical approaches. The application of the proposed model proved that the automation of processes, including criterion formation and the adjustment of pairwise comparison matrices, became a key element for eliminating subjectivity. The approach confirmed that the use of automated models contributed to reducing errors and increasing accuracy in tasks that required flexible parameter tuning. This aligns with the conclusions of current study, where it was also found that automating criterion formation and ensuring model flexibility are critical for increasing effectiveness. At the same time, the specific task with fuzzy criteria,

examined in detail by T.M. Nguyen *et al.*, complements the results of this study, demonstrating that AI can handle tasks of increased complexity.

Research by M.A. Alves *et al*. (2023) focused on the application of machine learning to largescale decision-making tasks. The authors noted that the automation of calculations allows for significantly accelerating the analysis process and increasing its efficiency when working with large sets of criteria and alternatives. Particular attention was paid to the challenges associated with processing large volumes of data, which posed difficulties for classical AHP. The use of machine learning algorithms, as noted in the study, not only sped up computations but also increased accuracy by reducing the probability of errors in large datasets. This is of great importance for the tasks considered in the current research, where scalability and the ability of algorithms to work with large information systems play a special role. Furthermore, emphasis on the optimisation of large volumes of criteria confirms the prospect of automation as one of the key tools for modernising multi-criteria analysis.

Research by S. Solaimani *et al*. (2024) focused on investigating the critical success factors for implementing AI in multi-criteria analysis, with an emphasis on integrating quantitative and qualitative approaches. Their main conclusion was that process automation, including consistency checking and adaptation to dynamic changes, significantly increases the accuracy and effectiveness of decision-making. At the same time, the authors highlighted that automation without ensuring process transparency can raise doubts regarding trust in the obtained results, which, in their opinion, is one of the key challenges in using modern algorithms. They also noted that integrating different data sources is critically important for reducing the risk of incomplete information, which affects the quality of multi-criteria analysis. These conclusions largely correlate with the results of this study. Specifically, the confirmation of the effectiveness of automatic consistency checking and its ability to reduce human errors aligns with the ideas proposed in this research.

An important addition is provided by the conclusions of V.A. Salomon & L.F. Gomes (2024), who investigated the role of consistency in increasing the accuracy of multi-criteria analysis. Their research paid significant attention to improving evaluation methods, particularly through the use of new algorithms for checking pairwise comparison matrices. It is especially important that the authors not only demonstrated the advantages of automation in detecting contradictions but also highlight how these algorithms can dynamically adapt to changes in data. Furthermore, the authors noted that automation contributes to reducing the probability of human errors, especially in tasks where experts may have differing views on the significance of criteria. This aligns with the conclusions of this research regarding the reduction of subjective influence thanks to the modernised approach with AI.

Another approach to combining AHP and AI was demonstrated by A.-A. Bouramdane (2023), who considered their integration for ensuring cybersecurity. The proposed model combined classical multi-criteria analysis approaches with automated risk assessment. This allowed for quickly and accurately assessing cyber threats and making optimal decisions regarding smart grid protection. The author emphasised that the use of AI contributes to accelerating the analysis process and reducing the time spent on consistency checking, which aligns with the conclusions of this study regarding the benefits of automation. A particular emphasis was placed on how automation allows for reducing the influence of the human factor in complex situations that require processing a large volume of data. A.-A. Bouramdane also underscored the adaptability of AI, which provides flexibility in dynamic conditions, representing an important addition to the conclusions of this study. At the same time, the application of automation in cybersecurity illustrates a more specific context that complements the results of this research, which is focused on supplier selection tasks. This confirms that the flexibility and adaptability of AI can be equally effective in various fields, including risk management and strategic decision-making.

A methodological approach developed by G. Marín Díaz *et al*. (2025) involved the integration of Explainable AI with AHP to improve business decisions. Particular attention was paid to the problem of transparency, emphasising the necessity of explaining results generated by automated systems. The authors' conclusions demonstrated that insufficient clarity of algorithms can hinder user trust in such systems, even if they provide high accuracy. Furthermore, the authors noted that integrating AI with AHP allows for considering complex interrelationships between criteria, which significantly increases accuracy and adaptability in decision-making. The proposed approach illustrated that the integration of Explainable AI can help resolve a key problem mentioned by S. Solaimani *et al*. (2024), related to trust in automated systems. The problem of transparency in automated systems is an important aspect that was considered less in-depth in this study. Thus, these studies expanded the discussion of the modernised approach, adding the perspective of ensuring trust in AI within multi-criteria analysis processes.

Furthermore, the integration of AHP into recommendation systems was investigated in detail by M.A. Akbar *et al*. (2023). The authors demonstrated that the combination of AI and AHP allows for effectively handling complex interrelationships between criteria while ensuring high accuracy and adaptability of the systems. Particularly interesting is their conclusion that automation not only increases the accuracy of recommendations but also minimises the risk of subjective errors in the criterion evaluation process. This echoes the conclusions of this research that the modernised approach eliminates dependence on expert evaluations, which can be a source of contradictions and errors in the classical approach. AI's ability to work with large volumes of data allows for scaling tasks, which is a significant advantage in multi-criteria analysis.

In conclusion, within the context of the modern use of AHP, the implementation of AI has allowed for significantly enhancing traditional approaches, providing automation, increased accuracy, and a reduction in the influence of subjective evaluations. These achievements have become particularly important in complex multi-criteria tasks where the number of criteria and alternatives increases significantly. At the same time, the integration of AI creates new challenges, such as ensuring process transparency and trust in automated systems. Nevertheless, contemporary research indicates that the adaptability and scalability of AI algorithms open up new prospects for AHP application in various fields, demonstrating its significance and potential for solving current decisionmaking tasks.

Conclusions

The research demonstrated that integrating AI tools into AHP allows for a significant improvement in its effectiveness, eliminating the main limitations of the classical approach. The implementation of machine learning algorithms and optimisation methods contributes to the automation of critically important stages of the process, such as constructing the hierarchical structure, determining weighting coefficients, and checking the consistency of pairwise comparison matrices. This, in turn, reduces the level of subjectivity in expert evaluations and significantly increases the consistency of decisions made. Automating the analysis process allows for a substantial reduction in the time expenditure required for evaluation, which is particularly important when working with large sets of alternatives and criteria. The proposed optimisation methods ensure the effective adjustment of contradictory evaluations, which contributes to the stability and reproducibility of the obtained results.

Despite the significant advantages of the automated approach, the application of AI in AHP presents certain technical challenges. In particular, the effectiveness of the algorithms is significantly dependent on the quality of the input data, which can limit the accuracy of decisions made in cases where the data are incomplete or contradictory. It is important to consider that the complexity of implementing AI models and configuring them can affect the speed at putting such solutions can be put into practice. Additional resources are necessary not only for training algorithms but also for maintaining their relevance in a changing environment. Automated analysis methods may not always be able to account for specific expert knowledge, which can affect the accuracy of individual evaluations in complex cases. Furthermore, automated methods can demonstrate insufficient flexibility when working with complex dynamic systems, where real-time adaptation of weighting coefficients is required. This necessitates further research into the development of adaptive mechanisms that will allow algorithms to adjust evaluations based on new input data without losing system stability.

At the same time, the results of the study confirmed that the most effective approach is combining expert analysis with AI capabilities. The hybrid method provides an optimal balance between process automation and control by specialists, allowing for increased accuracy and consistency of results without loss of flexibility. Further improvement of this approach can be aimed at integrating more complex self-learning algorithms and developing consistency-checking methods that will expand the application of AHP in the context of a rapidly changing decision-making environment.

Acknowledgements

None.

Funding

The study received no funding.

Conflict of Interest

None.

References

[1] Abdel-Basset, M., Mohamed, R., & Chang, V. (2024). A multi-criteria decision-making framework to evaluate the impact of Industry 5.0 technologies: Case study, lessons learned, challenges and future directions. *Information Systems Frontiers*. doi: 10.1007/s10796-024-10472-3.

[2] Akbar, M.A., Khan, A.A., & Huang, Z. (2023). Multicriteria decision making taxonomy of code recommendation system challenges: A fuzzy-AHP analysis. *Information Technology and Management*, 24(2), 115-131. doi: 10.1007/s10799-021-00355-3.

[3] Ali, R., Hussain, A., Nazir, S., Khan, S., & Khan, H.U. (2023). Intelligent decision support systems – an analysis of machine learning and multicriteria decision-making methods. *Applied Sciences*, 13(22), article number 12426. doi: 10.3390/app132212426.

[4] Alves, M.A., Meneghini, I.R., Gaspar-Cunha, A., & Guimarães, F.G. (2023). Machine learning-driven approach for large scale decision making with the analytic hierarchy process. *Mathematics*, 11(3), article number 627. doi: 10.3390/math11030627

[5] Andriichuk, O., Kadenko, S., & Tsyganok, V. (2024). Significance of the order of pair-wise comparisons in Analytic Hierarchy Process: An experimental study. *Journal of Multi-Criteria Decision Analysis*, 31(3-4), article number e1830. doi: 10.1002/mcda.1830.

[6] Araujo, T., Helberger, N., Kruikemeier, S., & De Vreese, C.H. (2020). In AI we trust? Perceptions about automated decision-making by artificial intelligence. *AI & Society*, 35(3), 611-623. doi: 10.1007/s00146-019-00931-w.

[7] Bouramdane, A.-A. (2023). Cyberattacks in smart grids: Challenges and solving the multi-criteria decision-making for cybersecurity options, including ones that incorporate artificial intelligence, using an analytical hierarchy process. *Journal of Cybersecurity and Privacy*, 3(4), 662-705. doi: 10.3390/jcp3040031.
[8] Ding, R.X., Palomares, I., Wang, X., Yang, G.R., Liu, B., Dong, Y., Herrera-Viedma, E., & Herrera, F. (2020). Large-Scale decision-making: Characterization, taxonomy, challenges and future directions from an Artificial Intelligence and applications perspective. *Information Fusion*, 59, 84-102. doi: 10.1016/j.inffus.2020.01.006.
[9] Dos Santos, V.R., Fávero, L.P., Moreira, M.Â., dos Santos, M., de Oliveira, L.D., de Araújo Costa, I.P., de Oliveira Capela, G.P., & Kojima, E.H. (2023). Development of a computational tool in the Python language for the application of the AHP-Gaussian method. *Procedia Computer Science*, 221, 354-361. doi: 10.1016/j.procs.2023.07.048.
[10] Dźwigoł, H. (2023). Multi-criteria decision analysis in quantitative research. *Scientific Papers of Silesian University of Technology. Organization & Management*, 184, 96-114. doi: 10.29119/1641-3466.2023.184.6.
[11] Fedorov, E., & Utkina, T. (2022). Method of clusterization of quasiperiodic signal based on clonal selection algorithm. *Bulletin of Cherkasy State Technological University*, 27(2), 11-21. https://doi.org/10.24025/2306-4412.2.2022.253905
[12] Goepel, K.D. (2018). Implementation of an online software tool for the analytic hierarchy process (AHP-OS). *International Journal of the Analytic Hierarchy Process*, 10(3), 469-487. doi: 10.13033/ijahp.v10i3.590.
[13] Gupta, S., Modgil, S., Bhattacharyya, S., & Bose, I. (2022). Artificial intelligence for decision support systems in the field of operations research: Review and future scope of research. *Annals of Operations Research*, 308(1), 215-274. doi: 10.1007/s10479-020-03856-6.
[14] Kim, K., & Kim, B. (2022). Decision-making model for reinforcing digital transformation strategies based on artificial intelligence technology. *Information*, 13(5), article number 253. doi: 10.3390/info13050253.
[15] Krenicky, T., Hrebenyk, L., & Chernobrovchenko, V. (2022). Application of concepts of the analytic hierarchy process in decision-making. *Management Systems in Production Engineering*, 30(4), 304-310. doi: 10.2478/mspe-2022-0039.
[16] Kumar, R. (2025). A comprehensive review of MCDM methods, applications, and emerging trends. *Decision Making Advances*, 3(1), 185-199. doi: 10.31181/dma31202569.
[17] Kuraś, P., Strzałka, D., Kowal, B., Organiściak, P., Demidowski, K., & Vanivska, V. (2024). REDUCE – a tool supporting inconsistencies reduction in the decision-making process. *Applied Sciences*, 14(23), article number 11465. doi: 10.3390/app142311465.
[18] Lande, D., Strashnoy, L., & Driamov, O. (2023). Analytic hierarchy process in the field of cybersecurity using generative AI. doi: 10.2139/ssrn.4621732.
[19] Mai, W. (2024). Developing an ethical framework for artificial intelligence in investment decision-making: A fuzzy analytic hierarchy analysis. In *Proceedings of the 5th management science informatization and economic innovation development conference*. Guangzhou: MSIEID. doi: 10.4108/eai.8-12-2023.2344816.
[20] Marín Díaz, G., Gómez Medina, R., & Aijón Jiménez, J.A. (2025). A methodological framework for business decisions with explainable AI and the analytic hierarchical process. *Processes*, 13(1), article number 102. doi: 10.3390/pr13010102.
[21] Merhi, M.I., & Harfouche, A. (2024). Enablers of artificial intelligence adoption and implementation in production systems. *International Journal of Production Research*, 62(15), 5457-5471. doi: 10.1080/00207543.2023.2167014.
[22] Moslem, S. (2024). A novel parsimonious spherical fuzzy analytic hierarchy process for sustainable urban transport solutions. *Engineering Applications of Artificial Intelligence*, 128, article number 107447. doi: 10.1016/j.engappai.2023.107447.
[23] Nazim, M., Mohammad, C.W., & Sadiq, M. (2022). A comparison between fuzzy AHP and fuzzy TOPSIS methods to software requirements selection. *Alexandria Engineering Journal*, 61(12), 10851-10870. doi: 10.1016/j.aej.2022.04.005.
[24] Nguyen, T.M., Nguyen, V.P., & Nguyen, D.T. (2024). A new hybrid Pythagorean fuzzy AHP and COCOSO MCDM based approach by adopting artificial intelligence technologies. *Journal of Experimental & Theoretical Artificial Intelligence*, 36(7), 1279-1305. doi: 10.1080/0952813X.2022.2143908.
[25] Pidchenko, S., Kucheruk, O., Drach, I., & Pyvovar, O. (2024). Multi-criteria model for selection of optical linear terminals based on FUZZY TOPSIS method. *Radioelectronic and Computer Systems*, 2024(1), 65-75. doi: 10.32620/reks.2024.1.06.
[26] Potomkin, M.M., Semenenko, O.M., Kliat, Y.O., & Sedliar, A.A. (2024). Comparing the results of alternative ranking obtained by several variants of the analytic hierarchy process. *Cybernetics and Systems Analysis*, 60(6), 970-977. doi: 10.1007/s10559-024-00733-z.
[27] Prasetyaningrum, I., Fathoni, K., & Priyantoro, T.T. (2020). Application of recommendation system with AHP method and sentiment analysis. *Telecommunication Computing Electronics and Control*, 18(3), 1343-1353. doi: 10.12928/telkomnika.v18i3.14778.
[28] Ren, Z., Xu, Z., & Wang, H. (2019). The strategy selection problem on artificial intelligence with an integrated VIKOR and AHP method under probabilistic dual hesitant fuzzy information. *IEEE Access*, 7, 103979-103999. doi: 10.1109/ACCESS.2019.2931405.

[29] Salomon, V.A., & Gomes, L.F. (2024). Consistency improvement in the analytic hierarchy process. *Mathematics*, 12(6), article number 828. doi: 10.3390/math12060828.
[30] Solaimani, S., Dabestani, R., Harrison-Prentice, T., Ellis, E., Kerr, M., Choudhury, A., & Bakhshi, N. (2024). Exploration and prioritisation of critical success factors in adoption of artificial intelligence: A mixed-methods study. *International Journal of Business Information Systems*, 45(4), 429-453. doi: 10.1504/IJBIS.2024.138052.
[31] Soori, M., Jough, F.K., Dastres, R., & Arezoo, B. (2024). AI-based decision support systems in Industry 4.0, a review. *Journal of Economy and Technology*. doi: 10.1016/j.ject.2024.08.005.
[32] Svoboda, I., & Lande, D. (2024). AI agents in multi-criteria decision analysis: Automating the analytic hierarchy process with large language models. *SSRN*. doi: 10.2139/ssrn.5069656.
[33] Tymchenko, O., Khamula, O., Vasiuta, S., Sosnovska, O., & Mlynko, O. (2022). A comparison of methods for identifying the priority hierarchy of influencing factors. In *IntelITSIS – 3d international workshop on intelligent information technologies and systems of information security* (pp. 228-237). Khmelnytskyi: CEUR.
[34] Wang, K., Ying, Z., Goswami, S.S., Yin, Y., & Zhao, Y. (2023). Investigating the role of artificial intelligence technologies in the construction industry using a Delphi-ANP-TOPSIS hybrid MCDM concept under a fuzzy environment. *Sustainability*, 15(15), article number 11848. doi: 10.3390/su151511848.
[35] Wongvilaisakul, W., Netinant, P., & Rukhiran, M. (2023). Dynamic multi-criteria decision making of graduate admission recommender system: AHP and fuzzy AHP approaches. *Sustainability*, 15(12), article number 9758. doi: 10.3390/su15129758.
[36] Zhou, D., Xue, X., Lu, X., Guo, Y., Ji, P., Lv, H., Ye, W., Hu, Y., Li, Q., & Cui, L. (2024). A hierarchical model for complex adaptive system: From adaptive agent to AI society. *ACM Transactions on Autonomous and Adaptive Systems*. doi: 10.1145/3686802.

Коригування показників методу ієрархій за допомогою інструментів AI

Михайло Клименко
Аспірант, асистент
Ужгородський національний університет
88000, пл. Народна, 3, м. Ужгород, Україна
https://orcid.org/0000-0002-6938-4941

Павло Федорка
Доктор філософії, доцент
Ужгородський національний університет
88000, пл. Народна, 3, м. Ужгород, Україна
https://orcid.org/0000-0002-9242-5588

Анотація. Це дослідження спрямоване на вдосконалення методу аналізу ієрархій (MAI) шляхом інтеграції алгоритмів штучного інтелекту (ШІ) для автоматичного коригування його показників, що дозволить підвищити точність, узгодженість і адаптивність методу. У межах роботи проведено концептуальний аналіз традиційного та ШІ-орієнтованого підходів. Методологія дослідження включала систематичний аналіз літератури, виявлення основних обмежень класичного методу, а також тестування можливостей ШІ для покращення узгодженості та точності вагових коефіцієнтів. Результати дослідження показали, що впровадження ШІ у MAI значно зменшує рівень суб'єктивності експертних оцінок, знижує потребу у ручному коригуванні матриць парних порівнянь та підвищує узгодженість ухвалених рішень. Зокрема, алгоритми оптимізації автоматично ідентифікують суперечливі оцінки та коригують їх без втручання людини, що скорочує час ухвалення рішень. Використання методів кластеризації допомагає автоматично групувати критерії та альтернативи за схожими характеристиками, зменшуючи кількість необхідних парних порівнянь. Застосування алгоритмів прогнозування вагових коефіцієнтів, заснованих на машинному навчанні, дає змогу адаптувати MAI до динамічних змін у даних, підвищуючи стабільність і відтворюваність результатів. Крім того, впровадження методів Explainable AI сприяє підвищенню прозорості процесу ухвалення рішень, дозволяючи пояснювати вплив кожного критерію на кінцевий результат. Аналіз також продемонстрував, що використання ШІ в багатокритеріальному аналізі дає змогу значно зменшити когнітивне навантаження на експертів, мінімізуючи вплив людського фактора та підвищуючи точність розрахунків. Проте, попри значні переваги, інтеграція ШІ у MAI потребує ретельного налаштування моделей, оскільки їх ефективність залежить від якості вихідних даних і пояснюваності отриманих рішень. Практичне значення отриманих результатів полягає у можливості використання запропонованих підходів для оптимізації процесів ухвалення рішень у бізнесі, державному управлінні та технічних науках, що сприятиме підвищенню ефективності аналітичних систем

Ключові слова: система підтримки прийняття рішень; рекомендаційна система; інформаційні моделі; штучний інтелект; аналіз даних; інформаційна технологія

Journal homepage: https://itce.com.ua/en

Information Technologies and Computer Engineering **Vol. 22, No. 1. 2025**

Article's History: Received: 25.12.2024 Revised: 13.03.2025 Accepted: 24.04.2025

UDC 656.02: 004.021 DOI: 10.63341/vitce/1.2025.115

Use of fuzzy sets in calculating the passenger capacity utilisation rate in conditions where it is impossible to collect objective data

Ivan Zora*

Postgraduated Student
Vinnytsia National Technical University
21021, 95 Khmelnytske Shose Str., Vinnytsia, Ukraine
https://orcid.org/0009-0000-2225-777X

Oleksandr Khoshaba

PhD in Technical Science, Associate Professor
Vinnytsia National Technical University
21021, 95 Khmelnytske Shose Str., Vinnytsia, Ukraine
https://orcid.org/0000-0001-5375-6280

Abstract. The tasks of planning the organisation of passenger transportation by urban transport in modern Ukrainian conditions face new challenges, in particular, with the complexity or even impossibility of obtaining accurate input data for calculations. The research focused on solving the problem of unavailability of accurate and up-to-date data for calculating the organisation of passenger transportation by urban transport by using fuzzy logic methods. It is assumed that in conditions of limited time for conducting field research or the impact of military operations that cause dynamic changes in passenger traffic through migration processes and allow obtaining data by traditional methods, the proposed approach will allow performing calculations with minimal error. On the example of the coefficient of passenger capacity utilisation on the stage of a transport route, which directly depends on the indicator of passenger occupancy, the possibility of expanding the mathematical model of passenger transportation in urban transport using fuzzy logic approaches is considered. In particular, this refers to replacing input values with a subjective assessment of an outsider in the form of using fuzzy sets. The theoretical study showed the possibility and expediency of using fuzzy sets to solve the problem of the lack of objective input data in calculating the passenger capacity utilisation rate. The general principles of forming universes of fuzzy sets when they are used in mathematical models of the organisation of passenger transportation in urban transport to level the subjectivity of input data are determined. The requirements for the degree of overlap of the accumulated functions of belonging of fuzzy sets of the permissible level of subdivision are described, which can be used to reduce the error of calculations and, accordingly, the dimension of universes of fuzzy sets. The dependence of the tensor bit depth of the initial results on the quantitative indicator of stages on the public transport route, which can be used as a basis for analysing the complexity of calculations, is determined. The general principles of working with fuzzy sets in this mathematical model are shown using the example of calculating the passenger capacity utilisation rate. The study can be useful for city administrations, transport companies, software developers, transport logistics experts, and scientists to optimise public transport operations in the face of a lack of objective data and dynamic changes

Keywords: public transport; organisation of transport routes; passenger capacity of transport; fuzzy logic; tensor

Introduction

The issue of calculating the organisation of passenger transportation on urban transport in Ukraine has become relevant due to the high level of urbanisation and priority focus on improving the efficiency and convenience of public transport, ensuring its superiority over private, in the development and modernisation of transport infrastructure in cities. However, the realities of war have created new challenges for the scientific and transport

Suggested Citation:
Zora, I., & Khoshaba, O. (2025). Use of fuzzy sets when calculating the passenger capacity utilisation rate in conditions where it is impossible to collect objective data. *Information Technologies and Computer Engineering*, 22(1), 115-123. doi: 10.63341/vitce/1.2025.115

*Corresponding author

community. Temporary changes in transport routes, changes in the population in areas of cities and in cities in general associated with both continuous shelling of Ukrainian cities (damage to infrastructure) and changes on the line of combat collision (evacuation of the population in frontline cities, return of the population to de-occupied cities) require an immediate response and adaptation of the transport system, which is not possible when using classical mathematical models that provide for the collection of data on passenger traffic by statistical methods and methods of field research, which is complicated, limited in time, and sometimes even impossible (for example, planning when returning the administration to just de-occupied city). It is in such cases that the expansion of existing mathematical models for organising passenger transportation in urban transport using fuzzy logic methods allows calculations based on subjective data with minimal error.

The solution of such problems is quite relevant in modern scientific discourse. In particular, it may be advisable to use Bayesian methods. However, according to P.I. Bidyuk *et al.* (2021), despite all the flexibility in modelling, these methods have significant disadvantages that make it impossible to use them to solve the problem at hand. The choice of model structure and calculation methods requires highly qualified specialists, and the quality of results largely depends on the correct choice of the a priori distribution.

T.W. Richardson *et al.* (2020) propose a missing data imputation method that combines normalising flows and Monte Carlo methods. According to the above data, the use of normalising streams allows accurately estimating the data distribution density, which contributes to better imputation of missing values. But at the same time, the performance of MCFlow algorithms may depend on the choice of hyperparameters, which requires additional experiments to optimise, which is a long process.

A large number of studies note the effectiveness of using fuzzy logic methods to solve problems of the absence of objective input data. R. Saatchi (2024) provides an overview of current developments in fuzzy logic and, in particular, fuzzy inference systems (FIS). The paper discusses the prospects for further development of fuzzy logic, in particular, in industrial processes. Methods of fuzzy logic are widely used in applied fields and in situations involving the presence of various kinds of uncertainties, when these uncertainties cannot be clearly formalised using methods of probability theory and mathematical statistics. Such uncertainties can be caused by the inability to mathematically clearly determine the values of parameters and the boundaries of their belonging, or by incompleteness of needs or uncertainty in the question of the impossibility of the occurrence of certain events.

Studies of the use of fuzzy logic methods in areas related to transport and traffic are becoming increasingly widespread. For example, V. Naumov *et al*. (2021) proposes a methodology for estimating passenger preferences when choosing a bus route in a public transport system, using a mathematical apparatus of fuzzy logic based on a number of subjective characteristics such as: delivery speed, comfort, cost, etc. The originality of the study lies in the application of fuzzy logic to model the subjective preferences of passengers, which allows considering uncertainty and variability in the choice of transport route. As a further development of approaches to replacing missing or fuzzy data, their replacement with linguistic variables is used. For example, a model was developed that evaluates infrastructure projects in road and rail transport, considering the criteria for sustainable development and quality of life. Using the fuzzy logic toolbox package in MATLAB, the authors not only created a knowledge base with linguistic variables, membership functions, and inference rules, but also tested the model on real projects, which confirmed its effectiveness in supporting decision-making (Kaczorek & Jacyna, 2022).

There are a number of studies describing the use of fuzzy logic for mathematical modelling of transport processes. Thus, for example, I. Medvediev *et al.* (2024) developed a mathematical model that allows assessing and managing risks in the conditions of uncertainty typical of modern agribusiness. The study focused on risks in the logistics route of grain transportation from Ukraine to Poland and allows considering subjective and objective risk factors, ensuring adaptation to changing supply chain conditions. X. Yang *et al*. (2022) developed a model that combines various criteria such as comfort, reliability, and accessibility to provide a comprehensive assessment of the operation of the transport system, using fuzzy logic to account for subjective and objective indicators.

Most of the previous studies combine the use of fuzzy logic to level out the absence, fuzziness, or subjectivity of input data. Understanding this allows expanding the searching for literature beyond the specialised topics of transport technologies and use research, approaches and developments in the field of applying fuzzy logic in solving problems from various scientific and applied fields. For example, J. Grosset *et al*. (2024) clearly describe the extended possibilities of using fuzzy logic to solve the input uncertainty problem. In particular, the influence and possibility of mathematical modelling of the process is revealed not only in conditions of fuzzy knowledge (fuzzy sets, linguistic variables, fuzzy decision rules), but also in conditions of fuzzy behaviour (fuzzy conclusions) and fuzzy interaction or roles. The effectiveness of using linguistic variables in solving a decision-making problem under a fuzzy model is shown in a study conducted for an automatic mail sorting line (Grebennik & Kovalenko, 2024). The researchers not only identified input and output linguistic variables, but also substantiated the importance of phasification and dephasisation when using complex multi-tertiary classification systems.

The series of books "Studies in Fuzziness and Soft Computing" edited by J. Katzpshik of Springer Nature Switzerland AG became a fundamental work in the field of studying, researching and describing methods,

approaches and implementations of fuzzy logic in various fields of science and technology. It includes publications on various topics in soft computing that cover fuzzy sets, rough sets, neural networks, evolutionary computing, probabilistic and provable thinking, multi-valued logic, and related fields. One of the books in the series under the authorship of T. Bhatia *et al*. (2023) describes a set of methods for finding solutions of the "more-for-less" type for various types of fuzzy transport planning problems. The book presents new methods for solving various types of problems, including symmetric balanced fuzzy transport planning problems, symmetric intuitive fuzzy transport planning problems with mixed constraints, and symmetric intuitive fuzzy linear fractional transport planning problems with mixed constraints. The book discusses in detail their applications using examples of representative tasks and discusses possible areas for further research.

Based on the above, it becomes clear that the problem of the impossibility of obtaining objective data for calculations exists in various fields of science and technology. When solving such problems, they rely on fuzzy logic methods, which demonstrate high efficiency. However, methods of involving fuzzy logic in each individual model are highly variable and require in-depth analysis and research to identify the optimal method for improving the existing mathematical model. The purpose of this study was to evaluate the feasibility of using fuzzy sets to calculate the passenger capacity utilisation rate in conditions of impossibility of obtaining objective data, as well as to determine promising directions for further research to improve the mathematical model of passenger transportation organisation in urban transport.

Materials and Methods

The main stages of the study included the analysis of modern methods of work, such as the Bayesian method, the Monte Carlo method, and fuzzy logic methods, in conditions of inability to obtain objective input data, and the determination of the indicative component (variable) of the mathematical model for the study of passenger transportation in urban transport. Such a component is most susceptible to the influence of fuzziness, and therefore, the consideration of methods of working with it is the most revealing. This allowed investigating and presenting approaches to the use of fuzzy logic methods in a mathematical model, developing an algorithm for working with data that have limited accuracy, and providing recommendations for the optimal choice of parameters for practical implementation of the model.

Calculating the organisation of passenger transportation in urban transport is a complex complex multi-criteria mathematical model that considers a large number of different factors. By comparative analysis of the mathematical models considered by V.V. Bilichenko *et al*. (2020), R.A. Khabutdinov & I.O. Fedorenko (2021), J. Hellekes & C. Winkler (2022), describing the organisation of passenger transportation on public transport, the passenger capacity utilisation rate on the stage as a key component of most models were identified and selected. In addition, in comparison with other indicators such as the average length of the stage, the density of traffic flow, the specific number of traffic lights on the route, it can contain up to 100% uncertainty throughout the route, based on setting the task conditions, since it is based on the filling indicator. The static passenger capacity utilisation rate on the stage was calculated using the equation:

$$\gamma_{ck} = \frac{F_k}{n_{rot} \cdot q_{avg}}, \quad (1)$$

where F_k – filling on k-th stage; n_{rot} – number of rotations performed by vehicles during peak hours and per day; q_{avg} – average passenger capacity of buses operating on this route. At the same time:

$$F_k = F_{k-1} + P_k - O_k, \quad (2)$$

where F_{k-1} – filling on the previous k-th stage (on the first stage, this value is zero); P_k and O_k – accordingly, the number of passengers who got on and off the bus at the bus stop.

To investigate the possibility and feasibility of using methods of fuzzy modelling and parameter phasing, the method of parameter modelling and the method of gradual approximation with theoretical verification were used to assess the stability of the model, optimise parameters and estimate the complexity of calculations. To provide a deeper analysis, membership functions were used to describe the uncertainty and variability in the transport models. In the study, universes of possible values of key parameters were created that consider the uncertainty of input data and allow adapting the model to the real conditions of urban transport functioning. The sensitive analysis method was used to determine the quantitative indicator of the complexity of calculations of an improved mathematical model. The calculations were based on the example of the passenger capacity utilisation rate on a transport stage.

Results and Discussion

If it is impossible to obtain up-to-date accurate data of filling parameters, they can be obtained as a subjective assessment of the observer in the form of a linguistic variable, or a fuzzy set. To calculate the passenger capacity utilisation rate based on a subjective estimate of the number of passengers who entered or left the vehicle at a stop, at least 3 fuzzy sets must be defined: few ($\tilde{A}$), many ($\tilde{B}$), full/all ($\tilde{C}$). For a conditional fleet of vehicles, where q_{avg} = 18 persons the corresponding universes can acquire the following values:

$$\cup_A = \{1, 2, 3, 4, 5, 6, 7, 8\};$$
$$\cup_B = \{7, 8, 9, 10, 11, 12, 13, 14, 15, 16\};$$
$$\cup_C = \{15, 16, 17, 18\}.$$

Therefore, there will be corresponding fuzzy sets that can be represented as follows:

$$\tilde{A} = \{a, \mu_A(a) \mid a \in \cup_A\};$$
$$\tilde{B} = \{a, \mu_B(a) \mid a \in \cup_B\};$$
$$\tilde{C} = \{a, \mu_C(a) \mid a \in \cup_C\}.$$

For a better understanding of the distribution of universes of the proposed fuzzy sets, they are shown in the graph of membership functions (Fig. 1) as an example. The X-axis represents the number of passengers who entered the vehicle according to the subjective assessment of the observer (variable *a*), while the Y-axis shows the value of the membership function of the corresponding fuzzy set μ.

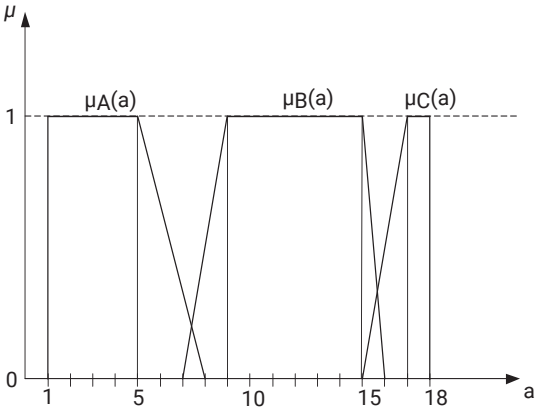

Figure 1. Graphs of functions belonging to fuzzy sets $\tilde{A}$ (little), $\tilde{B}$ (many) and $\tilde{C}$ (full)
Source: developed by the author based on the proposed as an example of fuzzy sets and their universes

In conditions of independent calculations (i.e., the absence of recursive use of previously computed data), such an approach would be optimal for solving the problem of input data fuzziness. However, when calculating the passenger capacity utilisation rate on a particular stage (equation 1), the fill indicator is used F_k (equation 2), which is directly dependent on the calculated filling rate in the previous stage (F_{k-1}).

For clarity, the value of the fullness indicator is substituted into the equation for calculating the passenger capacity utilisation rate on the stage, and the calculation problem is formulated at the initial stages of the route. For example, according to the subjective assessment of an outside observer, "few" passengers entered the bus at the starting point of the route, then the calculation of the passenger capacity utilisation rate on the 0-th stage will look like this:

$$\gamma_{c0} = \frac{P_0}{n_{rot} \cdot q_{abg}} = \frac{\{1,2,3,4,5,6,7,8\}}{n_{rot} \cdot q_{avg}}, \quad (3)$$

which does not cause any special difficulties, even with manual calculation and after calculation, it will represent a certain set of values . Let on the next stage, according to the observer's assessment – "many" went in and "few" came out. In this case, the equation statement will be as follows:

$$\gamma_{c1} = \frac{F_0 + P_1 - B_1}{n_{rot} \cdot q_{avg}} = \frac{\{1,2,3,4,5,6,7,8\} + \{7,8,9,10,11,12,13,14,15,16\} - \{1,2,3,4,5,6,7,8\}}{n_{rot} \cdot q_{avg}}. \quad (4)$$

Even at this stage, there is actually a transition to a 3-dimensional tensor of results , which directly depends on the conditional three-dimensional tensor of the fullness coefficient represented by a combination of fuzzy sets . Even at this iteration of calculations, the need to automate such calculations is clearly seen. Each subsequent iteration of calculations, each subsequent stage on a public transport route, will add a bit depth to the measurement of the fullness indicator and, as a result, the passenger capacity utilisation rate itself. On a conditional *k*-th stage, the problem of finding the passenger capacity utilisation rate takes the following form (if you entered – "few", left – "many" on the stage *k-1* and went in – "few", went out – "all" on the *k*-th stage):

$$\gamma_{ck} = \frac{\begin{Bmatrix} \{F_{k-2}\}\tilde{A}_1\tilde{B}_1 & \{F_{k-2}\}\tilde{A}_1\tilde{B}_2 & \cdots & \{F_{k-2}\}\tilde{A}_1\tilde{B}_n \\ \{F_{k-2}\}\tilde{A}_2\tilde{B}_1 & \{F_{k-2}\}\tilde{A}_2\tilde{B}_2 & \cdots & \{F_{k-2}\}\tilde{A}_2\tilde{B}_n \\ \cdots & \cdots & \ddots & \cdots \\ \{F_{k-2}\}\tilde{A}_n\tilde{B}_1 & \{F_{k-2}\}\tilde{A}_n\tilde{B}_2 & \cdots & \{F_{k-2}\}\tilde{A}_n\tilde{B}_n \end{Bmatrix}_{+\tilde{A}-\tilde{C}}}{n_{rot} \cdot q_{avg}}, \quad (5)$$

where the tensor dimension digit will be in the following dependence on the ordinal (quantitative) stage indicator:

$$n = k + (k-1) + 2 = 2k + 1. \quad (6)$$

That is, for example, the tensor will be 23-dimensional, which, in fact, leads such calculations to the field of "Big Data". However, there are several other factors that increase the complexity of the mathematical model. Returning to the consideration of graphs of functions belonging to fuzzy sets (Fig. 1), which were adopted to solve the problem of fuzzy input data to solve the problem of determining the passenger capacity utilisation rate, it is possible to clearly distinguish the coverage area of possible values of fuzzy data sets after accumulating their membership functions (Fig. 2), or rather zones of non-belonging, in which, in particular, some possible values of universes do not even reach a probability of 0.5.

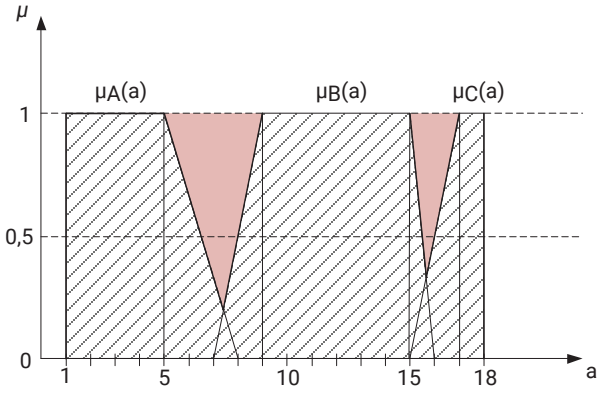

Figure 2. Result of accumulating membership functions of fuzzy sets
Note: $\tilde{A}$ (little), $\tilde{B}$ (many) and $\tilde{C}$ (full)
Source: compiled by the author based on Figure 1

Figure 2 shows that the selected fuzzy sets, or rather their universes, are not sufficient to solve the problem with a sufficient level of reliability. Since it is not appropriate to increase the coverage of universes and thereby expand the boundaries of sets, this disadvantage, which has a negative impact on the accuracy of calculations, must be solved by subdividing. This results in a larger set of fuzzy sets with a more complex function for accumulating membership functions, but with a much smaller calculation error. For example:

$$\tilde{A} = \{a, \mu_A(a) \mid a \in \{1, 2, 3, 4, 5\}\};$$
$$\tilde{B} = \{a, \mu_B(a) \mid a \in \{4, 5, 6, 7, 8\}\};$$
$$\tilde{C} = \{a, \mu_C(a) \mid a \in 7, 8, 9, 10, 11\}\};$$
$$\tilde{D} = \{a, \mu_A(a) \mid a \in \{10, 11, 12, 13, 14\}\};$$
$$\tilde{E} = \{a, \mu_B(a) \mid a \in \{13, 14, 15, 16, 17\}\};$$
$$\tilde{F} = \{a, \mu_C(a) \mid a \in \{16, 17, 18\}\}.$$

It is predicted that the proposed level of subdivision of fuzzy set universes will increase the minimum allowable probability level to 0.5. Figure 3 shows graphs of the membership functions of the proposed fuzzy sets.

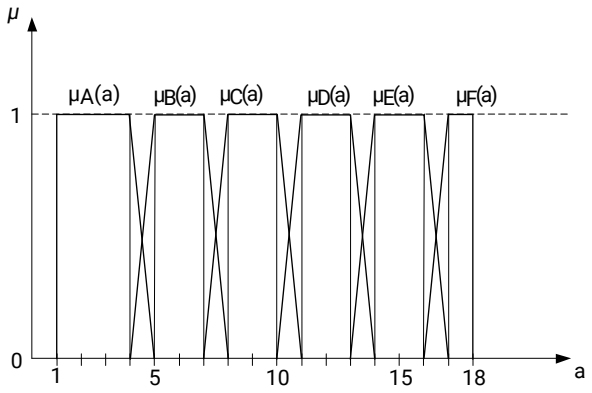

Figure 3. Graphs of functions belonging to fuzzy sets
Note: $\tilde{A}$ (few), $\tilde{B}$ (less than half), $\tilde{C}$ (half), $\tilde{D}$ (more than half), $\tilde{E}$ (many) and $\tilde{F}$ (full)
Source: developed by the author based on fuzzy sets and their universes proposed as an example

Figure 3 shows that the use of subdivision is appropriate to increase the coverage area of the accumulated graph of fuzzy set membership functions and, thereby, reduce the calculation error. But this process cannot be endless for two reasons. Firstly, at a certain level of subdivision, there is actually a transition from fuzzy sets back to clear values, which does not correspond to the task at hand. Secondly, the greater the fragmentation, the more difficult it will be for the side observer to give an estimate of the observation and, as a result, at a certain level, the observer will begin to hesitate between several sets, which will lead to the need to use both in calculations, and thus, in fact, bring the input values back to a lower level of subdivision.

In addition to the question of determining optimal universes, there is a question of possible uncertainty of another quantity – q_{avg}. The average passenger capacity of a fleet of vehicles may be unknown or not clearly defined. For example, the composition of a fleet can be determined by the model range, but not by the number of vehicles of a certain type. In this case, the equation for determining the passenger capacity utilisation rate (1) becomes inversely proportional to another fuzzy variable q_{avg}. Depending on the available variations, the number of fuzzy sets representing the variable, and their universes, the complexity and variability of the model increases. For clarity, a graph of the functions of belonging to a fuzzy variable is given $P_k(O_k)$ considering the variability of the variable q_{avg} (for the universe $\cup_Q = \{18, 24\}$) (Fig. 4).

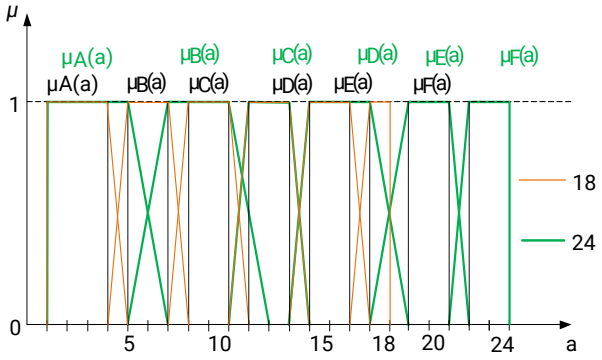

Figure 4. Graphs of functions belonging to fuzzy sets
Note: $\tilde{A}$ (few), $\tilde{B}$ (less than half), $\tilde{C}$ (half), $\tilde{D}$ (more than half), $\tilde{E}$ (many) and $\tilde{F}$ (full) considering the dependence on the fuzzy set $\tilde{Q}$
Source: developed by the author based on fuzzy sets and their universes proposed as an example

Figure 4 shows that the complexity of the model increases significantly as it expands with another fuzzy variable. Considering the fact that the dependence of the passenger capacity utilisation factor on the average passenger capacity of buses is inversely proportional, the dimension of the resulting tensor increases in proportion to the dimension of the fuzzy set universe . Thus, equation (3) takes the final form:

$$n = (2k+1) \cdot |\cup_Q|, \quad (7)$$

this clearly indicates the need to automate such calculations, due to their ultra-high complexity associated with large amounts of data, and/or a change in the approach of working with a fuzzy variable q_{avg}.

The results obtained demonstrate both the prospects of using fuzzy sets for modelling the passenger capacity utilisation rate, and certain problems associated with the complexity of calculations and the dependence of results on the choice of parameters of membership functions. This highlights the importance of automating calculations and further improving fuzzy data processing algorithms, in particular, to reduce the dimension of the resulting tensor and optimise the model. The proposed approach demonstrates the effectiveness of using fuzzy logic methods for working with data with limited accuracy. The results can be used to further optimise mathematical models in the field of public

transport, ensuring adaptation to real conditions of urban transport. This approach helps to improve the accuracy of modelling and efficiency of transport network management in conditions of fuzziness and uncertainty.

The expediency and practical significance of using fuzzy logic methods in modelling transport processes under variable and uncertain conditions, and limitations that should be considered to improve the accuracy and adaptability of models in real conditions, are reflected in many papers that confirm the results of the current study. For example, in a study on the use of a linguistic variable in the monitoring process (Kovtunov *et al.*, 2020), the dependence given in equation (6) was also derived using fuzzy logic. However, this relationship was used to determine the number of terms of a linguistic variable describing the process of monitoring transport communications.

In the dissertation work, I. Kara (2017) considered the use of fuzzy logic to represent vehicle fullness, where fuzzy sets were proposed and graphs of the corresponding membership functions were plotted. Comparison of the graph presented in the dissertation with Figure 1 provided in this study reveals the following. Similar shortcomings were identified, namely, insufficient accuracy of the proposed membership functions, due to the limitation of some values of the universes of the proposed fuzzy sets at the probability mark of 0.4. In contrast to the current study, no further work was carried out to improve accuracy in the paper.

In study by Q. Bao *et al*. (2024) used fuzzy logic as a key tool for solving the two-goal optimisation problem, which aims to consider the uncertainty and multi-factor planning of infrastructure for charging electric vehicles. The study shows the feasibility of using fuzzy logic when working with inaccurate, partially defined or subjective data. N. Jan *et al*. (2023) considered the use of interval-significant complex fuzzy sets for solving decision-making problems in transport strategy. The researchers proposed a model that allows evaluating complex, inaccurate, or contradictory information that occurs during the development of transport strategies in regional planning. The main focus is on multi-criteria tasks where there is a need to consider uncertainty in the input data.

In theoretical research by A. Calvi & S. Pozzi (2021) emphasised that fuzzy logic is a promising mathematical approach for modelling processes in transport engineering that are characterised by subjectivity, ambiguity, uncertainty, and inaccuracy. The main provisions of fuzzy logic systems were presented together with a detailed analysis of their application to solve various problems in transport engineering. Special attention was paid to the importance of fuzzy logic systems as universal approximators in solving transport problems, and the possibilities of further application of fuzzy logic in this area were considered.

S. Niroomand *et al*. (2024) proposed a new approach to solving transport problems with completely intuitive fuzzy parameters. An approach is proposed to transform an intuitive fuzzy transport problem into a multi-criteria one with clear parameters and apply a hybrid optimisation method to solve it. The results of computer experiments demonstrated the effectiveness of the proposed approach in comparison with the existing methods. The above methodology with high efficiency can be useful for solving real transport problems in conditions of uncertainty. In general, the approach to solving the fuzziness problem in the paper is close to the approach proposed in the current study. With the difference that guided by the problem of solving an intuitive-fuzzy problem, the researchers were forced to apply ranking functions in the development of defasification, which actually requires an individual approach to solving each individual problem and, accordingly, the need for careful selection of the appropriate function for a specific problem. Moreover, the approach proposed in this study, although it requires a similar individual approach, is reduced to the formation of universes and, accordingly, the level of subdivision, which is technically a simpler operation in comparison with the selection or formation of a ranking function.

A very close approach to solving the multi-criteria multi-product transport problem was proposed by M. Kar *et al.* (2018). The researchers proposed the use of trapezoidal fuzzy numbers, followed by the use of confidence theory to transform a problem with fuzzy parameters into a deterministic form. The study proposed the application of two approaches to optimisation: the fuzzy programming method and the global criteria method. By testing the example of a transport system for two types of products delivered to three destinations using two types of vehicles, the researchers demonstrated the effectiveness of a combined approach to reduce transportation costs and time. However, the model proposed in the paper was based on the use of trapezoidal fuzzy numbers to model parameters, while the model presented in this paper uses, but is not limited to, trapezoidal fuzzy sets as an example.

Comparison of the obtained results with the data of other studies devoted to the application of fuzzy logic methods in the absence of objective input data in the field of transport, showed that the proposed approach demonstrates compliance with global approaches to solving problems with fuzzy data. In comparison with some of the analysed methods that require the involvement of highly qualified specialists for the development of functions of belonging to fuzzy variables, the approach described in this paper allows the involvement of specialists with a lower level of qualification, sufficient to form universes and determine the necessary degree of subdivision, which is critical in modern conditions of personnel shortage in Ukraine. The results of the study showed the prospects of the proposed model, which not only ensures the accuracy of calculations, but also creates prerequisites for its scaling and implementation in real transport systems with minimal staff training costs.

Conclusions

The conducted research demonstrated the possibility and expediency of using fuzzy logic methods to solve the

problem of calculating the passenger capacity utilisation rate in the absence of objective input data. The developed mathematical model allows considering the subjective estimates of an outside observer, presented in the form of fuzzy sets, which significantly increases flexibility and reduces dependence on objective information, which can be difficult to obtain in modern conditions. The proposed model has shown efficiency when working with undefined or subjective data, providing a reduction in calculation error by accumulating membership functions and applying subdivision of fuzzy sets.

During theoretical studies to improve the accuracy of calculations, it was revealed that one of the main problems of applying fuzzy logic is to increase the complexity of calculations due to an increase in the dimension of tensors that occur in multi-stage calculations. This leads to the need to automate modelling and calculation processes. In addition, it was found that the use of a subdivision of fuzzy sets reduces the error of calculations, although at a certain stage this leads to a complication of estimates for outside observers and the impossibility of using a mathematical model for calculations in general.

The results of the study showed that the use of fuzzy logic is a promising area for solving the problems of organising passenger transportation by urban transport in conditions of dynamic changes in passenger flows, or the lack of objective input data. The developed model can be used to evaluate the operation of transport systems in conditions of military operations, migration processes, or other situations that make it difficult to collect accurate data.

Further research should be aimed at optimising the processes of subdivision of fuzzy sets, developing methods for automating calculations, and expanding the model to consider additional factors, such as fleet variability and route changes. It is also advisable to consider using other methods of fuzzy logic, such as linguistic variables or fuzzy sets of Type-2. This will improve the adaptability of transport systems to changing conditions and improve the accuracy of forecasting.

Acknowledgements
None.

Funding
The study received no funding.

Conflict of Interest
None.

References

[1] Bao, Q., Gao, M., Chen, J., & Tan, X. (2024). Location and size planning of charging parking lots based on EV charging demand prediction and fuzzy bi-objective optimisation. *Mathematics*, 12(19), article number 3143. doi: 10.3390/math12193143.

[2] Bhatia, T., Kumar, A., & Appadoo, S. (2023). *More-for-less solutions in fuzzy transportation problems*. London: Springer Nature. doi: 10.1007/978-3-031-30337-1.

[3] Bidyuk, P.I., Kalinina, I.O., & Gozhyi, O.P. (2021). *Bayesian data analysis*. Kherson: Book Publishing House FOP V.S. Vyshemyrskyi.

[4] Bilichenko, V.V., Tsymbal, S.V., & Tsymbal, O.V. (2020). Analysis of methods for determining the quantity and passenger capacity of rolling stock on urban passenger transport routes. *Bulletin of Mechanical Engineering and Transport, 2*, 11-18.

[5] Calvi, A., & Pozzi, S. (2021). Special issue: Fuzzy logic systems for transportation engineering. *Journal of Intelligent & Fuzzy Systems,* 41(6), 4705-4712. doi: 10.3233/JIFS-189957.

[6] Grebennik, I., & Kovalenko, O. (2024). A fuzzy decision-making model for an automatic postal sorting line. *ASU and Automation Devices*, 1(180), 16-26. doi: 10.30837/0135-1710.2024.180.016.

[7] Grosset, J., Oukacha, O., Fougères, A.-J., Djoko-Kouam, M., & Bonnin, J.-M. (2024). Fuzzy multi-agent simulation for collective energy management of autonomous industrial vehicle fleets. *Algorithms,* 17(11), article number 484. doi: 10.3390/a17110484.

[8] Hellekes, J., & Winkler, C. (2021). Incorporating passenger load in public transport systems and its implementation in nationwide models. *Procedia Computer Science*, 184, 115-122. doi: 10.1016/j.procs.2021.03.022.

[9] Jan, N., Gwak, J., Choi, J., Lee, S.W., & Kim, C.S. (2023). Transportation strategy decision-making process using interval-valued complex fuzzy soft information. *AIMS Mathematics*, 8(2), 3606-3633. doi: 10.3934/math.2023182.

[10] Kaczorek, M., & Jacyna, M. (2022). Fuzzy logic as a decision-making support tool in planning transport development. *Archives of Transport,* 61(1), 51-70. doi: 10.5604/01.3001.0015.8154.

[11] Kar, M.B., Kundu, P., Kar, S., & Pal, T. (2018). A multi-objective multi-item solid transportation problem with vehicle cost, volume, and weight capacity under fuzzy environment. *Journal of Intelligent & Fuzzy Systems,* 35(2), 1991-1999. doi: 10.3233/JIFS-171717.

[12] Kara, R. I. (2017). *Determination of passenger flows on urban routes using fuzzy logic and cellular subscriber transactions*. (Doctoral dissertation, Lviv Polytechnic National University, Lviv, Ukraine).

[13] Khabutdinov, R.A., & Fedorenko, I.O. (2021). Analysis of the influence of changing the passenger capacity utilization factor on transport energy-efficiency and motor vehicle emissions for city passenger transportation. *SWorldJournal*, 1(10-01), 76-89. doi: 10.30888/2663-5712.2021-10-01-041.

[14] Kovtunov, Y.O., Makogon, O.A., Isakov, O.V., Babkin, Y.V., Kalinin, I.V., & Lazuta, R.R. (2020). The use of the fuzzy logic mathematical apparatus for fuzzification and interactive monitoring of transport communications. *Modern Technologies in Mechanical Engineering,* 3(61), 65-72. doi: 10.26906/SUNZ.2020.3.064.

[15] Medvediev, I., Muzylyov, D., & Montewka, J. (2024). A model for agribusiness supply chain risk management using fuzzy logic. Case study: Grain route from Ukraine to Poland. *Transportation Research Part E: Logistics and Transportation Review,* 190, article number 103691. doi: 10.1016/j.tre.2024.103691.

[16] Naumov, V., Zhamanbayev, B., Agabekova, D., Zhanbirov, Z., & Taran, I. (2021). Fuzzy-logic approach to estimate the passengers' preference when choosing a bus line within the public transport system. *Communications – Scientific Letters of the University of Zilina*, 23(3), A150-A157. doi: 10.26552/com.C.2021.3.A150-A157.

[17] Niroomand, S., Allahviranloo, T., Mahmoodirad, A., Amirteimoori, A., Mršić, L., & Samanta, S. (2024). Solving a fully intuitionistic fuzzy transportation problem using a hybrid multi-objective optimization approach. *Mathematics*, 12(24), article number 3898. doi: 10.3390/math12243898.

[18] Richardson, T.W., Wu, W., Lin, L., Xu, B., & Bernal, E. A. (2020). MCFlow: Monte Carlo flow models for data imputation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 10500-10510). Seattle: IEEE. doi: 10.1109/CVPR42600.2020.01421

[19] Saatchi, R. (2024). Fuzzy logic concepts, developments and implementation. *Information,* 15(10), article number 656. doi: 10.3390/info15100656.

[20] Yang, X., Zhang, R., Li, Y., Yang, Y., Qu, D., Liu, T., & Zhao, B. (2022). Fuzzy-theory-based pedestrian dynamics models for studying the waiting passenger distribution at the subway platform. *Tunnelling and Underground Space Technology,* 129, article number 104680. doi: 10.1016/j.tust.2022.104680.

Застосування нечітких множин при розрахунку коефіцієнту використання пасажиромісткості в умовах неможливості збору об'єктивних даних

Іван Зьора

Аспірант
Вінницький національний технічний університет
21021, вул. Хмельницьке шосе, 95, м. Вінниця, Україна
https://orcid.org/0009-0000-2225-777X

Олександр Хошаба

Кандидат технічних наук, доцент
Вінницький національний технічний університет
21021, вул. Хмельницьке шосе, 95, м. Вінниця, Україна
https://orcid.org/0000-0001-5375-6280

Анотація. Задачі планування організації перевезення пасажирів міським транспортом у сучасних українських умовах стикаються з новими викликами, зокрема зі складністю або навіть неможливістю отримання точних вхідних даних для проведення розрахунків. Дослідження зосереджено на вирішенні проблеми недоступності точних і актуальних даних для розрахунків організації перевезення пасажирів міським транспортом шляхом використання методів нечіткої логіки. Передбачається, що за умов обмеженого часу для проведення натурних досліджень або впливу військових дій, що спричиняють динамічні зміни пасажиропотоків через міграційні процеси та унеможливлюють отримання даних традиційними методами, запропонований підхід дозволить виконати розрахунки з мінімальною похибкою. На прикладі коефіцієнту використання пасажиромісткості на перегоні транспортного маршруту, що прямо залежить від показника наповненості пасажирами, розглянуто можливість розширення математичної моделі організації перевезення пасажирів на міському транспорті за допомогою підходів нечіткої логіки. Зокрема, йдеться про заміну вхідних величин суб'єктивною оцінкою стороннього спостерігача у вигляді використання нечітких множин. Теоретичне дослідження показало можливість та доцільність використання нечітких множин для вирішення проблеми відсутності об'єктивних вхідних даних при розрахунках коефіцієнту використання пасажиромісткості. Визначено загальні принципи формування універсумів нечітких множин при їх використанні в математичних моделях організації перевезення пасажирів на міському транспорті з метою нівелювання суб'єктивності вхідних даних. Описано вимоги до ступеня перекриття акумульованими функціями належності нечітких множин допустимого рівня субдивізії, що може бути використано з метою зменшення похибки розрахунків та, відповідно, розмірності універсумів нечітких множин. Визначено залежність величини розрядності тензору вихідних результатів від кількісного показника перегонів на маршруті громадського транспорту, що може братися за основу при аналізі складності розрахунків. Показано загальні принципи роботи з нечіткими множинами в даній математичній моделі на прикладі розрахунку коефіцієнту використання пасажиромісткості. Дослідження може бути корисним міським адміністраціям, транспортним компаніям, розробникам програмного забезпечення, експертам з транспортної логістики та науковцям для оптимізації роботи громадського транспорту в умовах нестачі об'єктивних даних і динамічних змін

Ключові слова: громадський транспорт; організація транспортних маршрутів; пасажиромісткість транспорту; нечітка логіка; тензор

Journal homepage: https://itce.com.ua/en

Information Technologies and Computer Engineering **Vol. 22, No. 1. 2025**

Article's History: Received: 24.12.2024 Revised: 04.03.2025 Accepted: 24.04.2025

UDC 621.39:004.42 DOI: 10.63341/vitce/1.2025.124

Comparative analysis of load balancing methods based on SDN/NFV

Oleksandr Berestovenko*

Postgraduate Student
National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute"
03056, 37 Beresteiskyi Ave., Kyiv, Ukraine
https://orcid.org/0000-0003-4887-4674

Abstract. The purpose of the study was to determine the advantages and disadvantages of using load balancing methods in networks based on Software-Defined Networking and Network Functions Virtualisation technologies. Particular attention was paid to comparing the effectiveness of different approaches to balancing, including centralised and distributed methods, and the use of intelligent algorithms for load forecasting. The analysis helped to identify the advantages and disadvantages of each method, and their ability to adapt to changing network traffic conditions, considering such parameters as bandwidth, latency, packet loss, and energy efficiency. The paper discussed methods of load balancing in networks based on Software-Defined Networking and Network Functions Virtualisation technologies, which are important for ensuring the efficiency, scalability, and adaptability of modern networks. The key challenges faced by these technologies are described, such as the dynamism and unpredictability of traffic, resource optimisation, energy efficiency, and the integration of intelligent algorithms for load forecasting and energy consumption reduction. The study presents a comparison of various load balancing methods, including centralised and distributed traffic management, and the use of virtual balancers and adaptive traffic redirection algorithms. Particular attention was paid to analysing the impact of these methods on throughput, latency, packet loss, and energy efficiency under different traffic conditions. The role of machine learning in optimising load balancing processes, and the possibilities of integrating Software-Defined Networking and Network Functions Virtualisation into hybrid networks were considered. According to the results of the study, the use of balancing methods based on Software-Defined Networking/Network Functions Virtualisation can significantly improve network efficiency, reduce latency and increase throughput, while reducing energy consumption under high loads. Key results for Ukraine, where the integration of Software-Defined Networking/Network Functions Virtualisation into the telecommunications infrastructure can become the basis for improving the quality of services, optimising costs, and ensuring a high level of security in the context of digital transformation and infrastructure modernisation, are derived

Keywords: network functions virtualisation; intelligent algorithms; resource optimisation; hybrid structures; energy efficiency

Introduction

Modern network infrastructures, especially in large corporations and data centres, face unprecedented workload due to the growth of data volumes, the active introduction of cloud technologies and the Internet of Things (IoT). In the context of digital transformation, the issue of ensuring reliability, scalability, and efficiency of traffic management in networks is becoming particularly relevant. This is where innovative approaches such as Software-Defined Networking (SDN) and network functions based on NFV (Network Functions Virtualisation) come to the fore, offering radically new approaches to optimisation and load balancing.

Load balancing is one of the key tasks of the network infrastructure, because the efficiency of resource allocation depends not only on the quality of user service, but also on the performance of the network as a whole. Conventional approaches to load balancing are often limited by static rules, which is not always effective under variable network conditions and high traffic dynamics. Using SDN and NFV technologies to implement dynamic balancing techniques allows networks to automatically adapt to the current load, ensuring efficient resource allocation in real time. SDN technology separates network management from its

Suggested Citation:
Berestovenko, O. (2025). Comparative analysis of load balancing methods based on SDN/NFV. *Information Technologies and Computer Engineering*, 22(1), 124-133. doi: 10.63341/vitce/1.2025.124

*Corresponding author

physical infrastructure, helping to centrally monitor traffic and provide flexible load balancing based on user needs. NFV, in turn, allows migrating network functions to virtual platforms, which simplifies resource management and reduces dependence on hardware.

The problem is that different balancing methods based on SDN and NFV have their own characteristics and limitations. For example, some methods may be more effective for cloud environments, while others may be more effective for data centres or IoT infrastructure. There are also questions about reliability, scalability, security, and speed of response to changes. Thus, a comparative analysis of existing SDN/NFV-based load balancing methods is necessary to select the optimal approach that meets modern network infrastructure requirements.

The study by J. Billingsley *et al.* (2020) analysed adaptive balancing methods for cloud services, where the load often varies depending on the number of users. Researchers have found that SDN/NFV technologies significantly improve resource efficiency in cloud environments, improving performance. A. Das *et al.* (2025) considered security aspects related to load balancing, in particular, protection against DDoS attacks. Using SDN, the researchers created a system that not only distributes the load, but also identifies abnormal network activity. A.M. Zarca *et al*. (2019) analysed the effectiveness of various balancing methods for IoT networks, where the use of SDN/NFV proved useful for managing a large number of devices. They noted that NFV technology allows effectively managing different levels of device priority.

A. Filali *et al.* (2020) examined methods for predictive load balancing based on historical data. Using SDN and machine learning algorithms, they created a system that ca n predict peak load moments and allocate resources accordingly. Z. Song *et al*. (2019) reviewed balancing methods adapted for e-commerce systems that require high reliability and fast response to user requests. Research has shown that SDN/NFV-based methods can reduce request processing delays. M. Alenezi *et al*. (2019) focused on opportunities to reduce infrastructure costs through the use of SDN/NFV in telecommunications networks. This study proved that centralised control and virtualisation can reduce the need for physical hardware. The study by Q. Zhou *et al*. (2021) compared static and dynamic balancing methods. The researchers found that dynamic SDN-based methods distribute the load more efficiently in conditions of high traffic changes. L. Zhu *et al*. (2020) compared the impact of different protocols on SDN network performance. They found that the choice of protocol can significantly affect the speed of request processing, and in particular, OpenFlow protocols provide the highest performance. In addition, P.P. Ray & N. Kumar (2021) examined the impact of SDN/NFV on network flexibility in computing power allocation in multi-cloud environments. The results showed a significant reduction in infrastructure maintenance costs while increasing network performance.

An analysis of available studies showed that SDN/NFV-based load balancing methods have demonstrated their effectiveness, but there are still several aspects that need to be studied in more depth. Among the unexplored or under-covered topics were the adaptability of methods to heterogeneous network environments, such as multi-cloud environments or hybrid networks; energy optimisation, which is important in the context of growing demands for green technologies; and the use of artificial intelligence and machine learning for load forecasting and dynamic resource allocation in SDN/NFV networks.

The purpose of this study was to investigate the effectiveness of various load balancing approaches using SDN/NFV technologies by benchmarking them. To achieve this goal, the following tasks were solved: to compare the efficiency of various SDN/NFV load balancing methods in multi-cloud and hybrid environments; to evaluate the energy consumption of these methods and their compliance with modern energy efficiency requirements.

Materials and Methods

An integrated approach to the study of network resource management was applied to analyse load balancing methods in networks based on SDN and NFV technologies. This made it possible to determine the efficiency, flexibility and scalability of the network infrastructure in the context of modern requirements for telecommunications systems. The study used two key approaches to load balancing: centralised management and decentralised balancing. For the first approach, the SDN controller plays a central role, which makes decisions based on global network status information. In decentralised balancing, each node makes autonomous decisions using only local data. This approach provides greater reliability in the event of a central element failure, but requires more calculations on each node.

Machine learning algorithms that can predict loads based on historical data were used for modelling and analysis. Methods for analysing traffic patterns considering peak loads were integrated, which ensured optimisation of traffic routing in dynamic and complex network environments. In particular, regression, clustering, and adaptive learning algorithms were considered, which allowed improving the accuracy of forecasts. The effectiveness of balancing methods was evaluated based on quantitative parameters such as throughput, latency, packet loss, power consumption, and adaptability under variable load conditions. These indicators were selected as the most important for evaluating the performance of the network infrastructure. To obtain the results, simulations were used with the help of tools such as Mininet for modelling SDN networks and specialised Python libraries for implementing machine learning algorithms.

Special attention was paid to energy efficiency analysis, which is an important aspect for modern data centres. A comparison of different algorithms, such as Round Robin, Least Connection, Weighted Least Connection, Shortest Queue First and Adaptive Load Balancing, helped determine which method is best suited to balancing certain types of load under different network conditions to reduce energy consumption and make optimal use of network

resources. Various scenarios were modelled in experiments, in particular peak load conditions, to assess the system's behaviour in critical situations. Reliability and security issues were also considered, as a centralised SDN controller can become system vulnerability. To reduce such risks, approaches to implementing backup management mechanisms that ensure uninterrupted network operation even in the event of a central node failure were investigated. To assess the flexibility of network solutions, special attention was paid to multi-cloud and hybrid environments that require rapid adaptation to changes in network structure and traffic. Two scenarios for using SDN/NFV in such environments were modelled to determine optimal configurations.

All experiments were carried out considering the needs of modern telecommunications systems, in particular, the need to ensure high bandwidth, low latency, and high network availability, which is becoming increasingly important due to process automation. The obtained data were analysed for such parameters of efficiency of methods as adaptability to variable traffic, power consumption, and speed of response to loads. The analysis was based on a comparison of different approaches to load balancing, including centralised and decentralised methods, considering specific infrastructure requirements and system performance in real-world operating conditions.

Results

Load balancing in SDN and NFV-based networks is a complex and multidimensional task that involves a number of technical and organisational challenges. One of the main problems is the dynamism and unpredictability of traffic, because the volume of traffic often changes due to user demand, peak loads, or specific events, such as new product releases or network attacks. This creates a significant load on the system and requires constant dynamic balancing to avoid congestion and ensure stable network operation. The second important problem is resource optimisation: load balancing in an SDN/NFV environment should consider the available resources of each network node and optimise them as much as possible, considering the different requirements of virtualised network functions (VNF), such as CPU time, memory, and bandwidth. In addition, energy efficiency is an important aspect: load balancing should help to reduce power consumption, especially for large data centres. Using SDN/NFV can help to allocate resources more efficiently, but it requires specialised algorithms to reduce energy consumption (Tipantuna & Hesselbach, 2020). There is also the question of integrating intelligent algorithms, because modern networks are extremely complex and variable, which requires intelligent approaches such as machine learning. The implementation of such algorithms faces challenges related to computing costs, training models, and ensuring real-time operation, and raising issues of data security and privacy.

Another important challenge is scalability and flexibility, especially for multi-cloud and hybrid environments that require adaptive solutions to handle large volumes of traffic and provide a fast response to structural or load changes. In addition, reliability and security must be considered, since a centralised SDN controller increases network vulnerability and must be protected from threats that may disrupt the balancing or stable operation of network functions. Another important problem is latency and throughput: when implementing load balancing, it is necessary to minimise latency and maintain high network throughput, which is a difficult task with large volumes of traffic and interaction with different types of infrastructure (George, 2022). Integration of SDN and NFV is an important technological component of modern network systems, because both of these technologies are aimed at improving the flexibility, scalability, and efficiency of network infrastructure. When combined, SDN and NFV offer the ability to dynamically manage network resources and quickly adapt the network to changing needs, which is especially important in an era of growing traffic and performance requirements (Bonfim *et al*., 2019).

The basic principle of integration was to use an SDN controller to manage network traffic and routing, while NFV was responsible for virtualising network functions (such as routers, firewalls, load balancers) on public hardware. SDN allows centrally managing traffic routing, movement, and prioritisation using software that provides fast network configuration (Kaur *et al*., 2020). NFV, in turn, allows performing functions normally performed by specialised hardware on virtual machines or containers.

SDN and NFV integration provides several key benefits. Network management flexibility increases: network functions can be quickly migrated or scaled to meet traffic needs without the need for physical hardware replacement. This significantly reduces the cost of network deployment and maintenance, and allows quickly adapting to new business requests. SDN/NFV allows centrally monitoring network resources and providing efficient traffic management, which greatly simplifies load balancing and improves network performance. However, the integration of SDN and NFV is not without its challenges. Ensuring compatibility between components that may come from different manufacturers is challenging. The lack of generally accepted standards for certain aspects of SDN/NFV technologies can complicate deployment and lead to compatibility issues. The integration of these technologies increases security requirements, as a centralised SDN controller can become vulnerability in the infrastructure, creating network security risks. In addition, with the integration of SDN and NFV, the requirements for computing resources and monitoring systems are increasing. Managing and monitoring the status of virtualised functions requires high performance and reliable monitoring tools to identify and solve potential problems in real time. To reduce computing costs, artificial intelligence and machine learning technologies are being actively explored, which can optimise resource allocation and predict network loads. SDN/NFV-based load balancing methods are key components of modern network infrastructures, as they allow dynamic and efficient load

distribution between different network nodes, considering performance requirements and minimising the likelihood of delays or congestion (Nawaz *et al.*, 2024). One of the main approaches to load balancing in SDN is to centrally manage traffic through the SDN controller, which receives information about the state of the network from all its parts and optimally distributes traffic between network nodes. The controller can predict future load based on current and past trends, and based on this information, change traffic routes to avoid congestion (Lakhani & Kothari, 2020). This approach reduces latency and ensures efficient use of network resources.

Distributed load balancing is an alternative method in which traffic balancing occurs without centralised control. In this case, each network node independently determines how to distribute traffic based on local network status information. Distribution balancing uses routing and adaptive network algorithms, in particular traffic hashing, which allows evenly distributing the load between different routes without the need for a central controller, which reduces the load on the central system and increases scalability.

NFV technologies allow dynamically changing the infrastructure by virtualising network functions, such as routers or load balancers. This allows quickly scaling the network by adding or reducing the number of virtual functions depending on the current load. This approach not only improves network performance, but also significantly reduces physical hardware costs, since functions can be performed on public virtual platforms. The use of machine learning algorithms opens up new opportunities for load balancing, allowing the prediction of loads based on historical data and optimisation of traffic routing using intelligent systems that can adapt to changes. Machine learning algorithms can analyse traffic patterns and predict peak loads, which allows performing balancing in advance and reducing the risk of network congestion.

In complex hybrid networks that use both conventional technologies and SDN, load balancing must consider both virtual and physical resources. Balancing mechanisms can integrate different routing strategies and adapt to different types of traffic, which ensures high network performance and reliability. An important part of this approach is the use of policies that optimise resource usage and minimise congestion in certain parts of the network. Thus, the use of various load balancing methods based on SDN/NFV allows ensuring high efficiency of networks, increasing their scalability, and reducing the likelihood of congestion. The choice of a particular method depends on the specifics of the network and its needs for flexibility, speed of adaptation, and efficiency of resource management.

Load balancing is a key task in modern Software-Defined Networks (SDN) and Network Features Virtualisation (NFV). The efficiency of this process significantly affects the quality of network operation, including bandwidth, latency, packet loss, and overall power consumption. Different methods, such as centralised and decentralised management, virtual balancers, and adaptive redirection algorithms, show different performance depending on traffic characteristics and environmental conditions. Table 1 summarises the main comparison parameters, such as bandwidth, latency, packet loss, power consumption, and adaptability to load changes.

Table 1. Comparative characteristics of load balancing methods in SDN/NFV networks by key metrics

Method	Bandwidth (Gbps)	Delay (ms)	Packet loss (%)	Power consumption (W)	Flexibility for traffic changes
Least delay	8.5	15	2.5	250	Average
Adaptive redirection	9.2	12	1.8	300	High
Virtual balancer	10	10	1.2	280	High
Centralised management	7.5	20	3	200	Low
Decentralised management	8	18	2	220	Average

Source: created by the author based on N. Giri *et al*. (2018), M.F. Monir & D. Pan (2021)

The analysis showed that different load balancing methods have their advantages and disadvantages. The virtual balancer showed the best results in terms of bandwidth, minimising latency and packet loss. It was also highly adaptable to changes in network traffic, making it suitable for complex and dynamic networks. However, its power consumption has proven to be high, which can be a critical factor for some environments. Centralised and decentralised management methods have different levels of energy consumption, which is an important aspect for assessing their compliance with modern energy efficiency requirements. Centralised management, due to the centralisation of resources, can reduce energy consumption by optimising the use of equipment, although it is inferior in performance in large and distributed networks. On the other hand, decentralised management, despite increased energy consumption, has advantages in scenarios with high resource allocation, where energy efficiency is less critical compared to flexibility and adaptability.

The lowest latency method, while showing average adaptability, also meets energy efficiency requirements in stable traffic conditions, where the need to adapt to variable loads is lower. However, for more dynamic environments where high adaptability and scalability are required, additional energy saving mechanisms may be needed, such as the use of specialised hardware platforms or algorithms optimised for energy efficiency. Thus, the choice of balancing method should consider not only performance and adaptability, but also compliance with energy efficiency requirements, which include minimising energy

consumption while maintaining the required level of efficiency for different types of network environments. The results of the study can become the basis for improving methods of balancing and further optimising energy consumption in SDN/NFV networks that meet modern energy efficiency requirements.

With the introduction of SDN and NFV technologies, it became necessary to improve load balancing methods in networks to effectively manage traffic and ensure high quality of services. The main goal of such methods is to optimally distribute the load between servers, routers, and other network components to prevent congestion and reduce delays, especially in dynamic traffic conditions. There are several different approaches to load balancing, each with its own advantages and disadvantages, depending on the specifics of the network and its performance requirements. Load distribution in SDN/NFV networks can significantly improve the adaptability of the network, its ability to scale, and its response to changing load conditions. Since networks with SDN and NFV differ from conventional ones in that they can be configured programmatically to optimise traffic, the chosen balancing method should consider the dynamism and variability of traffic, and the resource requirements that are available in each specific situation.

Table 2 compares the main load balancing algorithms used in SDN/NFV networks, with a focus on important parameters such as response speed, real-time efficiency, scalability, and application in various environments. The table helps to assess how each algorithm behaves in a particular environment, which allows network developers and engineers to select optimal solutions for implementation in real-world conditions, which is an important stage for ensuring the smooth operation of critical network services, in particular, in the context of 5G and IoT development.

Table 2. Comparison of load balancing algorithms in SDN/NFV networks: efficiency and application in different network environments

Balancing algorithm	Network type	Load response time	Real-time efficiency	Scalability	Application
Round Robin	SDN/NFV	Moderate	Average	Low	Small and medium-sized networks
Least Connection	SDN/NFV	Fast	High	Average	Networks with uneven load
Weighted Least Connection	SDN/NFV	Fast	High	Average	Networks with different resource requirements
Shortest Queue First	SDN/NFV	Very fast	High	High	Networks with high bandwidth requirements
Adaptive Load Balancing	SDN/NFV	Very fast	Very high	Very high	Large networks, high-load environments

Source: created by the author based on S. Rout *et al*. (2020), T.G. Thajeel & A. Abdulhassan (2021)

Each of the presented algorithms has its own advantages and limitations. For example, the Round Robin algorithm is easy to implement, but its performance is limited in large or high-load networks. At the same time, Least Connection and Weighted Least Connection show better results when handling uneven loads, although they require more computing resources. The Shortest Queue First and Adaptive Load Balancing algorithms provide high real-time efficiency, making them optimal for large and dynamic networks. Since each of the methods has its own characteristics, it is important to choose an algorithm depending on the specifics of the network and its requirements. The use of these technologies allows optimising resources, reducing latency, and increasing network throughput. The use of intelligent algorithms for load forecasting also opens up new opportunities for achieving high efficiency in variable traffic conditions.

Modern information systems require not only high performance, but also reliability. The use of SDN/NFV technologies allows implementing adaptive backup mechanisms that increase fault tolerance, ensuring the continuity of services in different environments. Among the key redundancy mechanisms, there are three approaches: active-active, active-passive, and policy-based redundancy. The active-active model demonstrates the highest efficiency due to the simultaneous operation of several nodes, although it requires significant resources. Instead, the active-passive approach saves resources, but is more vulnerable to delays in switching. Policy-oriented redundancy offers flexibility, but depends on the effectiveness of the monitoring system.

In multi-cloud environments, SDN/NFV helps to optimise resources through platform integration. For example, dynamic traffic balancing between cloud providers minimises latency and distributes the load evenly. Another aspect is the centralised implementation of security features, such as virtual firewalls, which allow protecting data in distributed clouds. In hybrid environments, these technologies play a critical role in automating resource management. Integration allows smoothly switching between on-premises and cloud resources, reducing system downtime. Additionally, the use of virtualised backup functions ensures business continuity by allowing traffic to be quickly redirected in the event of a primary infrastructure failure.

The implementation of these scenarios illustrates the potential of SDN/NFV to overcome conventional network constraints. It provides high performance, flexibility, and reliability in multi-level environments, which is an important step towards the dynamic and adaptive infrastructures of the future. Table 3 shows a summary of the key benefits and challenges of each environment. This provides a clear picture of the practical implementation of SDN/NFV in multi-cloud and hybrid scenarios, which are the basis for sustainable network infrastructures.

Table 3. Comparison of SDN/NFV usage scenarios in multi-cloud and hybrid environments

Category	Multi-cloud environment	Hybrid environment
Backup mechanisms	Load balancing between providers; optimising data routes to ensure fault tolerance	Use on-premises reserves and private clouds to ensure continuity of work
Security	Centralised monitoring; tunnelling to protect data transfer between clouds	Access control to local data; integration with private cloud security policies
Scenario 1: Scalability	Additional resources are connected in real time to handle peak load	Use of cloud infrastructure as a reserve to improve productivity during peak times
Scenario 2: Recovery	Automatically redirects traffic to an alternative provider in the event of a failure	Local storage of critical data with the ability to automatically switch to the cloud in case of failures

Source: created by the author based on M.S. Bonfim *et al*. (2019), H.U. Adoga & D.P. Pezaros (2022)

In a multi-cloud environment, the use of SDN/NFV allows creating effective strategies for traffic management and security at all infrastructure levels. SDN allows centrally managing security policies, optimising resource usage to reduce costs and improve performance. NFV allows quickly implementing new network features, reducing dependence on physical devices and increasing flexibility. For example, in the process of automating the allocation of resources between different clouds, it is possible to reduce response time and improve the ability to scale.

However, in a hybrid environment where both public and private clouds are combined, SDN/NFV provides the ability to provide more reliable connection management between different cloud platforms. Due to integration with NFV, it is possible to organise effective monitoring and management of resources to ensure stability and security in the face of constant changes. One of the main advantages of this approach is the ability to ensure continuous operation of the enterprise, even if some resources are decommissioned or transferred to another platform to perform tasks. Therefore, the integration of SDN/NFV into multi-cloud and hybrid environments allows achieving a high level of adaptability, reducing costs, and improving the management and security characteristics of the infrastructure. This is especially important for organisations that work with large amounts of data and have a complex infrastructure that requires flexibility in load distribution and business process continuity. This development towards using SDN/NFV has huge potential to save resources and improve infrastructure efficiency. These technologies allow Ukrainian telecommunication operators to introduce new services, ensure optimal use of their networks and improve the quality of service, which is especially important in the context of the growing demand for digital technologies in the country.

For Ukraine, the issue of load balancing in SDN and NFV networks is of particular relevance in the context of digital transformation and infrastructure modernisation. Given the growing demand for high-quality Internet access, the introduction of these technologies can significantly improve the efficiency of traffic management in telecommunications systems. In addition, the introduction of SDN/NFV methods allows Ukrainian telecommunication operators to optimise costs, provide greater flexibility in deploying services, and adapt to the growing needs of both the commercial and public sectors. In addition, the development of SDN/NFV can become the basis for improving cybersecurity, which is especially important in the face of modern challenges. The participation of Ukrainian IT companies in the development of solutions based on these technologies contributes to their integration into the international innovation market. Given limited resources, the introduction of more energy-efficient traffic balancing techniques is also an important factor in reducing operating costs and maintaining environmental sustainability. Further research on the effectiveness of these methods in local conditions can become the foundation for their adaptation to the realities of Ukrainian network infrastructures.

However, these prospects require government support and a strategic approach to implementing SDN/NFV. It is important to train specialists who can develop, implement, and maintain such solutions, and ensure their compliance with cybersecurity standards. Thus, research and adaptation of SDN/NFV can contribute to a significant improvement in the network infrastructure of Ukraine, laying the foundation for further digitalisation of the economy and society.

Discussion

The study of load balancing methods in networks based on SDN and NFV technologies has shown that these approaches can significantly increase the efficiency of network management, reducing the likelihood of congestion and improving performance. A central aspect of the study was the comparison of two main methods: centralised and decentralised load balancing. The results showed that each of them has its own advantages and disadvantages, depending on the requirements for scalability, flexibility, and energy efficiency of the network.

Centralised management, where the SDN controller acts as a single decision centre, allows achieving high efficiency in conditions of stable and predictable traffic. However, in the face of rapid load changes, this method may have limitations due to the need to process large amounts of data centrally. The study by M.D. Tache *et al.* (2024) considered a centralised approach to load balancing in SDN networks using conventional routing algorithms. The results showed that centralised management is effective in networks with constant and predictable traffic, but its efficiency is significantly reduced at high load peaks. The researchers emphasise that centralised systems have

limits on the speed of response to changes in the network, which causes delays in processing requests. In this study, it was found that centralised balancing ensures stable network operation in environments with low traffic variability, but its efficiency is significantly reduced in conditions of high dynamics or under unexpected loads. F. Chahlaoui & H. Dahmouni (2020) pointed to the problem of optimising centralised load balancing in SDN networks, in particular, in the context of using conventional routing algorithms. The main objective of the study was to determine the effectiveness of the centralised approach in conditions of constant and predictable traffic, and to identify the limitations of this approach when working with high load peaks, such as traffic spikes or processing large amounts of data in real time. The researchers focused on centralised approaches under stable load, while the current study focused on flexible and adaptive algorithms that can effectively respond to variable conditions. In addition, the use of NFV and machine learning to optimise system performance under high traffic peaks was analysed.

On the other hand, decentralised load balancing, where each node of the network independently makes decisions about traffic routing, demonstrated a high level of adaptability to changes in load. In situations with extreme load or unstable traffic, this method can significantly improve response time and reduce the likelihood of congestion on specific nodes. Z. Nezami *et al*. (2021) investigated decentralised load balancing in SDN networks using multi-level traffic management techniques. They note that the decentralised approach has a significant advantage in the context of variable traffic, since each node can make its own decisions. However, the researchers also pointed to the problem of increasing delays and load on nodes with a very large number of connected devices. These findings partially coincide with the current ones, since the high level of adaptability of the decentralised approach to variable traffic was equally indicated. However, the current study highlights the additional difficulties associated with optimising power consumption and integrating with NFV technologies, which was not emphasised by researchers. However, decentralised systems require more sophisticated algorithmic provisioning to maintain network balancing, which can increase the load on individual nodes and cause delays with high traffic volumes.

X. Jiang *et al*. (2021) examined the effectiveness of using distributed balancing methods, including hashing-based algorithms for traffic distribution. They obtained a result that showed that distributed balancing gives better results in large, heavily loaded networks compared to centralised control. This confirms current conclusions about the importance of flexibility and scalability in networks, but the researcher did not consider power consumption and its impact on overall system efficiency. In the current results, energy efficiency was an important component.

One of the most important aspects of the study was the use of machine learning technologies to predict traffic and optimise balancing processes. Algorithms that can predict network loads can reduce the likelihood of overloads and optimise resource usage. S. Liang *et al*. (2020) examined the role of machine learning algorithms in load prediction for traffic balancing in SDN networks. They found that using machine learning can reduce the likelihood of network congestion by predicting the load using models based on historical data. However, they also noted that the accuracy of such forecasts depends on the quality and relevance of training data. This is especially true for complex and dynamic networks, where changes in traffic can be unpredictable. The results are completely consistent with current data on the importance of machine learning for load forecasting. Both studies pointed to limitations related to the relevance of data, but the current one has added an aspect of using hybrid approaches to improve adaptability to new situations, which the researchers did not provide.

The study also drew attention to aspects of energy efficiency in load balancing. The use of NFV technologies reduces power consumption by virtualising network functions, which reduces the need for specialised hardware. However, when using virtual load balancers, there may be an increase in power consumption compared to conventional hardware solutions. T.V.K. Buyakar *et al*. (2019) investigated the efficiency of using network function virtualisation for load balancing. They concluded that virtualisation can significantly reduce hardware costs and improve network scalability. However, virtualisation processes can be energy-intensive and require additional optimisation to maximise efficiency. The researchers' findings coincide with the current results regarding the energy intensity of virtualisation and the need to optimise it. However, in the current study, more attention is paid to integrating NFV with other technologies to improve energy efficiency, which was not indicated in the paper by T.V.K. Buyakar *et al*. This indicates the need for additional optimisation of energy aspects in the design of networks based on SDN and NFV. R. Moosavi *et al*. (2021) investigated the effectiveness of SDN and NFV integration in the context of energy conservation. They found that these technologies can reduce the overall power consumption of networks, but require detailed configuration to avoid unnecessary power consumption due to excessive virtualisation. The results of the study under consideration and the current one partially coincide in terms of energy consumption. However, the current study focuses more on load balancing, and it notes that energy aspects are only part of the overall network optimisation strategy. The integration of these technologies with other modern methods, such as multichannel or hybrid networks, opens up new opportunities for achieving greater flexibility and scalability. For example, using hybrid cloud environments with dynamic resource allocation capabilities can further improve load balancing efficiency.

U.K. Jena *et al*. (2022) focused on comparing different load balancing methods in hybrid networks that use both physical and virtual resources. They found that hybrid networks can provide high flexibility and adaptability, but the complexity of managing such networks increases significantly.

Comparison with current results shows similarities in understanding the importance of flexibility in load balancing, but emphasis has been added on using machine learning to improve the management of such networks, which is a more detailed approach in the current study. O. Pidpalyi (2024) analysed the prospects for integrating machine learning algorithms to optimise routes and traffic in the telecommunications industry. According to his conclusions, the use of these technologies helps to increase network performance by reducing latency and improving throughput by predicting traffic changes based on historical data. This is very similar to the current results, which also focused on using machine learning to predict the load, but the researcher did not consider factors such as power consumption and cost of calculations when implementing such algorithms. A.A. Ibrahim *et al*. (2020) investigated the application of load balancing methods using a centralised controller in SDN/NFV environments. The main focus was on reducing latency and power consumption, which are also important aspects in this paper. However, in contrast to the current results, the researcher did not factor in changing load conditions and mainly considered stable traffic, which limits their approach to dynamic networks. They also emphasised the importance of using less energy-intensive methods, such as those with the lowest latency, which is partly similar to current data on energy efficiency, but his method provided less flexibility when adapting to changes in traffic.

Based on the results of the discussion, it can be concluded that the optimal approach to load balancing depends on the specific conditions and requirements for the network. In the future, it is worth continuing to work on integrating machine learning to provide greater flexibility and efficiency in balancing processes. SDN and NFV technologies have great potential to improve network performance, but it is necessary to consider the specifics of infrastructure and traffic to maximise their efficiency.

Conclusions

The study was aimed at comparative analysis of load balancing methods based on SDN and NFV technologies, with an emphasis on their efficiency and compliance with modern requirements for energy efficiency and adaptability in dynamic traffic conditions. As a result of the analysis, it was revealed that centralised management methods have advantages in energy consumption, but they are inferior in performance in conditions of distributed resources. In addition, decentralised approaches, although require higher power consumption, provide high adaptability, which makes them optimal for scenarios with high resource allocation. The method of the lowest delay has shown its effectiveness in conditions of stable traffic, while meeting the requirements of energy efficiency.

The study showed that SDN/NFV is the most effective for such networks. They combine centralised and decentralised management architectures, providing a balance between power consumption and adaptability in a variable traffic environment. Such methods can provide high performance, reduce latency, and optimise power consumption depending on specific usage conditions. In addition, combined approaches help to reduce infrastructure costs and improve overall network efficiency. Overall, the study confirmed the importance of considering energy constraints and dynamic traffic requirements when choosing balancing methods in SDN/NFV networks, which can significantly increase their efficiency and productivity, and optimise energy and resource costs in the face of constant data growth.

The prospects for further research are to expand scenarios for testing load balancing methods in real-world environments, in particular in 5G, IoT, and cloud computing networks, and to implement machine learning for automatic traffic optimisation. In addition, an important area is to improve the energy efficiency of decentralised systems and integrate SDN/NFV with the latest technologies, such as blockchain and quantum computing.

Acknowledgements

None.

Funding

The study received no funding.

Conflict of Interest

None.

References

[1] Adoga, H.U., & Pezaros, D.P. (2022). Network function virtualization and service function chaining frameworks: A comprehensive review of requirements, objectives, implementations, and open research challenges. *Future Internet*, 14(2), 59. doi: 10.3390/fi14020059.

[2] Alenezi, M., Almustafa, K., & Meerja, K.A. (2019). Cloud based SDN and NFV architectures for IoT infrastructure. *Egyptian Informatics Journal*, 20(1), 1-10. doi: 10.1016/j.eij.2018.03.004.

[3] Billingsley, J., Miao, W., Li, K., Min, G., & Georgalas, N. (2020). Performance analysis of SDN and NFV enabled mobile cloud computing. In *GLOBECOM 2022 – 2022 IEEE Global Communications Conference* (pp. 1-6). Taipei: IEEE. doi: 10.1109/globecom42002.2020.9322530.

[4] Bonfim, M.S., Dias, K.L., & Fernandes, S.F. (2019). Integrated NFV/SDN architectures: A systematic literature review. *ACM Computing Surveys (CSUR)*, 51(6), article number 114. doi: 10.1145/3172866.

[5] Buyakar, T.V.K., Agarwal, H., Tamma, B.R., & Franklin, A.A. (2019). Prototyping and load balancing the Service Based Architecture of 5G core using NFV. In *2019 IEEE Conference on Network Softwarization* (pp. 228-232). Paris: IEEE. doi: 10.3233/jifs-189706.

[6] Chahlaoui, F., & Dahmouni, H. (2020). A taxonomy of load balancing mechanisms in centralized and distributed SDN architectures. *SN Computer Science*, 1, article number 268. doi: 10.1007/s42979-020-00288-8.
[7] Das, A., Nanda, P., Jain, R., Saini, T., Bhaskar, S., & Mohapatra, H. (2025). Security considerations of SDN networks during DDoS Attacks in load balancing. In *Human impact on security and privacy: Network and human security, social media, and devices* (pp. 123-140). Hershey: IGI Global. doi: 10.4018/979-8-3693-9235-5.ch007.
[8] Filali, A., Mlika, Z., Cherkaoui, S., & Kobbane, A. (2020). Preemptive SDN load balancing with machine learning for delay sensitive applications. *IEEE Transactions on Vehicular Technology*, 69(12), 15947-15963. doi: 10.1109/TVT.2020.3038918.
[9] George, J. (2022). Optimizing hybrid and multi-cloud architectures for real-time data streaming and analytics: Strategies for scalability and integration. *World Journal of Advanced Engineering Technology and Sciences*, 7(1), 174-185. doi: 10.30574/wjaets.2022.7.1.0087.
[10] Giri, N., Kukreja, V., Panchi, D., Sajnani, J., & Seedani, H. (2018). Performance evaluation of load balancing algorithms for SDN. In *2018 Fourth international conference on computing communication control and automation* (pp. 1-4). Pune: IEEE. doi: 10.1109/ICCUBEA.2018.8697762.
[11] Ibrahim, A.A., Hashim, F., Noordin, N.K., Sali, A., Navaie, K., & Fadul, S.M. (2020). Heuristic resource allocation algorithm for controller placement in multi-control 5G based on SDN/NFV architecture. *IEEE Access*, 9, 2602-2617. doi: 10.1109/ACCESS.2020.3047210.
[12] Jena, U.K., Das, P.K., & Kabat, M.R. (2022). Hybridization of meta-heuristic algorithm for load balancing in cloud computing environment. *Journal of King Saud University-Computer and Information Sciences*, 34(6), 2332-2342. doi: 10.1016/j.jksuci.2020.01.012.
[13] Jiang, X., Yang, H., Yang, Y., & Chen, Z. (2021). Cluster load balancing algorithm based on dynamic consistent hash. *Journal of Intelligent & Fuzzy Systems*, 41(3), 4461-4468. doi: 10.3233/jifs-189706.
[14] Kaur, K., Mangat, V., & Kumar, K. (2020). A comprehensive survey of service function chain provisioning approaches in SDN and NFV architecture. *Computer Science Review*, 38, article number 100298. doi: 10.1016/j.cosrev.2020.100298.
[15] Lakhani, G., & Kothari, A. (2020). Fault administration by load balancing in distributed SDN controller: A review. *Wireless Personal Communications*, 114(4), 3507-3539. doi: 10.1007/s11277-020-07545-2.
[16] Liang, S., Jiang, W., Zhao, F., & Zhao, F. (2020). Load balancing algorithm of controller based on SDN architecture under machine learning. *Journal of Systems Science and Information*, 8(6), 578-588. doi: 10.21078/JSSI-2020-578-11.
[17] Monir, M.F., & Pan, D. (2021). Exploiting a virtual load balancer with SDN-NFV framework. In *2021 IEEE international Black Sea conference on communications and networking* (pp. 1-6). Bucharest: IEEE. doi: 10.1109/BlackSeaCom52164.2021.9527807.
[18] Moosavi, R., Parsaeefard, S., Maddah-Ali, M.A., Shah-Mansouri, V., Khalaj, B.H., & Bennis, M. (2021). Energy efficiency through joint routing and function placement in different modes of SDN/NFV networks. *Computer Networks*, 200, article number 108492. doi: 10.1016/j.comnet.2021.108492.
[19] Nawaz, H., Ali, M.A., Rai, S.I., & Maqsood, M. (2024). Comparative analysis of cloud based SDN and NFV in 5g Networks. *The Asian Bulletin of Big Data Management*, 4(1), 206-216. doi: 10.62019/abbdm.v4i1.114.
[20] Nezami, Z., Zamanifar, K., Djemame, K., & Pournaras, E. (2021). Decentralized edge-to-cloud load balancing: Service placement for the Internet of Things. *IEEE Access*, 9, 64983-65000. doi: 10.1109/ACCESS.2021.3074962.
[21] Pidpalyi, O. (2024). Future prospects: AI and machine learning in cloud-based SIP trunking. *Bulletin of Cherkasy State Technological University*, 29(1), 24-35. doi: 0.62660/bcstu/1.2024.24.
[22] Ray, P.P., & Kumar, N. (2021). SDN/NFV architectures for edge-cloud oriented IoT: A systematic review. *Computer Communications*, 169, 129-153. doi: 10.1016/j.comcom.2021.01.018.
[23] Rout, S., Patra, S.S., Patel, P., & Sahoo, K.S. (2020). Intelligent load balancing techniques in software defined networks: A systematic review. In *2020 IEEE international symposium on sustainable energy, signal processing and cyber security* (pp. 1-6). Gunupur Odisha: IEEE. doi: 10.1109/iSSSC50941.2020.9358873.
[24] Song, Z., Sun, Y., Wan, J., Huang, L., & Zhu, J. (2019). Smart e-commerce systems: Current status and research challenges. *Electronic Markets*, 29, 221-238. doi: 10.1007/s12525-017-0272-3.
[25] Tache, M.D., Păscuțoiu, O., & Borcoci, E. (2024). Optimization algorithms in SDN: Routing, load balancing, and delay optimization. *Applied Sciences*, 14(14), article number 5967. doi: 10.3390/app14145967.
[26] Thajeel, T.G., & Abdulhassan, A. (2021). A comprehensive survey on software-defined networking load balancers. In *2021 4th international Iraqi conference on engineering technology and their applications* (pp. 1-7). Najaf: IEEE. doi: 10.1109/IICETA51758.2021.9717919.
[27] Tipantuna, C., & Hesselbach, X. (2020). NFV/SDN enabled architecture for efficient adaptive management of renewable and non-renewable energy. *Open Journal of the Communications Society*, 1, 357-380. doi: 10.1109/OJCOMS.2020.2984982.

[28] Zarca, A.M., Bernabe, J.B., Trapero, R., Rivera, D., Villalobos, J., Skarmeta, A., & Gouvas, P. (2019). Security management architecture for NFV/SDN-aware IoT systems. *IEEE Internet of Things Journal*, 6(5), 8005-8020. doi: 10.1109/JIOT.2019.2904123.
[29] Zhou, Q., Yu, J., & Li, D. (2021). A dynamic and lightweight framework to secure source addresses in the SDN-based networks. *Computer Networks*, 193, article number 108075. doi: 10.1016/j.comnet.2021.108075.
[30] Zhu, L., Karim, M.M., Sharif, K., Xu, C., Li, F., Du, X., & Guizani, M. (2020). SDN controllers: A comprehensive analysis and performance evaluation study. *ACM Computing Surveys (CSUR)*, 53(6), article number 133. doi: 10.1145/3421764.

Порівняльний аналіз методів балансування навантаження на основі SDN/NFV

Олександр Берестовенко
Аспірант
Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського»
03056, просп. Берестейський, 37, м. Київ, Україна
https://orcid.org/0000-0003-4887-4674

Анотація. Метою дослідження було визначення переваг та недоліків застосування методів балансування навантаження в мережах на основі технологій Software-Defined Networking та Network Functions Virtualization. Окрему увагу було приділено порівнянню ефективності різних підходів до балансування, зокрема централізованих та дистрибутивних методів, а також використанню інтелектуальних алгоритмів для прогнозування навантаження. Аналіз дозволив визначити переваги та недоліки кожного з методів, а також їхні можливості для адаптації до змінних умов мережевого трафіку, з урахуванням таких параметрів, як пропускна здатність, затримки, втрата пакетів та енергоефективність. У роботі розглянуто методи балансування навантаження в мережах на основі технологій Software-Defined Networking та Network Functions Virtualization, які є важливими для забезпечення ефективності, масштабованості та адаптивності сучасних мереж. Описано ключові виклики, з якими стикаються ці технології, такі як динамічність і непередбачуваність трафіку, оптимізація ресурсів, енергоефективність, а також інтеграція інтелектуальних алгоритмів для прогнозування навантаження і зменшення енергоспоживання. У дослідженні представлено порівняння різних методів балансування навантаження, включаючи централізоване та дистрибутивне управління трафіком, а також використання віртуальних балансувальників і адаптивних алгоритмів переспрямування трафіку. Особливу увагу приділено аналізу впливу цих методів на пропускну здатність, затримки, втрати пакетів і енергоефективність у різних умовах трафіку. Розглянуто роль машинного навчання в оптимізації процесів балансування навантаження, а також можливості інтеграції Software-Defined Networking та Network Functions Virtualization у гібридні мережі. Згідно з результатами дослідження, використання методів балансування на основі Software-Defined Networking/Network Functions Virtualization дозволяє значно підвищити ефективність мереж, зменшити затримки та збільшити пропускну здатність, при цьому знижуючи енергоспоживання в умовах високих навантажень. Виведено ключові результати для України, де інтеграція Software-Defined Networking/Network Functions Virtualization у телекомунікаційну інфраструктуру може стати основою для підвищення якості послуг, оптимізації витрат та забезпечення високого рівня безпеки в умовах цифрової трансформації та модернізації інфраструктури

Ключові слова: віртуалізація функцій мережі; інтелектуальні алгоритми; оптимізація ресурсів; гібридні структури; енергоефективність

ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ ТА КОМП'ЮТЕРНА ІНЖЕНЕРІЯ

Науково-технічний журнал

Том 22, № 1, 2025

Заснований у 2004 р. Виходить 3 рази на рік

Оригінал-макет видання виготовлено
у редакційно-видавничому відділі Вінницького національного технічного університету.

Відповідальний редактор:
В. Белзецька

Підписано до друку 24.04.2025 р. Формат 60*84/8
Умовн. друк. арк. 15,7
Наклад 100 примірників

Адреса видавництва:
Вінницький національний технічний університет
21021, вул. Хмельницьке шосе, 95, м. Вінниця, Україна
тел/факс: +38 (0432) 65-19-03
E-mail: info@itce.com.ua
https://itce.com.ua/uk

INFORMATION TECHNOLOGIES AND COMPUTER ENGINEERING

Scientific and Technical Journal

Vol. 22, No. 1, 2025

Founded in 2004. Published three times per year

The original layout of the publication is made
in the publishing department of Vinnytsia National Technical University

Managing editor:
V. Belzetska

Signed for print 24.04.2025. Format 60*84/8
Conventional printed pages 15.7
Circulation 100 copies

Publishing Address:
Vinnytsia National Technical University
21021, 95 Khmelnytske Shose Str., Vinnytsia, Ukraine
тел/факс: +38 (0432) 65-19-03
E-mail: info@itce.com.ua
https://itce.com.ua/en