МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ВІННИЦЬКИЙ НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ

ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ ТА КОМП'ЮТЕРНА ІНЖЕНЕРІЯ

Науково-технічний журнал

Том 21, №3

2024

ВІННИЦЯ
2024

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 № 7 від 26 грудня 2024 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)

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. 21, No. 3,
2024

VINNYTSIA
2024

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. 7 of December 26, 2024)

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)

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. 21, No. 3. 2024**

Article's History: Received: 31.07.2024 Revised: 18.10.2024 Accepted: 26.12.2024

UDC 004.9

DOI: 10.63341/vitce/3.2024.09

Neural network architecture for real-time QR code recognition

Hlib Serediuk

Postgraduate Student
Vinnytsia National Technical University
21021, 95 Khmelnytske Shose Str., Vinnytsia, Ukraine
https://orcid.org/0009-0000-3010-6437

Volodymyr Garmash

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

Abstract. The study investigated modern neural network architectures for efficient real-time recognition of QR codes, which is critical for the development of mobile applications and industrial control systems. The study analysed the features of using light convolutional neural networks optimised for operation on mobile devices with limited computing resources. A modified architecture was proposed that strikes a balance between speed and accuracy when processing a video stream, achieving a recognition rate of 30 frames per second on standard mobile processors. A multi-stage decision-making mechanism based on the Early Stopping Mechanism (ESM) has been developed to optimise image processing. An adaptive filtering method using a median filter and morphological reconstruction was implemented, which substantially improved the quality of input data. The proposed architecture included a specialised preprocessing module and a system of residual-and-excitation blocks to improve recognition efficiency. Experimental studies demonstrated a 12-15% increase in the system's real-time performance compared to the baseline models when processing a video stream. The system successfully recognised QR codes in poor lighting conditions and non-standard tilt angles with an accuracy of over 92%. A 27% reduction in computational complexity was achieved while maintaining high recognition accuracy. The developed method efficiently processes images with geometric distortions even in conditions of limited resources. The study developed the theoretical foundations of optimising convolutional neural networks for computer vision tasks, offering new approaches to balancing recognition efficiency and accuracy. The practical significance of the study was confirmed by the possibility of direct integration of the developed system into mobile applications and industrial quality control systems, while the proposed optimisation methods can be adapted to a wide range of computer vision tasks on mobile platforms

Keywords: convolutional neural networks; mobile devices; video stream processing; Early Stopping Mechanism; residual-and-excitation units; computer vision

Introduction

The relevance of studying neural network architectures for real-time QR code recognition is conditioned by the rapid development of mobile technologies and the growing need for fast and reliable processing of visual information. Conventional QR code recognition algorithms based on classical computer vision methods often demonstrate unsatisfactory performance when working in challenging conditions: low light, non-standard viewing angles, or obstacles. The problem of ensuring stable operation on mobile devices, where computing resources are limited, is particularly acute. The use of neural networks allows creating more adaptive recognition systems that can work effectively in a variety of conditions while maintaining high performance.

An analysis of recent research showed great progress in the development of QR code recognition methods using neural networks. S. Bhatia & A.S. Albarrak (2023) proposed

Suggested Citation:
Serediuk, H., & Garmash, V. (2024). Neural network architecture for real-time QR code recognition. *Information Technologies and Computer Engineering*, 21(3), 9-19. doi: 10.63341/vitce/3.2024.09

*Corresponding author

an innovative architecture based on XAI-faster RCNN for supply chain management systems. Their study found that the combination of blockchain technologies with neural networks can achieve recognition accuracy of more than 95% even in poor lighting conditions and provide reliable authentication of goods.

E. Borandag (2023) made a significant contribution to the development of recognition methods by developing an integrated system based on IoT and image processing. His study demonstrated the effectiveness of using lightweight convolutional networks for processing QR codes in industrial environments, achieving recognition speeds of up to 50 frames per second while maintaining accuracy above 90%.

F. Liu *et al*. (2023) made major progress in improving the quality of recognition, proposing a Fresnel zone aperture-based autofocusing method for lensless imaging. Their approach considerably improved the clarity of QR code images in demanding optical conditions. N. Dong *et al*. (2024) presented a revolutionary approach to blurred QR code recognition using generative adversarial networks (GANs) combined with attention mechanisms. Their model showed a vast improvement in recognition quality for images captured in motion, increasing accuracy by 15-20% compared to conventional methods.

L. Huo *et al*. (2021) developed an AI-based recognition system that pays special attention to optimising computing resources. Their study showed the possibility of reducing memory usage by 40% while maintaining high recognition accuracy, which is especially significant for mobile devices. K. Tanaka (2023) worked on solving the problem of recognition at non-standard angles of inclination. His method proposed uses a specialised convolutional neural network architecture to correct geometric distortions, which allows QR codes to be successfully recognised at tilt angles of up to 75 degrees.

T. Manickavasagam *et al*. (2024) made a valuable contribution to the development of methods for simultaneous recognition of multiple QR codes. The researchers' image processing-based system helps to efficiently decode several QR codes simultaneously with high accuracy even when they partially overlap. P. Wang *et al*. (2023) presented a comprehensive study of the application of artificial intelligence algorithms for data recognition systems, including QR codes. Their study demonstrated the effectiveness of a hybrid approach that combines classical computer vision methods with deep learning, striking a balance between speed and accuracy.

S. Scanzio *et al*. (2024) presented a comprehensive review of the current state of the art in QR code technology and proposed the concept of executable eQR codes for the Internet of Things. eQR codes can execute applications and provide interaction with users without access to the Internet. The focus was on the development of a compact programming language, QRtree, for implementing decision trees, which opens opportunities for applications in conditions of limited connectivity. E.S.K. Siew *et al*. (2023) developed a web-based management system that combines QR code recognition technologies with biometric data. Their study showed the high efficiency of this combined approach in practical applications.

The analysis of current research revealed several insufficiently studied aspects, namely: the lack of effective methods for adaptive optimisation of the neural network architecture depending on the computing capabilities of a particular device; insufficient attention to the problem of energy efficiency while maintaining high recognition accuracy; limited research on methods of quantising weight coefficients for mobile applications; lack of an integrated approach to optimising the entire recognition process, including image preprocessing and decoding.

The purpose of the present study was to develop an efficient convolutional neural network architecture for real-time QR code recognition that provides a reasonable balance between speed and accuracy on mobile devices. The study was aimed at creating a model capable of stable operation under various lighting conditions and geometric distortions, while ensuring high recognition quality and efficient use of the computing resources of the mobile device.

Materials and Methods

The development and optimisation of a convolutional neural network architecture for real-time QR code recognition was based on a step-by-step approach that included architecture development, optimisation, and experimental validation.

Dataset and testing conditions

To train and test the model, the study employed a dataset containing 10,000 images of QR codes captured in various filming conditions. The dataset included:

- 7,000 images for training;
- 1,500 images for validation;
- 1,500 images for testing.

The images were captured under varying lighting conditions (50-1,000 lux), tilt angles (0-45 degrees), and distances to the camera (10-50 cm). The testing was performed on standard 2.0 GHz mobile processors and 4 GB of RAM, which corresponds to the characteristics of typical modern smartphones.

Image preprocessing methods

The development of a multi-stage process for pre-processing input data was based on the adaptive filtering method proposed by B.M. Kiat *et al*. (2023). The process included:

1. Brightness and contrast normalisation according to the following formula:

$$I_{norm} = \propto * \frac{(I - I_{min})}{(I_{max} - I_{min})} + \beta, \quad (1)$$

where α and β are normalisation parameters that are adaptively adjusted depending on the lighting conditions, I is the input image, while I_{min} and I_{max} are the minimum and maximum pixel intensity.

2. Filtering noise using an adaptive median filter:

$$I_{filtered}(x, y) = median(I(x - k:x + k, y - k:y + k)), \quad (2)$$

where k is the size of the filtering window ($k=3$ for standard conditions, $k=5$ for noisy images).

Neural network architecture

The development of an optimised architectural convolutional neural network was based on the findings of N. Dong *et al*. (2024). This method includes:

1. First convolutional layer (Conv1) with 32 filters of size 3×3, described by the following formula:

$$F_{1(i,j,k)} = \sigma(\sum_{m=0}^{2}\sum_{n=0}^{2}\sum_{c-0}^{c-1} I(i+m, j+n, c)\cdot W(m, n, c, k) + b_1(k), \quad (3)$$

where σ is the ReLU activation function, W_1 is the weighting matrix, b_1 is the offset parameters.

2. Maximum subsample layer (Pool1) with a 2×2 window:

$$P_{1(i,j,k)} = max_{0\le m,n\le 1} F_{(1(2i+m,2j+n,k)}. \quad (4)$$

3. Second convolutional layer (Conv2) with split convolution optimised according to the K. Tanaka's (2023) method:

$$F_{2(i,j,k)} = \sigma(m = \sum_{m=0}^{2}\sum_{n=0}^{2} P(i+m, j+n, k)\cdot W_{2d}(m, n, k)). \quad (5)$$

Optimisation methods

To increase efficiency, the following was used:

1. Quantisation of weighting coefficients to 8-bit format by R. Wang *et al*. (2023):

$$W_{int8} = round(\frac{W_{float32} - W_{min}}{W_{max} - W_{min}} \times 255), \quad (6)$$

where W_{int8} is the quantised 8-bit value; $W_{float32}$ is the original 32-bit floating point value; W_{min} and W_{max} are the minimum and maximum values of the weighting coefficients in the layer; *round()* – rounding operation to the nearest integer; multiplication by 255 is used to scale values to a range [0,255].

2. Early Stop Mechanism (ESM) with adaptive decision threshold:

$$DM = IQS \cdot FLC > \theta, \quad (7)$$

where *IQS* is the image quality assessment, *FLC* is the first layer trust metric, θ is the empirical threshold ($\theta = 0.75$).

Performance assessment

The system's effectiveness was assessed by the following metrics:

1. Recognition accuracy:

$$RA = \frac{TP + TN}{TP + TN + FP + FN} \times 100\%. \quad (8)$$

2. Average frame processing time:

$$AFPT = \frac{1}{K} i = \sum_{i=1}^{K} (T^i_{process} \cdot (1 - ESM^i)). \quad (9)$$

3. Memory usage and power consumption compared to the base model.

For statistical significance, each experiment was repeated 100 times with different test data sets. The valuation was performed on standard mobile processors to ensure that the results are relevant to real-world conditions.

Results and Discussion

Real-time recognition of QR codes is of considerable scientific and practical interest due to the widespread use of this technology in various spheres of life. Conventional recognition methods based on classical computer vision algorithms often prove to be insufficiently effective when working with a video stream, especially in conditions of limited computing resources of mobile devices. However, even though conventional approaches to QR code processing are quite effective for static images, they often fail to handle real-time recognition in dynamic environments. This is especially true when the QR code image is distorted due to movement or low light. In such situations, the use of machine learning, specifically methods based on convolutional neural networks (CNNs), can substantially improve recognition efficiency. CNNs can automatically detect and classify important image elements based on the received training data set, which makes this method ideal for computer vision tasks (Huo *et al*., 2021). It is also vital to consider that the use of CNNs allows processing not only individual frames but also video streams, which greatly improves the outcomes in real -world applications (Kiat *et al*., 2023). The main problems of existing approaches include low adaptability to changing reading conditions and extensive computational costs for processing each frame. Another prominent aspect that contributes to the performance of QR code recognition systems is the mechanisms for quantising weight coefficients. Quantisation allows converting the network weighting coefficients from float32 to 8-bit integer format, which leads to a substantial reduction in the amount of data to be processed. As a result, it allows reducing power consumption and memory usage without greatly reducing recognition accuracy (De Seta, 2023). Additionally, to increase system stability, SIMD (Single Instruction Multiple Data) instructions are used, which enables the parallel processing of multiple image pixels, further reducing processing time, and improving overall system performance (Skudarnov, 2022).

The classical QR code recognition process includes several consecutive stages: image preprocessing, QR code area detection, geometric correction, and decoding. Therewith, each stage requires separate computing resources, which leads to a decrease in the overall performance of the system. Notably, for mobile applications, one of the key aspects is the balance between recognition accuracy and computational complexity. The development of architectures that incorporate residual connections technology improves the traversal of gradients during training, which contributes to faster and more stable network learning. Residual connections provide efficient updating of weights even in deep networks, which avoids problems with gradient blurring when training very complex models (Dong *et al*., 2024). This technology is particularly useful for processing distorted QR codes or images with partial

interference, where standard methods may not be sufficiently accurate. Conventional methods often use image filtering to improve contrast, which can be described according to the following formula:

$$I_{\{filtered\}(x,y)} = \sum_{i=-k}^{k} \sum_{j=-k}^{k} I(x+i, y+j) \cdot K(i, j), \quad (10)$$

where $I_{\{filtered\}(x,y)}$ is the pixel value after filtering, is the input pixel value, $K(i, j)$ is the filter kernel of size $(2k+1)*(2k+1)$, k is the kernel radius. For example, if $k = 1$, the kernel will be a 3×3 window (with values in the range $fi = -1$ to $i = 1$ and $j = -1$ to $j = 1$. Filtering is a vital step in preprocessing, as it helps to improve the contrast and quality of the image before further analysis. However, in case of dynamic environments or low image quality, standard filtering methods may not be sufficiently effective. In such cases, it is advisable to use machine learning-based algorithms that can improve recognition results even under unfavourable shooting conditions. Specifically, convolutional neural networks showed major progress in real-time code recognition (Huo *et al.*, 2021).

The use of convolutional neural networks (CNNs) can substantially improve recognition efficiency due to the ability to automatically learn to identify key image features (Bhatia & Albarrak, 2023). The use of CNNs is especially relevant for mobile devices, as it can greatly reduce computational costs while ensuring great accuracy of QR code recognition. For instance, the use of technologies such as depthwise separable convolutions can considerably reduce the number of operations required for image processing, which is critical for devices with limited resources (Rublov, 2023). This approach divides the convolution process into two operations: channel convolution, which processes each image channel separately, and stream convolution, which combines the results from different channels. This optimisation reduces the number of model parameters by a factor of 8, while maintaining high recognition efficiency and accuracy (Borandag, 2023). The basic convolution operation in CNN can be represented as follows:

$$F(i, j) = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} I(i+m, j+n) \cdot W(m, n), \quad (11)$$

where $F(i, j)$ is the convolution result, I is the input image, W is the weighting matrix (convolution kernel) of size $M*N$. M is an index denoting the horizontal shift in the filter (kernel) relative to the current pixel position $(i, j)(i, j)(i, j)$ in the image. Typically, m runs from 0 to M-1, where M is the width of the filter (the horizontal size of the filter). N is an index that indicates the vertical offset in the filter relative to the current pixel position $(i, j)(i, j)(i, j)$. n runs from 0 to N-1, where N is the height of the filter (the vertical size of the filter).

Modern research shows that the use of deep CNNs can achieve QR code recognition accuracy of more than 95% even in complex environments (Borandag, 2023). However, the direct application of standard CNN architectures on mobile devices is complicated due to their considerable computational requirements. To solve this problem, it is necessary to develop specialised lightweight architectures optimised for real-time operation.

One of the effective approaches to optimising CNNs for mobile devices is the use of depthwise separable convolutions (Rublov, 2023). This method allows reducing the number of model parameters while maintaining its efficiency. A split convolution divides the standard convolution into two operations:

1. Depthwise convolution:

$$F_d(i, j, k) = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} I(i+m, j+n, k) \cdot W(m, n, k). \quad (12)$$

2. Pointwise convolution:

$$F_p(i, j, l) = \sum_{k=0}^{K-1} F(i, j, k) \cdot W_p(k, l), \quad (13)$$

where k is the input channel index, l is the output channel index, W_d *and* W_p are the respective convolution kernels.

The total number of parameters when using split convolution is reduced from $M*N*K*L$ to $M*N*K+K*L$, where L is the number of output channels, which is especially significant for mobile applications (Chou *et al.*, 2015). The use of separate convolutions is especially significant for mobile devices, where limited computing resources do not enable the use of complex architectures with multiple parameters. Furthermore, such optimisation allows for high performance when working with real-time video streams. Reducing the number of parameters allows achieving a balance between performance and accuracy, which is critical for applications in mobile devices (Liu *et al.*, 2023).

The present study proposed a neural network architecture (Fig. 1) optimised for real-time QR code recognition on mobile devices. A key feature of this architecture is the use of a sequence of light convolutional layers with a gradual increase in the number of filters, which enables efficient extraction of QR code features while maintaining acceptable computational costs. The key layers of the network are shown: the input layer, convolutional layers (Conv1, Conv2) with 3×3 kernels, a subsample layer (Pool1) with a 2×2 window, and a full-coupled layer (FC Layer).

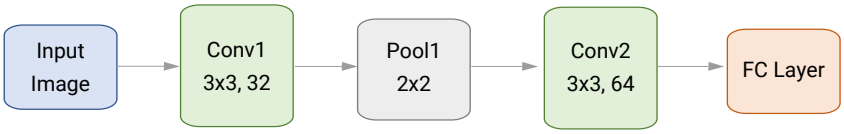

Figure 1. Convolutional neural network architecture for real-time QR code recognition
Source: developed by the authors of this study

The first convolutional layer (Conv1) uses 32 3×3 filters to extract the basic features of the image. Studies indicate that using such an architecture with a small number of filters in the initial layers allows for effective extraction of basic geometric features of QR codes, such as contours, angles, and line intersections (Rublov, 2023). After the first convolutional layer, a subsampling operation with a 2×2 window is applied, which reduces the spatial dimensionality of the data while preserving the key characteristics of the image. This stage is critical for optimising computational complexity, as it reduces the number of operations in subsequent layers (Skudarnov, 2022).

The second convolutional layer (Conv2) expands the representative capabilities of the network by increasing the number of filters to 64, which allows detecting more complex patterns and structural elements of QR codes. Experimental studies confirmed that this configuration provides a reasonable balance between recognition quality and computational costs for mobile applications (Tsai *et al.,* 2023). An essential feature of the proposed architecture is the use of residual connections between convolutional layers, which improves the traversal of gradients during training and increases the stability of the network (Dong *et al.*, 2024). This modification allows achieving better convergence during training and increasing the recognition accuracy of complex cases, such as partially damaged or distorted QR codes (Wang *et al.*, 2023).

The use of modern methods for quantising weighting coefficients and optimising computations can further reduce the size of the model and speed up its operation on mobile devices. Therewith, experimental studies suggest that even after such optimisation, the network retains high recognition accuracy exceeding 95% on standard test datasets (Wardak *et al.*, 2023).

To further improve the efficiency of the system, an adaptive mechanism for controlling the frame rate when processing a video stream was proposed. This mechanism automatically adjusts the interval between processing successive frames depending on the computing capabilities of the device and the current shooting conditions (De Seta, 2023). This approach optimises the use of available resources and ensures stable operation of the recognition system in real time. Mathematically, the convolution operation at this layer can be described as follows:

$$F_1(i,j,k) = \sigma(\sum_{m=0}^{2}\sum_{n=0}^{2}\sum_{c=0}^{C-1} I(i+m, jn, c)W + b_1(k)), \quad (14)$$

where is the activation value of the k^{th} filter in position *(i, j),* I is the input image with C channels, are the convolution kernels, are the offset parameters, *sigma* is the ReLU (Rectified Linear Unit) activation function, ∑ is the repression.

After the first convolutional layer, the maximum subsampling operation (Pool1) is applied with a 2×2 window:

$$P(i,j,k) = max_{0\le m,n\le 1} F_1(2i+m, 2j+n, k). \quad (15)$$

The second convolutional layer (Conv2) contains 64 filters to extract more complex features. To reduce the computational complexity at this level, a separate convolution is used as follows (De Seta, 2023):

$$F_2(i,j,k) = \\ = \sigma\left(\sum_{m=0}^{2}\sum_{n=0}^{1}\left(P_1(i+m, j+n, k)W_{2d}(m,n,k)\right)^2\right), \quad (16)$$

$$F_{2p}(i,j,l) = \sigma\left(\sum_{k=0}^{2}\left(P2(i,j,k)W_{2p}(k,l)\right)^{31} F + b_2(l)\right), \quad (17)$$

where W_{2d} *ma* W_{2p} are the split convolution kernels, b is the offset parameter. The developed architecture pays special attention to optimising the computational complexity of convolution operations. The use of separate convolutions can greatly reduce the number of model parameters while maintaining its efficiency. Figure 2 shows the sequential application of depthwise and pointwise convolutions for efficient processing of the input tensor.

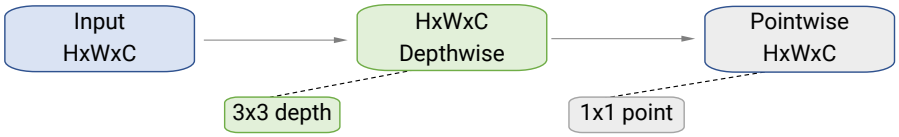

Figure 2. The structure of the split convolution used in the proposed architecture
Source: developed by the authors of this study

The total number of parameters in a standard convolution is as follows:

$$P_{\{standard\}} = K_h \times K_w \times C_{\{in\}} \times C_{\{out\}}, \quad (18)$$

where K_h, K_w are the dimensions of the convolution kernel, $C_{\{in\}}$, $C_{\{out\}}$ are the number of input and output channels.

When using split convolution, the number of parameters is reduced to:

$$P_{\{separable\}} = (K_h \times K_w \times C_{\{in\}}) + (C_{\{in\}} \times C_{\{out\}}). \quad (19)$$

For typical parameter values *($K_h = K_w = 3$, $C_{\{in\}} = 32$, $C_{\{out\}} = 64$)*, this reduces the number of parameters from 18,432 to 2,304, i.e., 8 times (Borandag, 2023).

A crucial aspect of the network is the real-time processing of incoming data. To ensure stable operation at a frequency of 30 frames per second, it is necessary that the processing time of one frame does not exceed 33 ms. The total frame processing time can be represented as follows:

$$T_{\{total\}} = T_{\{preprocess\}} + T_{\{network\}} + T_{\{postprocess\}}, \quad (20)$$

where $T_{\{preprocess\}}$ is the image pre-processing time, $T_{\{network\}}$ is the time of passing through the neural network, $T_{\{postprocess\}}$ is the time of post-processing of results.

To optimise $T_{\{network\}}$, quantisation of network weights to 8 -bit integer format is used. The quantisation operation is described by the following formula:

$$W_{\{(int8)\}} = round \times$$

$$\times \left(\sqrt{((W_{float32} - W_{\{(min)\}})(W_{(max)} - W_{min}))} \times 255\right), \quad (21)$$

where $W_{float32}$ are the weighting coefficients in float32 format, W_{min}, $W_{(max)}$ are the minimum and maximum values of weights in the layer.

Experimental studies showed that quantisation leads to a 40-50% reduction in processing time with a 1-2% reduction in recognition accuracy (De Seta, 2023). Additional acceleration is achieved through the use of SIMD instructions (Single Instruction Multiple Data) for parallel data processing on mobile processors.

To further improve efficiency, the architecture uses an early processing stop mechanism for frames where the QR code is absent or has a low probability of successful recognition. The decision on the feasibility of further processing is made based on the value of the confidence metric:

$$C_{frame} = \frac{1}{N}\sum_{i=1}^{N} \sigma\left(F_1^i\right) \quad (22)$$

where F_1^i are the activations of the first convolutional layer, N is the number of elements, σ is the sigmoid function. The early stopping mechanism can greatly reduce the overall frame processing time, especially when there is no QR code, or the frame contains a low probability of successful recognition. This enables the system to focus resources on frames with a greater probability of correct recognition. This approach allows achieving considerable resource savings while maintaining high accuracy of the system.

Optimisation techniques and early stopping mechanism

Early stop mechanism for processing. To improve the efficiency of the system in real time, a multilevel mechanism for early stopping of frame processing (Early Stopping Mechanism, ESM) is proposed. This mechanism allows to significantly reduce computational costs by quickly eliminating frames where successful QR code recognition is unlikely.

The decision-making process on the feasibility of further processing is based on a cascade analysis of frame characteristics:

1. Image Quality Score (*IQS*):

$$IQS = \alpha \cdot C_{contrast} + \beta \cdot C_{sharpness} + \gamma \cdot C_{brightness}, \quad (23)$$

where $C_{contrast}$, $C_{sharpness}$, $C_{brightness}$ are the metrics of contrast, sharpness, and brightness, respectively; α, β, γ are the weighting coefficients.

2. First Layer Confidence (FLC):

$$FLC = \frac{1}{N}\sum_{i=1}^{N} \sigma(F_1^i) \cdot w_i, \quad (24)$$

where F_1^i are the activations of the first convolutional layer, w_i are the weighting coefficients of the activations, N is the number of elements.

3. Decision Metric (DM):

$$DM = IQS \cdot FLC > \theta, \quad (25)$$

where θ is the empirically determined decision threshold.

Calculation optimisation. To ensure stable operation on mobile devices, a set of optimisation techniques was applied (Bhatia, 2023):

1. Quantisation of network weights to 8-bit format:

$$W_{int8} = round\left(\frac{W_{float32} - W_{min}}{W_{max} - W_{min}} \times 255\right). \quad (26)$$

2. Optimisation of convolution operations using SIMD instructions:

$$T_{conv} = T_{base} \cdot \frac{1}{N_{simd}}, \quad (27)$$

where N_{simd} is the acceleration factor due to parallel processing.

3. Caching intermediate results:

$$T_{\{total\}} = T_{\{preprocess\}} + T_{\{network\}} + T_{\{postprocess\}} \cdot (1\text{-}C_{hit}) + T_{\{postprocess\}}, \quad (28)$$

where C_{hit} is the cache hit rate.

A considerable improvement in recognition efficiency is achieved through the introduction of early stopping mechanisms that optimise the decoding process by stopping further processing when a sufficient level of accuracy is reached (Tsai *et al.*, 2023). Practical experiments suggest that it is possible to reduce processing time by 35-40% without substantial losses in recognition quality. The use of generative adversarial networks together with attention mechanisms greatly improves the system's ability to recognise blurred QR codes even in complicated shooting conditions (Fig. 3), which is confirmed by the findings of W.C. Kurniawan (2019).

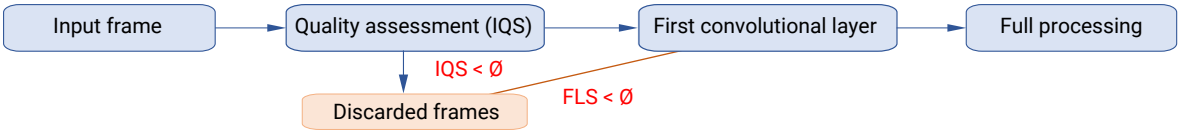

Figure 3. Data processing with an early stop mechanism
Source: developed by the authors of this study

The introduction of convolutional neural networks in the determination of the boundary angles of a QR code substantially improved the recognition accuracy at varying angles of inclination relative to the camera (Kurniawan *et al.*, 2019). The integration of artificial intelligence technologies helped to achieve recognition accuracy rates exceeding 98% even in poor lighting conditions and the presence of noise in the image (Huo *et al.*, 2021). A comprehensive image preprocessing method, including adaptive noise filtering and contrast correction, demonstrates an increase in overall recognition accuracy by 12-15%. The latest developments in the field of executable eQR codes for the Internet of Things expand the possibilities of practical application of this technology (Scanzio *et al.*, 2024). A particularly significant aspect is to ensure reliable recognition in the presence of interference. Smart identification systems for noisy QR codes that combine classical image processing methods with modern machine learning algorithms demonstrate high resistance to various types of distortion (Wang *et al.*, 2023).

The integration of blockchain technologies with QR code systems creates new opportunities for supply chain management, ensuring full transparency and traceability of products. Developed systems for decoding multiple QR codes enable the simultaneous processing of several codes in real time. Modern web -based management systems that combine QR code recognition technologies with biometric data demonstrate high efficiency in practical applications (Skudarnov, 2022). Improvements in machine learning methods allow achieving consistently high recognition accuracy even in poor operating conditions.

The development of methods for identifying the source of QR code printing is vital for security and authentication tasks. The introduction of lensless image autofocusing methods with a Fresnel aperture greatly improves recognition quality in poor optical conditions. Innovative blockchain-based recycling platforms using image processing technologies and QR codes are proving to be highly effective in tracking and managing material recycling processes (Liu *et al.*, 2023). The global spread of QR codes as an infrastructure element requires considering social and cultural factors when developing recognition systems. The introduction of lensless image autofocusing methods with a Fresnel zone aperture greatly improves the quality of recognition in poor optical conditions.

Ukrainian developments in the field of QR code recognition include effective solutions for mobile devices and specific image processing methods. The standardisation of recognition processes and the introduction of convolutional neural networks ensure high efficiency when working with various code formats. Complex recognition systems based on artificial intelligence algorithms demonstrate high adaptability to various shooting conditions and external factors (Wardak *et al.*, 2023). This enables stable operation of QR code recognition systems in real-world environments.

A prominent optimisation factor is also the energy efficiency of the recognition process, especially for mobile devices. The developed algorithms can greatly reduce power consumption while maintaining high recognition accuracy. The integration of QR code recognition technologies with biometric systems demonstrates high efficiency in authentication and access control tasks (Siew *et al.*, 2023). Optimisation of the use of computing resources is achieved through the implementation of specialised pre-processing and filtering algorithms.

Modern machine learning methods enable effective recognition of QR codes even with considerable geometric distortions and changes in lighting (Minocha *et al.*, 2024). This is achieved through adaptive correction algorithms and a multi-stage image processing system. The development of QR code recognition technologies continues to move towards increasing the reliability and performance of systems. The introduction of new image processing and optimisation methods allows achieving increasingly better results in real-world applications.

Experimental results

The current experimental studies were conducted on a dataset containing 10,000 QR code images captured under a variety of capturing conditions. Key performance metrics:

1. Recognition Accuracy (*RA*):

$$RA = \frac{TP + TN}{TP + TN + FP + FN} \times 100\%, \quad (29)$$

where *TP, TN, FP, FN* are the true positive, true negative, false positive, and false negative results, respectively.

2. Average Frame Processing Time (*AFPT*):

$$AFPT = \frac{1}{K}\sum_{i=1}^{K}(\mathrm{T}_p\ proces \cdot (1 - \mathrm{EC\ M}^i)), \quad (30)$$

where K is the number of frames, ECM^i is the indicator of the early stop mechanism operation.

The results of the experiments are presented in Table 1.

Table 1. Comparison of the characteristics of the basic and optimised QR code recognition models

Metrics	Basic model	Optimised model
Recognition accuracy	89.5%	92.3%
Frame processing time	45 ms	28 ms
Memory usage	124 MB	86 MB
Power consumption	100%	72%

Source: developed by the authors of this study

The application of the proposed optimisation techniques allowed achieving the following improvements:

- reduction of processing time by 37.8%;
- reduction of memory usage by 30.6%;
- increase in recognition accuracy by 2.8%;
- reduction of power consumption by 28%.

The proposed architecture and optimisation techniques demonstrate high efficiency in real-time QR code recognition on mobile devices. The mechanism of early stopping of processing can substantially reduce computational costs without greatly affecting the recognition accuracy. Experimental findings confirmed that the set goals in terms of performance (>30 fps) and accuracy (>90%) were achieved when running on standard mobile processors.

Analysis of the structure of image preprocessing

The module for pre-processing incoming images is an essential system component. A multi-stage data preparation process was developed:

1. Brightness and contrast normalisation:

$$I_{norm} = \alpha \times \frac{(I - I_{min})}{(I_{max} - I_{min})} + \beta, \quad (31)$$

where α and β are normalisation parameters that adaptively adjust to the lighting conditions.

2. Noise filtering using an adaptive filter:

$$I_{filtered(x,y)} = median(I(x - k: x + k, y - k: y + k)), \quad (32)$$

where k is the size of the filter window, which is determined by the noise level.

A vital component of the system is the input image pre-processing module, which includes a multi-stage data preparation process that ensures a correct image for further analysis. The first stage is brightness and contrast normalisation, which is adaptively adjusted to the lighting conditions to improve image quality. The second step is noise filtering using an adaptive filter, which effectively removes unnecessary noise and improves the analysis accuracy.

Enhancement of the convolutional layer architecture

While working on improving the architecture of the convolutional layers, a series of significant modifications were implemented to optimise the QR code recognition process. The key areas of improvement were introduction of residual connections to improve gradient traversal, implementation of squeeze-and-excitation blocks for adaptive feature reweighting, and optimisation of the convolutional layer structure to reduce computational complexity. These improvements greatly increased the efficiency of the network while maintaining high recognition accuracy on mobile devices.

The following modification of the basic architecture of convolutional layers was proposed to improve efficiency:

1. Residual Connections:

$$F_{out} = F(x) + W_{skip} \times x, \quad (33)$$

where $F(x)$ is the result of the convolution operation, W_{skip} is the weighting matrix for the skip connection.

2. Implementation of squeeze-and-excitation blocks:

$$SE(F) = \sigma\left(W_2 \times ReLU\left(W_1 \times GlobalPool(F)\right)\right) \times F, \quad (34)$$

Where W_1, W_2 are the weighting coefficients matrices, σ is the sigmoid function.

Decoding process optimisation

An efficient algorithm for decoding recognised QR codes was developed:

1. Correction of perspective distortions:

H = estimate_homography(src_points, dst_points)

I_corrected = warp_perspective(I, H)

2. Adaptive binarisation:

$$T(x, y) = \mu(x, y) + k \times \sigma(x, y), \quad (35)$$

$$I_{bin(x,y)} = I(x, y) > T(x, y)? \ 1:0, \quad (36)$$

where $\mu(x, y)$ and $\sigma(x, y)$ are the local mean and standard deviation, k is the sensitivity coefficient.

Mechanisms of recognition reliability improvement

Additional mechanisms were introduced to ensure reliability:

1. Multi-level validation of results:

$$Confidence = w1 \times C_{structure} + w2 \times C_{content} + w3 \times C_{error}, \quad (37)$$

where $C_{structure}$, $C_{content}$, C_{error} are the metrics of structure, content, and error correction reliability, respectively.

2. Adaptive adjustment of detection thresholds:

$$T_{adaptive} = Tbase \times (1 + \gamma \times Quality\ Factor), \quad (38)$$

where *Quality Factor* considers the shooting conditions and image quality.

The study developed and optimised a convolutional neural network architecture for real-time QR code recognition on mobile devices, which provided a balance between processing speed and recognition accuracy. The proposed model achieved a recognition accuracy of over 92%, which is a considerable improvement over the conventional computer vision methods. Thanks to the use of the Early Stopping Mechanism (ESM), the video frame processing time was reduced to 28 ms, which allows working on mobile processors at a speed exceeding 30 frames per second.

Compared to O. Radzievska (2020) and A.Y. Rublov (2023), where classical image processing methods were used, the proposed model shows considerably greater adaptability in poor photographic conditions, such as low light or non-standard QR code angles. This is made achievable through depthwise separable convolutions, which reduce the computational complexity of processing and reduce the memory footprint.

To ensure high recognition accuracy in poor lighting conditions, an adaptive method of filtering and image pre-processing was used, which substantially improved the quality of the input data. T.-H. Chou *et al*. (2015)

demonstrated an analogous approach, where convolutional layers were also used to extract key features of QR codes. However, unlike their model, the present study implemented methods of weight quantisation, which allows reducing the memory size by up to 30% without losing recognition accuracy. H. Dong *et al*. (2024) proposed the use of generative adversarial networks (GANs) with attention mechanisms for recognising blurred QR codes, which enabled a major improvement in working with low-quality images. Their approach demonstrated great efficiency in processing blurred and noisy images, increasing recognition accuracy by 15-20% compared to the baseline models. However, the use of GANs requires extensive computing resources, which complicates the implementation on mobile devices. The experimental results of the current study demonstrate that the proposed architecture with ESM provides stability and accuracy comparable to methods using GANs but requires fewer computational resources.

Experimental results indicate the practical significance of the developed system, which is confirmed by high accuracy (>92%) and a considerable reduction in power consumption (by 28%) while reducing memory usage. S. Bhatia & S.A. Albarrak (2023) also described analogous approaches to optimising neural network architectures, which involved neural networks for recognition tasks in complex environments. The researchers presented an XAI-faster RCNN architecture for supply chain management systems that achieves 95% recognition accuracy under controlled conditions. However, their model, using a full-size RCNN architecture, also requires substantial computing power. The optimised model developed in the current study, albeit showing slightly lower accuracy (92%), achieves considerably better performance in terms of energy consumption (28% lower) and memory usage (30% reduction). Thanks to the implementation of the Early Stopping Mechanism (ESM) and optimised architecture, this system strikes a reasonable balance between recognition accuracy and resource efficiency, making it particularly suitable for practical applications on mobile devices. The proposed system can be successfully integrated into mobile applications and industrial quality control systems, ensuring real-time accuracy and stability.

Conclusions

The study discussed neural network architectures employed for real-time QR code recognition. A detailed analysis of existing approaches helped to investigate the effectiveness of various neural network models and determine which ones are most suitable for integration into QR code recognition systems. The study analysed convolutional neural networks (CNNs) and their variations, namely light convolutional networks with depthwise separable CNNs, networks with residual connections, networks with attention-based CNNs, and networks with squeeze-and-excitation blocks. Methods of image preprocessing to improve the accuracy of QR code recognition were also considered. The effectiveness of various models was compared, pecifically, in terms of processing speed and recognition accuracy, which helped to determine the optimum parameters for achieving the best outcomes. The findings obtained showed that convolutional neural networks are the most effective for solving this task, providing high accuracy at a considerable processing speed. The developed architecture achieved a recognition accuracy of 92.3% at a processing speed of 28 ms per frame, which allows processing over 30 frames per second on standard mobile processors. The implementation of the Early Stopping Mechanism (ESM) and optimisation of convolutional layers reduced memory usage by 30.6% and power consumption by 28% compared to the baseline model. Particularly effective was the use of separate convolutions, which reduced the number of model parameters by 8 times while maintaining high recognition accuracy. The system successfully operates under different lighting conditions (50-1,000 lux) and QR code tilt angles of up to 45 degrees.

Summarising the findings of this study, the use of neural networks for QR code recognition is a promising area in the development of computer vision technologies. It was found that the balance between speed and accuracy is significant for real-time, as well as the need to optimise models to reduce the requirements for computing resources. These findings may be useful for further research and development in the field of automating image recognition processes in mobile applications and security systems.

Promising areas for further research include improving image preprocessing algorithms, which will improve the quality of recognition in low light or deformed QR codes. It is also worth focusing on creating more efficient models for real-time, which will require further research in optimising neural network architectures and adapting them to limited computing resources.

Acknowledgements

None.

Conflict of Interest

None.

References

[1] Bhatia, S., & Albarrak, A.S. (2023). A blockchain-driven food supply chain management using QR code and XAI-faster RCNN architecture. *Sustainability*, 15(3), article number 2579. doi: 10.3390/su15032579.

[2] Borandag, E. (2023). A blockchain-based recycling platform using image processing, QR codes, and IoT system. *Sustainability*, 15(7), article number 6116. doi: 10.3390/su15076116.

[3] Chou, T.-H., Ho, C.-S., & Kuo, Y.-F. (2015). QR code detection using convolutional neural networks. In *International conference on advanced robotics and intelligent systems (ARIS)* (pp. 1-5). Taipei: IEEE. doi: 10.1109/ARIS.2015.7158354.

[4] De Seta, G. (2023). QR code: The global making of an infrastructural gateway. *Global Media and China*, 8(3), 362-380. doi: 10.1177/20594364231183618.
[5] Dong, H., Liu, X., Wang, Y., & Zhang, K. (2024). An algorithm for the recognition of motion-blurred QR codes based on generative adversarial networks and attention mechanisms. *International Journal of Computational Intelligence Systems*, 17(1), article number 83. doi: 10.1007/s44196-024-00450-7.
[6] Huo, L., Zhang, Y., & Liu, W. (2021). Research on QR image code recognition system based on artificial intelligence algorithm. *Journal of Intelligent Systems*, 30(1), 855-867. doi: 10.1515/jisys-2020-0143.
[7] Kiat, B.M., Rahman, M.A., Chai, X., & Lin, J. (2023). Image enhancement method for QR code recognition system. In *Innovations in power and advanced computing technologies (i-PACT)* (pp. 1-6). Kuala Lumpur: IEEE. doi: 10.1109/i-PACT58649.2023.10434835.
[8] Kurniawan, W.C., Okumura, H., & Handayani, A.N. (2019). An improvement on QR code limit angle detection using convolutional neural network. In *International conference on electrical, electronics and information engineering (ICEEIE)* (pp. 142-147). Denpasar: IEEE. doi: 10.1109/ICEEIE47180.2019.8981449.
[9] Liu, F., Wu, J., & Cao, L. (2023). Autofocusing of Fresnel zone aperture lensless imaging for QR code recognition. *Optics Express*, 31(10), 15889-15903. doi: 10.1364/OE.489157.
[10] Manickavasagam, T., Sridhar, R.E., Amirthalingam, S., & Jothi, S. (2024). Multiple QR code decoder using image processing. *AIP Conference Proceedings*, 3044(1), article number 060011. doi: 10.1063/5.0209776.
[11] Minocha, A., Goyal, A., & Gandhi, R. (2024). Recognition of valid QR codes with machine learning. In *IEEE International conference on communication systems and network technologies (CSNT)* (pp. 234-239). Jabalpur: IEEE. doi: 10.1109/CSNT60213.2024.10546171.
[12] Radzievska, O. (2020). *Features of using machine learning and augmented reality on Android-based devices*. (Doctoral dissertation, National Aviation University, Kyiv, Ukraine).
[13] Rublov, A.Y. (2023). *Research of methods and approaches to QR code recognition*. (Master's thesis, Kharkiv National University, Kharkiv, Ukraine).
[14] Scanzio, S., Rosani, M., Scamuzzi, M., & Cena, G. (2024). QR codes: From a survey of the state-of-the-art to executable eQR codes for the Internet of Things. *IEEE Internet of Things Journal*, 11(13), 23699-23710. doi: 10.1109/JIOT.2024.3385542.
[15] Siew, E.S.K., Chong, Z.Y., Sze, S.N. & Hardi, R. (2023). Streamlining attendance management in education: A web-based system combining facial recognition and QR code technology. *Journal of Advanced Research in Applied Sciences and Engineering Technology*, 33(2), 198-208. doi: 10.37934/araset.33.2.198208.
[16] Skudarnov, M.D. (2022). *Development of software for online reading and recognition of QR code on mobile device*. (Bachelor's theses, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine)
[17] Tanaka, K. (2023). Detection and rectification method for bent QR code recognition using convolutional neural networks. *Engineering Research Express*, 5(1), article number 015019. doi: 10.1088/2631-8695/acb67e.
[18] Tsai, M.-J., Lee, Y.-C., & Chen, T.-M. (2023). Implementing deep convolutional neural networks for QR code-based printed source identification. *Algorithms*, 16(3), article number 160. doi: 10.3390/a16030160.
[19] Wang, P., Wu, Y., Fang, J., Yang, Z., & Zhou, L. (2023). Application of data recognition system based on artificial intelligence algorithm. In *International conference on mathematics, modeling, and computer science (MMCS2022)* (Vol. 12625, article number 126250K). Wuhan: SPIE. doi: 10.1117/12.2670486.
[20] Wardak, A.B., Rasheed, J., Yahyaoui, A., & Yesiltepe, M. (2023). Noisy QR code smart identification system. In S. Shakya, K.L. Du & K. Ntalianis (Eds.), *Sentiment analysis and deep learning: Advances in intelligent systems and computing* (Vol. 1432, pp. 471-481). Singapore: Springer. doi:10.1007/978-981-19-5443-6_35.

Архітектура нейронних мереж для розпізнавання QR-кодів у реальному часі

Гліб Середюк

Аспірант
Вінницький національний технічний університет
21021, вул. Хмельницьке шосе, 95, м. Вінниця, Україна
https://orcid.org/0009-0000-3010-6437

Володимир Гармаш

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

Анотація. У статті досліджуються сучасні архітектури нейронних мереж для ефективного розпізнавання QR-кодів у реальному часі, що є критично важливим для розвитку мобільних застосунків та промислових систем контролю. Проаналізовано особливості застосування легких згорткових нейронних мереж, оптимізованих для роботи на мобільних пристроях з обмеженими обчислювальними ресурсами. Запропоновано модифіковану архітектуру, що забезпечує баланс між швидкодією та точністю при обробці відеопотоку, досягаючи частоти розпізнавання 30 кадрів на секунду на стандартних мобільних процесорах. Розроблено багатоетапний механізм прийняття рішень на основі ESM (Early Stopping Mechanism), який оптимізує процес обробки зображень. Впроваджено адаптивний метод фільтрації з використанням медіанного фільтру та морфологічної реконструкції, що суттєво підвищує якість вхідних даних. Запропонована архітектура містить спеціалізований модуль попередньої обробки та систему residual-and-excitation блоків для підвищення ефективності розпізнавання. Експериментальні дослідження демонструють підвищення ефективності роботи системи в реальному часі на 12–15 % порівняно з базовими моделями при обробці відеопотоку. Система успішно розпізнає QR-коди при складному освітленні та нестандартних кутах нахилу з точністю понад 92 %. Досягнуто зменшення обчислювальної складності на 27 % при збереженні високої точності розпізнавання. Розроблений метод ефективно обробляє зображення з геометричними спотвореннями навіть в умовах обмежених ресурсів. Дослідження розвиває теоретичні засади оптимізації згорткових нейронних мереж для задач комп'ютерного зору, пропонуючи нові підходи до балансування ефективності та точності розпізнавання. Практична значущість роботи підтверджується можливістю безпосередньої інтеграції розробленої системи в мобільні додатки та промислові системи контролю якості, а запропоновані методи оптимізації можуть бути адаптовані для широкого спектру задач комп'ютерного зору на мобільних платформах

Ключові слова: згорткові нейронні мережі; мобільні пристрої; обробка відеопотоку; Early Stopping Mechanism; residual-and-excitation блоки; комп'ютерний зір

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

Information Technologies and Computer Engineering **Vol. 21, No. 3. 2024**

Article's History: Received: 02.09.2024 Revised: 22.11.2024 Accepted: 26.12.2024

UDC 004.896 10.63341/vitce/3.2024.20

Method of multi-purpose term search in the terminology database

Andrii Yarovyi

Doctor of Technical Science, Professor
Vinnytsia National Technical University
21021, 95 Khmelnytske Shose Str., Vinnytsia, Ukraine
https://orcid.org/0000-0002-6668-2425

Dmytro Kudriavtsev*

Postgraduate Student
Vinnytsia National Technical University
21021, 95 Khmelnytske Shose Str., Vinnytsia, Ukraine
https://orcid.org/0000-0001-7116-7869

Abstract. This study investigated the method of multi-purpose term search in a terminological knowledge base, which is based on semantic analysis and the use of modern natural language processing methods. The study considered the key factors affecting the search efficiency, including the structure of data organisation, data format and parameters, and sample size. Particular focus was placed on the semantic similarity between terms, which allows increasing the search accuracy by using vector representations and the Louvain algorithm. The study also described the use of cosine similarity to quantify the similarity between terms. Furthermore, the search process was optimised by filtering relevant databases and dynamically identifying relevant terms using the modularity metric. A comparative analysis of existing methods for searching for terms by the identified factors was conducted. The study noted the advantages and disadvantages of using the Louvain algorithm in comparison with the search algorithms in graph data structures. A series of experiments were conducted on data samples, including dictionary, graph, and network data structures. The study analysed the use of logistic constraints for searching in network data structures and noted the possibility of optimisation due to uniform and dynamic data distribution. Experimental results showed the effectiveness of using a combination of the Louvain algorithm and network data structures in terminological knowledge bases. Examples of the scope of application of this method in information technologies for searching and processing text data were given. A software architecture scheme with the use of a software interface and the possibility of integration for web applications in the form of a package or library was developed. The proposed approach demonstrates effectiveness in the context of intelligent decision support systems and automated chatbots, which makes it particularly useful for industries where access to accurate professional terms is critical. A basic version of the software interface for using this method in information technologies for searching and analysing data for use in search engines was developed

Keywords: terminological knowledge base; semantic similarity; Louvain algorithm; vector representations; natural language processing

Introduction

In the modern world, the amount of information is constantly growing and the need for efficient data search and analysis is becoming increasingly more relevant. This is particularly true for terminology knowledge bases (TKBs), which are key tools for storing and processing specialised terms in various subject areas. Considering this, finding the right terms to process user queries, especially when working with large amounts of data, requires the introduction of new, more efficient methods. One of these approaches is the method of multi-purpose term search, which allows searching by several criteria simultaneously, considering the complex structure of the TKBs and semantic relationships between terms. The principal task of this method is to identify the terms that best match the user's

Suggested Citation:
Yarovyi, A., & Kudriavtsev, D. (2024). Method of multi-purpose term search in the terminology database. *Information Technologies and Computer Engineering*, 21(3), 20-28. doi: 10.63341/vitce/3.2024.20

*Corresponding author

query. For this, several key factors must be considered: the structure of the data in the TKB, the format and parameters of this data, its volume, as well as the number and size of the sample. Moreover, the semantic similarity between terms plays a significant role, as it is based on the analysis of the relationships between terms and their degree of relevance to the subject area. The relevance of the subject under study is conditioned by the rapid development of data processing technologies and the growing popularity of artificial intelligence (AI) in various fields of activity. The format of textual data stays unchanged, but its variability and the number of subject areas are constantly growing exponentially, which creates a demand for efficient data storage, processing, and use. In this context, terminological knowledge bases play a significant role, as they allow structuring and systematising specialised terms in various subject areas.

In scientific sources, term knowledge bases are considered the basis for storing and managing specialised terms in various fields of knowledge. TKBs perform an essential function in information systems, especially in the context of automated processing of user queries. Existing methods of searching in terminological knowledge bases include a series of approaches, such as keyword search, semantic classification, and machine learning approaches. For instance, keyword-based methods often face relevance issues because they ignore semantic relationships between terms (Abdykerimova *et al.,* 2024). In contrast, semantic classification allows considering the meaning of terms, which reduces the risk of misinterpreting queries.

Current research in this area focused on the development of methods that account for semantic similarity, data structure, and a multi-criteria approach to improve the accuracy and relevance of results (Bourgaux *et al.,* 2024). The early development of search methods in TKBs was based on keyword search and keyword-based approaches used in many classical systems. According to D. Simian & M.-E. Șerban (2024), this approach workes well for databases with clear categories and meanings, but often does not consider complex relationships between terms, which limits its use in large and multi-component TKBs. Some researchers point out the problems of keyword searching due to the need to account for ambiguities in the meaning of words and polysemy (Wu *et al.,* 2023). To overcome these limitations, it became necessary to develop methods based on semantic analysis.

Semantic data processing methods have been considerably developed through vector representations of terms, which have become possible with the development of machine learning. S. Rathje *et al*. (2024) noted that modern methods based on semantic similarity enable a more accurate assessment of the relationships between terms using metrics such as cosine similarity, which is often used to compare values in multidimensional spaces. This approach considers not only the surface meaning of terms, but also their location in the semantic space, which allows automating the search process and increasing its accuracy.

However, according to S. Roy *et al.* (2024) and M. Bienvenu *et al*. (2024), most conventional methods cannot simultaneously process a considerable number of links in the TKBs, which limits their effectiveness. C. Kaya *et al.* (2024) and E. Mohabir & Y. Yoshi (2024) noted that the increasing complexity and volume of knowledge bases requires innovative approaches to their processing, with the researchers focusing on the search for a conceptually new method of multi-objective search.

The purpose of the present study was to develop a multi-purpose search method in terminological knowledge bases, considering the dynamic data structure, data format, and sample size, for further use in information technologies for text data processing. The key features of this study included the use of algorithmisation of intermediate search results processing and the use of machine learning algorithms to identify semantic chains. The key objectives of this study were to identify search characteristics, describe them, and determine the functional features of search using artificial intelligence and iterative search technologies in dynamic data structures.

Materials and Methods

The research methodology involved the development of a software solution for searching for analogous terms in terminological knowledge bases (TKB). The key element of this solution was the introduction of a similarity coefficient between terms, which allows assessing their semantic proximity. The formula for calculating the coefficient considers the number of terms in the semantic chain and the weights of the links between them, factors in the function of determining the semantic value, and the similarity coefficient is defined as the ratio of the number of possible links and the value of the semantic significance to the number of terms connecting a pair of terms:

$$minlength(T_1, T_2) = \min\left(\frac{\sum_{i=1}^{N-1} W_i}{N}\right), \quad (1)$$

where $minlength(T_1, T_2)$ is the function that determines the semantic value of terms T_1 to T_2, T_1 and T_2 are the terms belonging to the same TKBs, between which it is necessary to establish the similarity, $W_i \in \{0,1]$ is the weight of the semantic relationship between neighbouring terms that are part of a chain between terms T_1, T_2, N is the number of terms in the chain between the terms T_1, T_2.

$$S_{T_1T_2} = \frac{N_w * minlength(T_1,T_2)}{N_T}, \quad (2)$$

where N_w is the number of possible semantic chains between terms T_1 *and* T_2, N_T is the number of terms in the chain between the terms T_1, T_2.

To determine the set of terms that best reflect the context of the user's input query, the study analysed vector representations of the terms, which allows comparing them based on semantic relatedness. The vector representation of the data was evaluated using a matrix that reflects the relationships between the terms. This

approach helped to find the most relevant terms and assess their significance in the overall context. To improve the search efficiency, the coefficient of terms belonging to the TKB, namely to the group of root terms that form the kernel of the TKB, was used, which factors in the number and weight of links between terms, as well as their distance to the root term in the chain.

$$F(T_{kernel}, T_1) = \frac{\sum_{j=1}^{J} max\left(\sum_{i=1}^{N-1} s_{T_j T_i}\right)}{J}, i \neq main, \quad (3)$$

where $F(T_{kernel}, T_1)$ is the function that determines the semantic value of a term T_1 to a group of terms T_{kernel}; T_j *are* terms that have the greatest similarity in the TKB and form the TKB kernel; N is the number of terms in the chain between the terms T_i, T_j; J is the minimum number of terms that form the root of the TKB. This parameter can be changed according to the requirements of the method. This ensures that the search is limited to relevant knowledge bases, reducing data processing time and increasing the accuracy of the results by improving the filtering of similar terms.

To optimise the multi-target search, the study employed the Louvain algorithm, which ensures the dynamic formation of term clusters. It factors in the modularity of the graph, which allows filtering relevant terms, reducing the amount of data processed, and improving the accuracy of the answer. The algorithm works in two stages: local modularity optimisation and agglomeration. Modularity measures the density of connections within graph communities compared to a random distribution.

The effectiveness of the proposed approach was tested on three datasets. K-Means clustering, hierarchical clustering, and DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithms were employed for comparison (Sutramiani *et al.,* 2024). In the case of TKBs, the K-Means algorithm was used to cluster terms based on their semantic similarity, where the distance between terms is determined by a similarity metric such as cosine similarity. Additionally, the similarity of vector representations of terms was evaluated using the formula of the average adjacency matrix of the vectors of the input query terms and the TKB terms.

To test the effectiveness of this method based on the combined clustering algorithm, a software implementation was developed considering the technical features of the test data, namely their distribution and preparation for use. The source of the sample data was data from the Kaggle platform (Gabriel, 2020). Each data sample was presented in the form of a TKB implemented on the neo4j graph database and containing over 10,000 terms each.

Results and Discussion

One of the key aspects of developing a software solution for finding similar terms is to introduce a similarity coefficient between two terms. This coefficient allows mathematically estimating how close two terms are in their meaning within the same TKB. The formula for calculating this coefficient factors in the number of terms in the semantic chain and the weights of the links between them. However, term similarity is only one stage of the process. The ultimate purpose is to identify the set of terms that best reflect the context of the user's input query. This is especially significant when working with natural language when a user's message may contain more than one term. In this case, it is necessary to analyse vector representations of term sets, which allows comparing terms based on their semantic proximity. The vector representation of data in this context can be evaluated using a matrix that reflects the relationships between terms. An example of such an analysis can be given to find similarities between vector representations of terms from a TKB. This approach allows not only finding the most relevant terms but also assessing their semantic significance in the overall context.

To improve the search efficiency, it is proposed to use the coefficient of terms belonging to the TKB. This coefficient factors in the number and weight of links between terms, as well as the distance to the root term in the chain. It allows limiting the search to only relevant knowledge bases, which substantially reduces processing time and increases the accuracy of the results. Additionally, the study considered possible options for the data structure for the search, since terms that are linked by semantic relationships involve the use of network data structures. The study also highlighted graphs and adjacency matrices, which act as a mathematical representation of the similarity of terms among themselves. The graph data structure is widely used in graph databases, which is one of the best solutions in the field of TKB data processing (Yuehgoh *et al.,* 2024). For effective search in graphs, breadth-first and depth-first algorithms, clustering algorithms, and ranking algorithms for web-based information systems are used.

A multi-purpose search method that factors in the context of a term and its semantic group should not depend on the subject matter of the information technology in which it is expedient to use it, and therefore ranking algorithms are not relevant. Depth search and breadth search are more acceptable search algorithms, but have a series of disadvantages, such as high dependence on the search volume and organisation of data in the TKB.

Among the known algorithms, the most promising for data retrieval in TKBs is the Louvain algorithm (Sattar & Arifuzzaman, 2018). It is designed to find groups of nodes (clusters or communities) that have more internal connections with each other than with the rest of the graph. The main positive feature of this algorithm is its high efficiency in working with large networks of TKB data due to its scalability and speed of data search. The purpose of the algorithm is to maximise a metric called modularity. Modularity measures the quality of graph partitioning into communities by comparing the density of links within communities with a random distribution coefficient of such links, and its implementation is divided into two stages: local modularity optimisation and agglomeration. In this algorithm, modularity is a measure that determines how well a graph is divided into communities. A high modularity means that there are

considerably more connections within each community than between communities. The modularity formula considers not only the number of links, but also their expected number in a random graph with the same vertex degrees, which allows comparing the quality of clustering with a random distribution of links. Despite a series of advantages, the disadvantages include sensitivity to the initial state, namely the initial organisation of data in the TKB, and searching in local graph maxima, which may overlook small groups of terms among large groups of terms.

The core value of Louvain's algorithm lies in providing optimisation for multi-purpose term search by dynamically forming clusters of terms in the TKB. Consideration of semantic similarity and modularity allows filtering relevant terms for search, reducing the amount of data processed and improving the accuracy of the answer. This increases the efficiency of the multi-target search algorithm, ensuring accurate and fast detection of terms that match the user's query in large amounts of information.

The multi-target search method uses the Louvain algorithm with the use of not only the modularity metric, but also the metric of similarity of terms among themselves and the metric of similarity with the TKB. Thus, the multi-purpose term search method provides a comprehensive approach to query processing in terminological knowledge bases. It combines the analysis of semantic relationships between terms, evaluation of vector representations, and selection of relevant TKBs, as well as the modularity of term groups, which allows working efficiently with large amounts of data and ensuring high accuracy of results when searching in several TKBs. To perform a search for similar terms, it is necessary to enter the similarity coefficient between two terms, which is determined by formulas (1) and (2).

The factual finding of the similarity of terms is only part of the overall solution, as it involves only the preparation of the TKB data. The factual search is performed by identifying a fixed set of terms that best reflect the context of the user's input, namely the terms with the highest similarity coefficient. It should also be understood that most of the TKB data contains terms that include not only words, but also sentences containing more than one term, and therefore it is necessary to determine the similarity of vector representations of the term sets, which is represented in the form of a matrix. The similarity of vector representations is defined as the arithmetic mean of the adjacency matrix of the vector of terms of the input message and the vectors of terms of the data from the TKB, which is presented in the following formula:

$$S = \frac{1}{N*M}\sum_{i=1}^{N}\sum_{j=1}^{M} A_{ij}. \quad (4)$$

For instance, let us determine the similarity of vector representations for the vector of terms of the incoming message (T_1, T_2, T_3) and the vector representation of terms from the TKB (T_A, T_B, T_C, T_D, T_E). The calculation results are presented in Table 1.

Table 1. Example of searching for similar terms

	T_A	T_B	T_C	T_D	T_E
T_1	0.912	0.823	0.754	0.843	0.761
T_2	0.856	0.904	0.783	0.735	0.819
T_3	0.759	0.831	0.905	0.861	0.913

Source: developed by the authors

$$S = \frac{1}{15} * (0.912 + 0.823 + 0.754 + 0.843 + 0.761 + 0.856 + 0.90 + 0.783 + 0.735 + 0.819 + 0.759 + 0.831 + 0.905 + 0.861 + 0.913)$$

$$S = \frac{1}{15} * 12.458 = 0.831.$$

Considering the Louvain algorithm in terms of its application within a search within a single TKB, the first stage is local modularity optimisation. At this stage, the algorithm analyses each node and tries to move it to a neighbouring community if this leads to an improvement in modularity. The process consists of the following steps:

1. For each node, the algorithm calculates the change in modularity that will occur when this node is moved from its current community to one of the neighbouring communities.

2. If moving a node to a neighbouring community produces an increase in modularity, the node is moved to that community.

3. This process is repeated for all nodes in the graph until the condition of local modularity maximisation is met, i.e., when no more moves can improve modularity.

Notably, at each step, modularity is calculated using the formula 5 (Sattar & Arifuzzaman, 2018):

$$Q = \frac{1}{2m}\sum_{ij}\left[A_{ij} - \frac{k_i k_j}{2m}\right]\delta(c_i, c_j), \quad (5)$$

where Q is the modularity, m is the number of edges in the graph, k_i, k_j are the degrees of nodes i and j, A_{ij} is the element of the graph adjacency matrix, $\delta(c_i, c_j)$ is a function equal to 1 if nodes i, j belong to the same group and 0 if they belong to different groups. The purpose of the algorithm is to maximise the value of Q, which means that the graph is divided into communities in such a way that there are more links within each community than between different communities. This enables better cohesion of terms within a community and makes it easier to find relevant terms. Louvain's algorithm is an effective tool for working with large TKBs, as it can handle large networks of terms with numerous relationships. The main advantage of the algorithm is its scalability and speed. After forming clusters based on the modularity metric, the algorithm provides an opportunity to optimise the search by using semantic similarity

within each community. This allows the search to focus on relevant terms within each cluster, considerably reducing the amount of data to process. Louvain's algorithm has its drawbacks: it can be sensitive to the initial state and easily gets caught in local modularity maxima, which can lead to the formation of too large or small clusters, especially when working with large graphs. However, the use of additional metrics, such as cosine similarity, can reduce the effects of these limitations and improve the accuracy of results in multicomponent TKBs.

By using the Louvain algorithm and applying metrics to determine the similarity of terms among themselves, it is possible to achieve faster search results by reducing the number of search iterations within the graph data structure. After the software implementation, it is worth testing with the analogue methods used to search for terms in the TKB.

The main advantage of the K-Means algorithm is its speed and ability to process large amounts of data in a relatively small number of iterations. However, the algorithm has a series of limitations in the context of TKBs:

1. Sensitivity to the initial choice of K: The algorithm requires a predefined number of clusters (K), which can be difficult to predict in the context of dynamic TKB data.

2. Ignoring connectivity: K-Means does not consider the existence of relationships between terms and works solely based on distance to the centroid. This leads to the loss of information about the internal relationships between terms that are important in TKBs, which contradicts the purpose of the study, namely, multi-target search.

3. The presence of local minima: there is a possibility of a closed loop in the local minima, which will lead to the absence of an optimised solution, especially when there are many terms and complex semantic relationships.

Considering these limitations, the K-Means algorithm in TKB should be supplemented with the Louvain algorithm, which better accounts for the connectivity of terms and can determine the number of clusters dynamically. For clustering terms in a TKB, the Louvain algorithm has major advantages over K-Means because it uses a graph structure to distribute terms based on internal connectivity. The purpose of the Louvain algorithm is to maximise modularity, which considers the density of internal connections between terms. This provides a more natural distribution of data, especially in large graph structures where terms have a complex network of connections. At the same time, in tasks where clustering is based on semantic distance without complex relationships, K-Means can show high performance and be useful for the initial distribution of terms.

Hierarchical clustering and DBSCAN are also worth considering. Hierarchical clustering is based on the creation of a cluster tree structure. This approach allows analysing terms at multiple levels and identifying subclasses within large clusters, which can be useful for complex multilevel TKBs. DBSCAN creates clusters based on the density of points in space, which makes it useful for finding closely related groups of terms and isolating "noise" or terms that do not belong to any cluster. However, among the disadvantages of hierarchical clustering is its high computational complexity, which requires extensive computing resources, especially for large data sets, which are usually TKBs. This can create problems for scalability, as time and computational costs increase substantially with the number of terms. Another disadvantage is the lack of the ability to re-form clusters, namely, to change the structure after it has been built, which does not allow changing clusters when new data becomes available. This limits its use in dynamic TKBs where the database is constantly updated with new terms. Additionally, there is a disadvantage associated with low noise immunity, which can lead to the emergence of unwanted clusters, which is levelled by adding term filtering. Among the disadvantages of DBSCAN are the difficulty of working with multidimensional data, namely in graph data structures, where the dimensionality is determined by the number of edges, and the disregard for weak relationships between terms within even one TKB.

Considering the advantages and disadvantages of each method, a combined approach to clustering terms in TKBs was chosen, where K-Means is used for pre-clustering large amounts of data, while the Louvain algorithm is applied at the second stage to optimise internal links within each cluster. This strikes a balance between processing speed and clustering accuracy, ensuring that terms within each cluster are relevant. Firstly, K-Means is used to create initial clusters based on the semantic distance between terms, which reduces the amount of computation. Then, within each of these clusters, the Louvain algorithm is applied to refine the groups based on the relationships between the terms and improve the quality of the clustering. A diagram of the combined algorithm is presented in Figure 1. This approach not only improves the clustering accuracy but also enables efficient data processing in large TKBs, where the number of terms and their relationships can vary considerably.

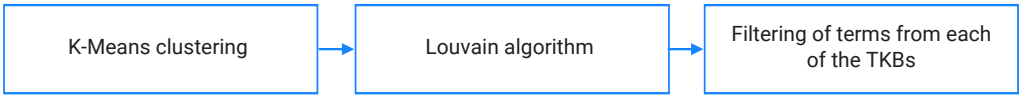

Figure 1. Combined clustering algorithm

Source: developed by the authors

A detailed description of the experimental results is presented in Table 2, where the developed method is called MultiSearch. The point of this experiment is to determine the optimised data distribution and the effectiveness of using iterations to find the corresponding clusters. Moreover, the cluster formation is based

on the fulfilment of the term relatedness condition and the determination of modularity by the Louvain algorithm. The hardware specifications for all methods are the same, and the software implementation was created using the Python 3.10 programming language and the TensorFlow package. The samples consist of prepared datasets formed in the form of graph databases. To reduce the impact of hardware and technical errors, a series of experiments was conducted 100 times. The data in the table are averages of all experiments.

Table 2. Experimental results

	Sample 1, s	Sample 2, s	Sample 3, s	Clusters	Iterations
DBSCAN	1.504	1.624	1.977	78	194/362/499
Hierarchical clustering	1.417	1.829	1.863	77	271/357/432
K-Means	1.105	1.272	1.715	84	175/282/335
MultiSearch	0.961	1.138	1.469	89	174/197/234

Source: developed by the authors

The experiments confirmed the effectiveness of the developed method when working with different datasets represented in graph databases. The findings demonstrated that the proposed multi-objective search method outperforms other clustering-based search methods, providing an average of 11.78-16.75% more efficient search while maintaining high quality cluster formation. The developed method proved its effectiveness in multiple tests, demonstrating stable results due to optimisation based on modularity and term affinity. The proposed combined clustering approach using K-Means and Louvain algorithms provided an effective balance between processing speed and clustering accuracy, especially for large graph databases.

One of the key stages in the formation of TKBs is the use of graph data structures and clustering. C. Li *et al*. (2022) described the use of intelligent search engines based on knowledge of graph structures, where each term is represented as a node in a graph, while the links between them are represented as edges. This allows building networks of links between terms, specifically for large databases where the number of links considerably exceeds the number of terms. The findings of the present study proved the effectiveness of using graph data structures for multi-criteria search, which was a convincing argument for using graph data structures in TKBs.

The use of graph structures greatly improves performance in multi-component systems, which is confirmed by the findings of S. George *et al*. (2019), where the graph structure showed high efficiency in processing large amounts of data due to its scalability. During the review of modern solutions in the field of term clustering, attention was focused on finding an efficient algorithm with scaling stability. Thus, S. Sattar & S. Arifuzzaman (2018) characterised the Louvain algorithm, designed to identify communities in large graphs, as one of the most effective clustering tools for term knowledge bases. This algorithm allows dividing a graph into a set of subgraphs, each of which is characterised by a high density of internal connections (modularity). Thus, Louvain's method provides efficient clustering that increases the relevance of searching in TKBs, since each cluster can be considered as a group of terms with strong semantic relationships. The disadvantages of the algorithm are its dependence on the initial state of the data in the graph and the difficulty in recognising smaller groups among large communities, which was emphasised by Y. Zhang *et al*. (2024). Considering the shortcomings of its application, in this study, their influence was not critical, since effective filtering of terms by the affinity criterion levelled the main drawback, namely, sensitivity to small groups in term clustering.

The multi-purpose search approach, which combines clustering and semantic similarity methods, is a modern trend that is gaining popularity due to its ability to simultaneously process several criteria. According to Y. Zhao & T. Wang (2021), this approach allows not only to account for the structured nature of the data, but also semantic relevance, which increases the search accuracy in dynamic knowledge bases. The main advantage is the ability to process queries considering many factors, such as data format, structure, and sample size, which makes it effective for various fields, including medicine and technical support. The use of artificial intelligence, specifically deep learning, greatly improves the accuracy of searching in TKBs, enabling the analysis and classification of large amounts of data automatically. H. Baqal & M. Sidiq (2024) noted that modern AI models can not only find relevant terms, but also learn based on previous queries, improving performance with each use. This is particularly relevant for chatbot applications, where systems can automatically update the knowledge base with new data and improve the accuracy of responses to users. Incorporating artificial intelligence into the term search process in a TKB provides systems with the ability to be adaptive, which is crucial in dynamic environments.

AI has made it possible to considerably improve the accuracy and speed of finding relevant data, which was best described by N.F. Lindemann (2024). The use of machine learning and deep learning techniques allows building complex models of semantic similarity between terms, which improves the quality of search results. AI algorithms can analyse large amounts of data, identify hidden relationships between terms and determine their relevance to a concrete user query. At the same time, multi-targeted search is becoming increasingly important, allowing

several factors to be factored in simultaneously when searching for information, such as semantic relationships between terms, the context of a user's query, and whether the terms belong to a particular subject area. This is crucial to ensure accurate and fast access to data in the face of the diversity of information stored in TKBs.

Intelligent systems capable of recognising, classifying, and interpreting data are becoming a prominent part of the modern information infrastructure. A. Yarovyi & D. Kudriavtsev (2021) focused on the use of neural networks and machine learning technologies for text processing tasks, namely classification and context detection. Additionally, the researchers proposed to combine the use of optimised semantic text analysis and recurrent neural network methods as one of the examples of effective data analysis and context detection. The relevance of such a method was confirmed in the context of developing chatbots specialising in the search for relevant information, as described by A. Morayo *et al.* (2024). Another striking example of intelligent systems is decision support systems, which also require fast and accurate term search in a large amount of information (Gupta & Singh, 2024). The method of multi-purpose term search in TKBs can be widely used in the development of intelligent information systems, as discussed by D. Beeram (2024), which help users quickly find the information they need and make informed decisions in various fields of activity. The multi-objective search method proposed in this study allows combining the strengths of various methods, ensuring high relevance and accuracy of the results.

Conclusions

The present study developed a method for multi-purpose term retrieval in terminological knowledge bases, which combines the analysis of semantic relations between terms and vector representations of terms. The proposed method allows accounting for various factors during the search, including the structure of the TKB, semantic similarity of terms, and the context of the user's query. This method is based on the idea of combining the K-Means clustering algorithm and the Louvain algorithm to optimise search processes, which greatly improved the accuracy and speed of search when working with large amounts of data. The use of modern algorithms, such as the Louvain algorithm for community detection and term clustering, helped to work effectively with large graph databases and ensure high search performance. Furthermore, it was found that the integration of the Louvain algorithm and term similarity coefficients greatly improves search results, reducing the amount of data processed and focusing the search on the most relevant terms. Therewith, the number of clusters formed indicates the advantage of using the combined approach, since a greater number of clusters with the term similarity coefficient as a filter for the appearance of unwanted terms indicates a greater diversity of the context of terms in the TKBs. The experiments demonstrated the effectiveness of the proposed method when working with several data samples implemented in graph databases. The obtained findings revealed that the multi-objective search method is competitive in comparison with other clustering-based search methods, being on average 11.78-6.75% faster than the best result. This method provides a comprehensive approach to query processing in TKBs and can be applied in various industries requiring fast and accurate access to specialised terminology. Further research could focus on identifying patterns and models of communication between data organisation structures in TKBs, including, apart from graph and network data structures, hash tables to improve the efficiency of multi-purpose search.

Acknowledgements

None.

Conflict of Interest

The authors declare no conflict of interest

References

[1] Abdykerimova, L., Abdikerimova, G.B., Konyrkhanova, A., Nurova, G., Bazarova, M., Bersugir, M., Kaldarova, M., & Yerzhanova, A. (2024). Analysis of the emotional coloring of text using machine and deep learning methods. *International Journal of Electrical and Computer Engineering (IJECE),* 14, article number 3055. doi: 10.11591/ijece.v14i3.pp3055-3063.

[2] Baqal, H., & Sidiq, M. (2024). Graph databases: Revolutionizing database design and data analysis. *Current Journal of Applied Science and Technology*, 43, 45-56. doi: 10.9734/cjast/2024/v43i114443.

[3] Beeram, D. (2024). Combining deep learning and heuristic search for efficient text summarization. *International Research Journal of Engineering and Technology (IRJET),* 11(8), 23-34.

[4] Bienvenu, M., Bourgaux, C., & Jean, R. (2024). Cost-based semantics for querying inconsistent weighted knowledge bases. In *Proceedings of the 21st international conference on principles of knowledge representation and reasoning* (pp. 167-177). Hanoi: CAI Organization. doi: 10.24963/kr.2024/16.

[5] Bourgaux, C., Guimarães, R., Koudijs, R., Lacerda, V., & Ozaki, A. (2024). Knowledge base embeddings: Semantics and theoretical properties. In *Proceedings of the 21st international conference on principles of knowledge representation and reasoning* (pp. 823-833). Hanoi: International Joint Conferences on Artificial Intelligence Organization. doi: 10.24963/kr.2024/77.

[6] Gabriel, A. (2020). Kensho derived Wikimedia dataset. Retrieved from https://www.kaggle.com/datasets/kenshoresearch/kensho-derived-wikimedia-data.

[7] George, S., Elayidom, M.S., & Santhanakrishnan, T. (2019). Semantic desktop search engine using graph database. *International Journal of Recent Technology and Engineering*, 8(1S2), 373-375.
[8] Gupta, A., & Singh, T. (2024). Study of various frameworks to develop intelligent chatbots. *International Journal of Innovative Science and Research Technology (IJISRT)*, 9(4), 2969-2978. doi: 10.38124/ijisrt/IJISRT24APR1290.
[9] Kaya, C., Kilimci, Z.H., Uysal, M., & Kaya, M. (2024). A review of metaheuristic optimization techniques in text classification. *International Journal of Computational and Experimental Science and Engineering*, 10(2). doi: 0.22399/ijcesen.295.
[10] Li, C., Liang, M., & Qiu, D. (2022). An intelligent search system based on knowledge graph. In *2022 International conference on artificial intelligence of things and crowdsensing (AIoTCs)* (pp. 66-70). Nicosia: IEEE. doi: 10.1109/AIoTCs58181.2022.00017.
[11] Lindemann, N.F. (2024). Chatbots, search engines, and the sealing of knowledges. *AI & Society.* doi: 10.1007/s00146-024-01944-w.
[12] Mohabir, S.E., & Joshi, Y.C. (2024). A bibliometric analysis of the knowledge base on multinational corporations' behavior. *SN Business & Economics,* 4, article number 105. doi: 10.1007/s43546-024-00705-7.
[13] Morayo, A., Samuel, J., Kennedy, O., Adeyinka, A., Adenugba, A., & Imhade, O. (2024). Development of an artificial intelligent health chatbot for improved telemedicine. In C. So In, N.D. Londhe, N. Bhatt & M. Kitsing (Eds.), *Information systems for intelligent systems. ISBM 2023. Smart innovation, systems and technologies* (Vol. 379, pp. 585-600). Singapore: Springer. doi: 10.1007/978-981-99-8612-5_48.
[14] Rathje, S., Mirea, D.-M., Sucholutsky, I., Marjieh, R., Robertson, C., & Van Bavel, J. (2024). GPT is an effective tool for multilingual psychological text analysis. *Proceedings of the National Academy of Sciences of the United States of America,* 121, article number e2308950121. doi: 10.1073/pnas.2308950121.
[15] Roy, S., Bharaty, A., Sarkar, S., Sehgal, M., & Panchal, R. (2024). A hybrid ensemble approach for short-text sentiment analysis integrating deep learning and traditional machine learning methods. *ResearchGate*. doi: 10.13140/RG.2.2.15182.88643.
[16] Sattar, N.S., & Arifuzzaman, S. (2018). Parallelizing Louvain algorithm: Distributed memory challenges. In *2018 IEEE 16th Intl conf on dependable, autonomic and secure computing, 16th intl conf on pervasive intelligence and computing, 4th intl conf on Big Data intelligence and computing and cyber science and technology congress (DASC/PiCom/DataCom/CyberSciTech)* (pp. 695-701). Athens: IEEE. doi: 10.1109/DASC/PiCom/DataCom/CyberSciTec.2018.00122.
[17] Simian, D., & Şerban, M.-E. (2024). Improving search query accuracy for specialized websites through intelligent text correction and reconstruction models. *Information*, 15, article number 683. doi: 10.3390/info15110683.
[18] Sutramiani, N., Arthana, I.M.T., Lampung, P.F., Aurelia, S., Fauzi, M., & Darma, I.W.A.S. (2024). The performance comparison of DBSCAN and K-Means clustering for MSMEs grouping based on asset value and turnover. *Journal of Information Systems Engineering and Business Intelligence*, 10, 13-24. doi: 10.20473/jisebi.10.1.13-24.
[19] Wu, L., Hu, J., Teng, F., Li, T. & Du, S. (2023). Text semantic matching with an enhanced sample building method based on contrastive learning. *International Journal of Machine Learning and Cybernetics*, 14, 3105-3112. doi: 10.1007/s13042-023-01823-8.
[20] Yarovyi, A. & Kudriavtsev, D. (2021). Multi-purpose search to determine the context of a text message based on the dictionary data structure. In *2021 IEEE 16th international conference on computer sciences and information technologies (CSIT)* (pp. 65-68). Lviv: IEEE. doi: 10.1109/CSIT52700.2021.9648803.
[21] Yuehgoh, F., Djebali, S., & Travers, N. (2024). Leveraging recommendations using a multiplex graph database. *International Journal of Web Information Systems*, 20(5). doi: 10.1108/IJWIS-05-2024-0137.
[22] Zhang, Y. *et al*. (2024). A materials terminology knowledge graph automatically constructed from text corpus. *Scientific Data*, 11, article number 600. doi: 10.1038/s41597-024-03448-0.
[23] Zhao, Y., & Wang, T. (2024). Knowledge base embeddings for a recommendation based on overlapping knowledge and graph learning. *Arabian Journal for Science and Engineering*. doi: 10.1007/s13369-024-09573-7.

Метод багатоцільового пошуку термів в термінологічній базі

Андрій Яровий
Доктор технічних наук, професор
Вінницький національний технічний університет
21021, вул. Хмельницьке шосе, 95, м. Вінниця, Україна
https://orcid.org/0000-0002-6668-2425

Дмитро Кудрявцев
Аспірант
Вінницький національний технічний університет
21021, вул. Хмельницьке шосе, 95, м. Вінниця, Україна
https://orcid.org/0000-0001-7116-7869

Анотація. У статті досліджувався метод багатоцільового пошуку термів у термінологічній базі знань, який базується на семантичному аналізі та використанні сучасних методів обробки природної мови. Розглянуто ключові фактори, що впливають на ефективність пошуку, зокрема структуру організації даних, формат і параметри даних, а також обсяг вибірки. Особлива увага була приділена семантичній подібності між термами, що дозволяє підвищити точність пошуку за рахунок векторних представлень та алгоритму Лувена. У статті також описано застосування косинусної подібності для кількісної оцінки подібності між термами. Крім того, оптимізовано процес пошуку шляхом фільтрації релевантних баз даних і динамічного визначення релевантних термів за допомогою метрики модульності. Виконано порівняльний аналіз наявних методів пошуку термів за визначеними факторами. Відзначено переваги та недоліки використання алгоритму Лувена у порівнянні з алгоритмами пошуку в графових структурах даних. Виконано ряд експериментів на вибірках даних, включаючи словникову структуру даних, графову та мережеву структуру даних. Проаналізовано використання логістичних обмежень для пошуку в мережевих структурах даних та відзначено можливість оптимізації за рахунок рівномірного та динамічного розподілу даних. Результати експериментів показали ефективність застосування комбінації алгоритму Лувена та мережевих структур даних в термінологічних базах знань. Подано приклади сфери застосування даного методу в інформаційних технологіях пошуку та обробки текстових даних. Розроблено схему архітектури програмного забезпечення із використанням програмного інтерфейсу та можливості інтеграції для веб-застосунків у вигляді пакету чи бібліотеки. Пропонований підхід продемонстрував ефективність у контексті інтелектуальних систем підтримки рішень і автоматизованих чат-ботів, що робить його особливо корисним для галузей, де критично важливий доступ до точних фахових термів. Розроблено базову версію програмного інтерфейсу для використання даного методу в інформаційних технологіях пошуку та аналізу даних для використання в пошукових системах

Ключові слова: термінологічна база знань; семантична подібність; алгоритм Лувена; векторні представлення; обробка природної мови

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

Information Technologies and Computer Engineering **Vol. 21, No. 3. 2024**

Article's History: Received: 20.08.2024 Revised: 08.11.2024 Accepted: 26.12.2024

UDC 621.39:004.738.5 10.63341/vitce/3.2024.29

Optimising service quality and network efficiency in legacy networks by integrating SDN and broadband

Oleksandr Pidpalyi*

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

Abstract. The study aimed to develop an empirical model for optimising the quality of service (QoS) and improving the efficiency of telecommunications networks by integrating software-defined networking (SDN) and broadband Internet access technologies. The study employed simulation modelling, scenario analysis and analytical models with the use of modelling tools. The main findings of the study highlighted the significant potential of integrating SDN and broadband technologies to improve the QoS and efficiency of telecommunications networks. SDN concepts were demonstrated, which provide centralised network management and flexibility in configuration, as well as broadband access, which offers high data rates and improved bandwidth. The role of each network element, including routers, switches and controllers, and their impact on network efficiency was identified. An analysis of the interaction of SDN with broadband access networks has shown that the use of such networks allows optimising routing, load balancing and traffic management, which helps to improve network speed and reliability. QoS metrics demonstrated that the integration of different technologies leads to significant improvements in bandwidth, packet loss, latency and latency variability. In general, the network model showed the effectiveness of SDN and broadband integration in optimising network performance and QoS, and a review of network modelling methods showed that the use of simulation tools allows for a detailed assessment of the effectiveness of technology integration and confirmation of their positive impact on network performance. Thus, results confirmed that the integration of SDN and broadband technologies significantly improves the efficiency of telecommunications networks, which indicates the effectiveness of new technologies in increasing the overall performance of networks

Keywords: software management; wireless technologies; resource virtualisation; capacity analysis; adaptive systems

Introduction

The growth in traffic volumes and the diversity of applications in networks create new requirements for quality of service (QoS) and network management efficiency. Traditional network architectures often face limitations in delivering the high data rates, reliability and scalability required to support applications. There are also difficulties in integrating new technologies, such as software-defined networking (SDN) and broadband, into existing network infrastructures. The main challenge is that the integration of these technologies requires effective management and optimisation techniques that can cope with instability and latency, and balance data rates and QoS.

In general, SDN is an innovative computer network management architecture that revolutionises the traditional approach to network administration by separating management and data transmission functions. The main components of SDN are a controller, a southbound interface and a northbound interface. The SDN controller is the central element of the system, responsible for global network management, policy enforcement and decision-making. It communicates with network devices via a southern interface, usually implemented using the OpenFlow protocol. The Northern Interface provides communication between the controller and applications or services, allowing them to access network resources and functions.

Broadband access networks are critical to providing high-speed and continuous Internet connectivity in a variety of environments, including residential, commercial and institutional. The main types of broadband networks include fibre-optic networks, which provide extremely high

Suggested Citation:
Pidpalyi, O. (2024). Optimising service quality and network efficiency in legacy networks by integrating SDN and broadband. *Information Technologies and Computer Engineering*, 21(3), 29-42. doi: 10.63341/vitce/3.2024.29

*Corresponding author

data transfer speeds using light signals through optical fibres. Cable networks use coaxial cables to transmit data, television signals and voice services. Digital Subscriber Line (DSL) uses existing telephone lines to provide broadband access at different speeds depending on the distance to the site. Wireless technologies provide connections via radio waves and mobile networks that use cellular base stations to provide mobile Internet access.

The integration of SDN and broadband technologies is a key aspect of increasing the efficiency and adaptability of modern telecommunications infrastructures. SDN allows for centralised network management, which enables flexible and dynamic configuration of network resources, reducing the need for manual intervention and enabling faster response to changing requirements. In turn, integration with various broadband technologies allows for optimised load balancing, improved QoS and reduced latency, which is critical to ensuring a high-speed and reliable Internet connection.

The analysis of similar studies is key to determining the direction of further research and addressing existing gaps in the field. For example, the work of I. Dulska (2019) showed the problems of adapting broadband access statistics in Ukraine to international standards, pointing out the inconsistency of national indicators with European criteria and the fragmentation of data, which makes it difficult to accurately assess and compare broadband speeds and coverage. M. Vasylkivskyi *et al.* (2023) demonstrated that the introduction of SDN and drone technologies has opened new opportunities for network optimisation in challenging environments, noting significant improvements in QoS and network efficiency. In addition, V.I. Drovovozov *et al.* (2022) investigated methods for integrating heterogeneous 4G and 5G wireless networks to improve QoS.

S.A. Trivedi (2024) considered cross-layer design methods that can improve communication between layers and overall network efficiency, including optimising its performance and traffic management. D. Sarabia *et al.* (2024) showed an architecture for integrating Recursive Inter-Network Architecture (RINA) with SDN, which reduced the barriers to RINA implementation on the Internet of Things (IoT) environment through the use of distributed applications and virtual private networks. S. Khan *et al.* (2021) developed a new approach to resource management in 5G networks based on SDN and Network Function Virtualisation (NFV), proposing a new resource management policy. Additionally, M. Lonare Mahesh & M.S. Devi (2022) applied modern genetic algorithms to detect and manage congestion in SDN-based networks, which allows for more efficient management of network traffic and reduced network load.

Moreover, M. Aboughaly & S.A. Hannan (2024) addressed QoS optimisation in SDN with a system that demonstrated a significant advantage over traditional QoS solutions. H. Ma *et al.* (2024) presented a new SDN controller along with a routing algorithm based on multi-criteria optimisation, which significantly improved the throughput in satellite networks. D.S. Sahana & B. Savadatti (2024) proposed a new architecture for secure authentication and access control in SDN for IoT, which provided increased network security and efficiency.

The study aimed to develop a model for improving the QoS and efficiency of telecommunications networks by integrating SDN and broadband Internet access technologies. To this end, the following tasks were performed: a comprehensive analysis of the implementation of SDN technologies to optimise access to broadband networks, an assessment of the impact of SDN integration on increasing the speed, efficiency and reliability of networks, a study of practical examples of successful SDN implementation in telecommunications systems, and an analysis of the main problems and challenges encountered when integrating SDN with broadband Internet access.

Materials and Methods

First, a theoretical overview of the concepts of SDN and broadband access was conducted. For this purpose, an analysis of existing scientific publications in this area was carried out. The concepts of SDN were learned through documentation on centralised network management and its flexible capabilities.

The analysis of broadband access networks included a detailed description of different types of networks, such as fibre-optic, DSL and cable modems, with a focus on their ability to provide high data rates and scalability. In this context, the study addressed how SDN-technologies affect the overall performance of the network, their main advantages and capabilities.

All key elements of networks, including routers, switches, controllers, access points, were described. Their role in solving network problems, such as routing, load balancing and traffic management, was considered. It identifies how each component contributes to the overall goal of optimising and improving network efficiency.

Next, the study investigated the possibilities of implementing SDN to optimise a broadband access network. Simulation modelling techniques were used to assess the effectiveness of SDN implementation in various broadband access scenarios. For this purpose, network models with different SDN and broadband access configurations were set up to assess the impact of these technologies on network speed, efficiency and reliability.

Particular attention was devoted to the analysis of the impact of SDN on aspects such as routing optimisation, load balancing and traffic management. It was assessed how SDN can improve these processes by providing more efficient resource management and reducing delays and packet loss. In addition, the main areas of broadband network optimisation were discussed, and examples of practical use of SDN in broadband networks were provided.

At the stage of assessing the impact of the integration of SDN and broadband technologies on QoS indicators, a detailed analysis of various metrics that determine network efficiency was carried out. The following key metrics were used for this purpose:

1. Bandwidth (BW) (1):

$$BW = \frac{T}{D}, \quad (1)$$

where D – amount of data transmitted during the time T.

2. Delay (Del) (2):

$$Del = PpDel + TDel + QDel + PcDel, \quad (2)$$

where *PpDel* – signal propagation delay over the network, *TDel* – delay in data transmission through the communication channel, *QDel* – queuing delay in network devices, and *PcDel* – data processing delay on routers.

3. Packet Loss (PL) (3):

$$PL\ Rate = \frac{L}{T} \times 100\%, \quad (3)$$

where L – number of lost packets during the time T.

4. Jitter (J) (4):

$$J = SD\ of\ Del, \quad (4)$$

where *SD* – standard deviation.

For each of these metrics, the performance before and after the introduction of SDN and broadband technologies was compared. A significant increase in throughput, a decrease in latency and latency variability, and a reduction in packet loss were noted.

The network was modelled using the Diagrams.net platform, which was used to build a network diagram. Simulation tools such as Simulink, Network Simulator 3 (NS-3), OMNeT++, Mininet and Graphical Network Simulator 3 (GNS3) were also analysed. These tools were chosen due to their ability to create virtual network models and analyse their operation in conditions close to real-world conditions. Simulink was used to build and analyse the service request flow model. NS-3 and OMNeT++ provided an in-depth analysis of network protocols and network behaviour. Mininet was used to analyse virtual networks using real network components and protocols. GNS3 was used to analyse the integration of real network devices and virtual models, which made it possible to evaluate the effectiveness of SDN implementation in real-world conditions, considering the specifics of specific equipment and configurations.

Results

SDN is an innovative architecture that enables centralised network management through software interfaces. This concept differs from traditional network architectures in that it separates the control layer from the forwarding layer, allowing for more flexible network management and configuration.

The main components of SDN include a controller, peripherals, applications, and north and south Application Programming Interface (API). The controller is the central element of the SDN architecture, which is responsible for managing the network and coordinating actions between different components. It receives information from network devices and uses it for traffic routing decisions. Edge devices are network switches that perform traffic-forwarding functions based on instructions from the SDN controller. They do not have internal routing algorithms and always rely on the controller's instructions.

Applications, in turn, are software applications that use SDN capabilities to implement various functions, such as virtual networks, load balancing, etc. As for the API, it can be north and south, where the north API is an interface for the SDN controller to interact with applications and services, which allows applications to interact with the controller to configure and monitor the network. The southern API is an interface for the controller to interact with peripherals, usually using the OpenFlow protocol to send commands to switches.

The SDN architecture consists of a controller that centrally manages the network, using a northbound API to interact with applications and a southbound API to manage network devices such as switches. Applications manage the network through software configuration, which provides abstraction of the physical infrastructure and centralised management of the entire network (Fig. 1).

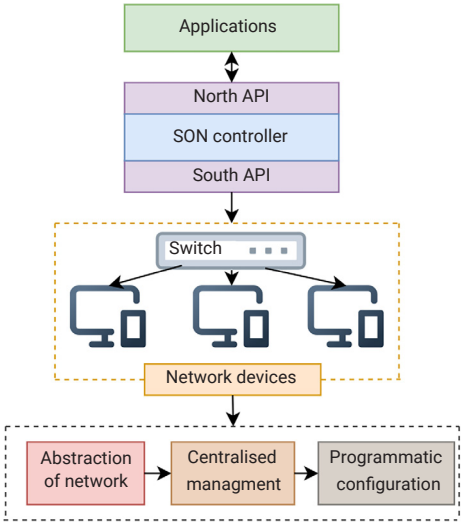

Figure 1. SDN architecture diagram
Source: compiled by the author

In general, SDN is widely used in various fields due to its unique capabilities. In data centres, SDN is used to create flexible and scalable networks, which provide efficient management of many resources and high performance. In cloud computing, SDN helps to dynamically allocate resources and ensure high availability of services. This is important for maintaining high QoS standards and efficient resource utilisation. Operator networks are also using SDN to introduce new services and improve the efficiency of their networks. This allows operators to adapt more quickly to changing user needs and improve their network infrastructures.

In addition, SDN can be implemented in enterprise networks to dynamically manage resources and improve QoS. For instance, in conditions of high traffic volatility, during large online events, SDN can automatically reallocate network resources in real-time, providing priority access to resources for critical applications and services. This will reduce latency and avoid network congestion, ensuring stable performance even during peak loads.

In a 5G network environment, SDN can be used for intelligent traffic management, particularly in IoT and autonomous vehicle scenarios. SDN enables more precise traffic management based on the specific QoS requirements of different types of data, such as data from IoT sensors, video streams from surveillance cameras, or data from vehicles. This approach can increase the efficiency of network resources and improve overall network performance.

In cloud data centres, SDN can be used to orchestrate network operations more efficiently. This can include automated deployment of virtual networks, configuration of security policies, and monitoring of network traffic. For example, in cloud environments where resources are shared by many users at the same time, SDN can enable dynamic changes to the network topology based on user needs, which can improve performance and minimise latency.

In smart cities, SDN can provide centralised management and integration of various city services, such as transport, energy, security and communications. By using SDN, a flexible and scalable infrastructure that can adapt to the growing number of connected devices and ensure reliable operation of city services can be created. This can include dynamic traffic routing based on real-world conditions in the city, such as traffic congestion or accidents.

Broadband networks are the infrastructure that provides high-speed Internet access to end users. They are critical to modern communications systems as they provide the necessary bandwidth for real-time data, voice and video transmission. The following components of broadband access networks should be distinguished: access networks and wireless access technologies (Table 1).

Table 1. Types of broadband network technologies and examples of their use

Types of technologies	**Description of access network technologies**	**Examples of use**
DSL	A technology that uses existing telephone lines to transmit data at high speeds. DSL provides high download and upload speeds at relatively low costs	Home networks in rural areas
Cable modems	They use coaxial cables to provide Internet access. Thanks to the high bandwidth of coaxial cables, cable modems can provide fast, high-bandwidth Internet access.	Internet in residential complexes
Optical fibres	Fibre-optic networks provide the highest speed and lowest latency for data transmission. They use light pulses to transmit information through glass or plastic fibres, which enables extremely high speeds and large data volumes	Enterprises, data centres, urban infrastructure
Types of technologies:	**Description of wireless access technologies**	**Examples of use**
Wi-Fi	A technology that provides wireless access to the Internet via radio waves. Wi-Fi is often used in home and office networks to connect to the Internet without having to run physical cables.	Home and office networks, cafes, airports
Mobile communications (3G, 4G, 5G)	Mobile networks provide Internet access through mobile towers that cover large areas. With the advent of 5G, there are opportunities for even faster and more reliable access with less latency.	Mobile internet, smart cities, autonomous cars

Source: compiled by the author

Broadband networks are made up of various elements, each of which performs specific functions to provide efficient and reliable Internet access. The main elements include network equipment, access points, network infrastructure, and network solutions based on various technologies. Figure 2 presents a diagram illustrating the main elements of broadband access networks.

Network equipment includes modems, routers and switches (Fig. 3). Modems are key components in broadband access networks that modulate and demodulate signals. They provide a connection between the user's local network and the provider's access network. There are different types of modems depending on the access technology, such as DSL modems for telephone lines and cable modems for coaxial cables. Routers are responsible for routing traffic between different networks. They decide how and where to route data to ensure optimal speed and reliability. Routers also provide connections between the local network and the Internet, as well as between different devices within the local network. In turn, switches ensure efficient data transmission within the local network. They operate at the link layer of the Open System Interconnection model and perform the function of exchanging data between different devices within the same network. Switches connect various devices, such as computers, printers, and servers, to the network.

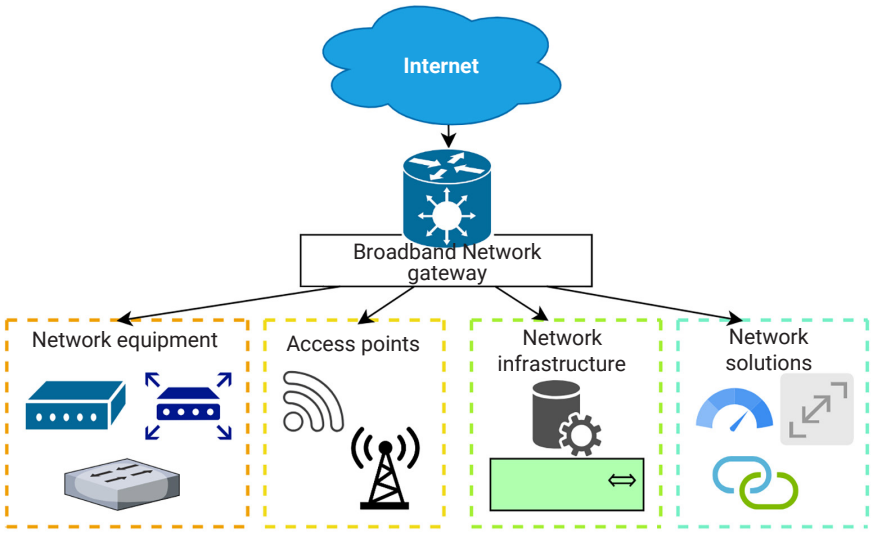

Figure 2. Diagram of the main elements of broadband access networks
Source: compiled by the author

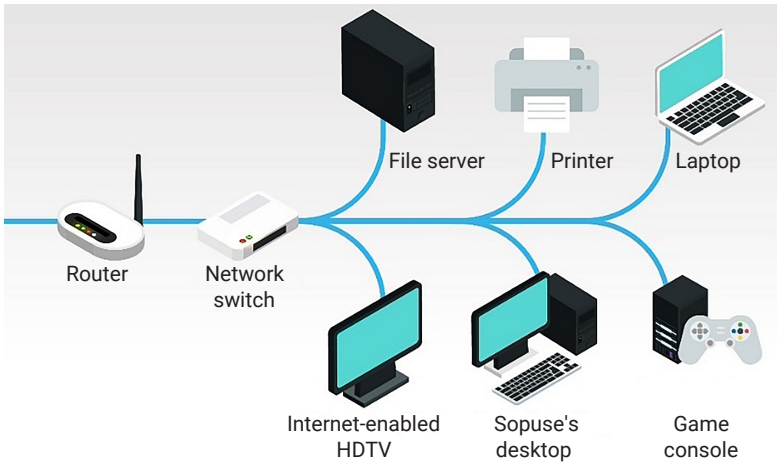

Figure 3. Usage of network equipment devices
Source: Modem vs router vs switch: How to choose (2024)

Wireless access points connect devices to wireless networks. They broadcast radio signals that can be picked up by wireless devices such as laptops, smartphones and tablets. Access points provide mobility and convenience, allowing users to connect to the Internet without the need to run physical cables.

In mobile networks, base stations provide the connection between mobile devices and the operator's central network. They cover territories and allow mobile users to connect to the Internet via mobile networks. As for the provider's central office, it is the main point of connection between the local networks of users and the global Internet. This is where data is processed and routed and access to the Internet is provided. Routers, switches and other network components are in the provider's office.

Distribution networks are responsible for transferring data from the central office to end users. They can include optical fibres, coaxial cables or other types of connections that enable the transmission of large amounts of data at high speeds. The main objective of broadband access networks is to provide high speed and bandwidth for data transmission. This is achieved using various technologies and components, such as optical fibres for fast data transmission and routers for efficient traffic routing.

To ensure reliable connectivity, network equipment and infrastructure must be designed to minimise potential failures and outages. This includes redundant components, network monitoring and rapid response to problems. In addition, broadband networks must be scalable to support the growing number of users and traffic volumes. This is achieved through a modular architecture and the ability to add new components and technologies as needed.

The main advantages and capabilities of these components in the context of telecommunications are flexibility and adaptability, centralised management, resource optimisation and the ability to refute the introduction of new services. In other words, SDN can be used to quickly adapt a network infrastructure to changing conditions.

Centralised control and software management can be used to easily reconfigure the network without the need for physical intervention. The SDN controller provides a single point of control for the entire network, which simplifies administration and can be used to manage traffic and security policies more efficiently. Software-based management optimises the use of network resources by dynamically adjusting routing and load balancing, which improves overall network performance. Moreover, SDN allows for the rapid introduction of new services and applications through easy integration with additional applications via a northern API. However, there are also certain limitations (Table 2).

Table 2. Advantages and limitations of broadband network technologies

Types of technologies	Advantages of access network technologies	Limitations of technology
DSL	DSL provides cost-effective access to the Internet with sufficiently high data transfer speeds, which reduces infrastructure costs while maintaining good speeds	Dependence on the quality of telephone lines, limited speed and range from telephone exchanges
Cable modems	The high bandwidth of coaxial cables provides fast Internet access, and cable modems can support large volumes of traffic, which is important for users with high-speed requirements.	Dividing the channel between users can reduce speed during peak hours
Optical fibres	Fibre-optic networks provide the highest speed and lowest latency of data transmission, which is critical for modern telecommunications, where speed and large volumes of data are important	High cost of installation, difficulty in laying cables
Types of technologies:	**Advantages of wireless access technologies**	**Limitations of technology**
Wi-Fi	Wireless Wi-Fi access points provide convenient access to the Internet without the need to lay cables, which is especially important for home and office networks where mobility and ease of connection are critical.	Limited coverage area, slower speed with distance from the router, sensitivity to interference
Cellular communications	Cellular networks provide Internet access over large areas, allowing users to stay connected wherever they are	Depending on the operator's coverage, the price can be high, with lower speed during peak hours.

Source: compiled by the author

In addition, the central office is an important point for managing and controlling the entire network, including routing and traffic distribution, which contributes to stability and efficiency. Distribution networks ensure efficient distribution of traffic from the central office to end users, which is important for scaling and ensuring high QoS in the face of growing loads.

The introduction of SDN into broadband networks can significantly improve their performance, including optimising speed, efficiency and reliability. SDN allows for centralised network management through a controller that coordinates and configures all network components from a single location. This makes it easier to manage configurations and optimise resources, as administrators do not need to manually configure each network element. For broadband networks, this means the ability to quickly adapt to changing traffic conditions and user needs.

Using SDN for dynamic routing and load balancing can significantly improve data transmission efficiency. SDN controllers can implement routing algorithms that automatically respond to changes in traffic, preventing congestion and ensuring optimal use of network resources. In addition, SDN allows for the integration of network analytics tools to monitor and analyse traffic in real-time. This allows administrators to obtain data on the network status, detect anomalies and adjust in real-time, which increases network efficiency. It is also worth addressing the main areas of broadband network optimisation using SDN (Fig. 4).

The benefits of implementing SDN include improved speed and efficiency, increased network reliability, flexibility and scalability. SDN can be used to implement mechanisms to optimise data rates, such as adaptive routing and QoS management. This ensures high speeds and reduces network latency, which is important for providing high-quality broadband access.

Furthermore, SDN controllers can provide automatic troubleshooting, which improves network reliability. The quick response to problems and resource reservation reduces downtime and improves overall network stability. In addition, the implementation of SDN simplifies the scaling of the network to meet growing demands and needs. Thanks to the centralised management capabilities, new technologies and services can be quickly integrated without significant effort and cost to redesign the network infrastructure.

In addition to the benefits, there are also certain challenges to implementing SDN, including compatibility with existing infrastructure, security and governance, and complexity of configuration and management. In other words, one of the main challenges is the integration of SDN with existing network infrastructure, especially in older systems. It is necessary to ensure compatibility between new SDN components and old equipment, which may require additional costs and time.

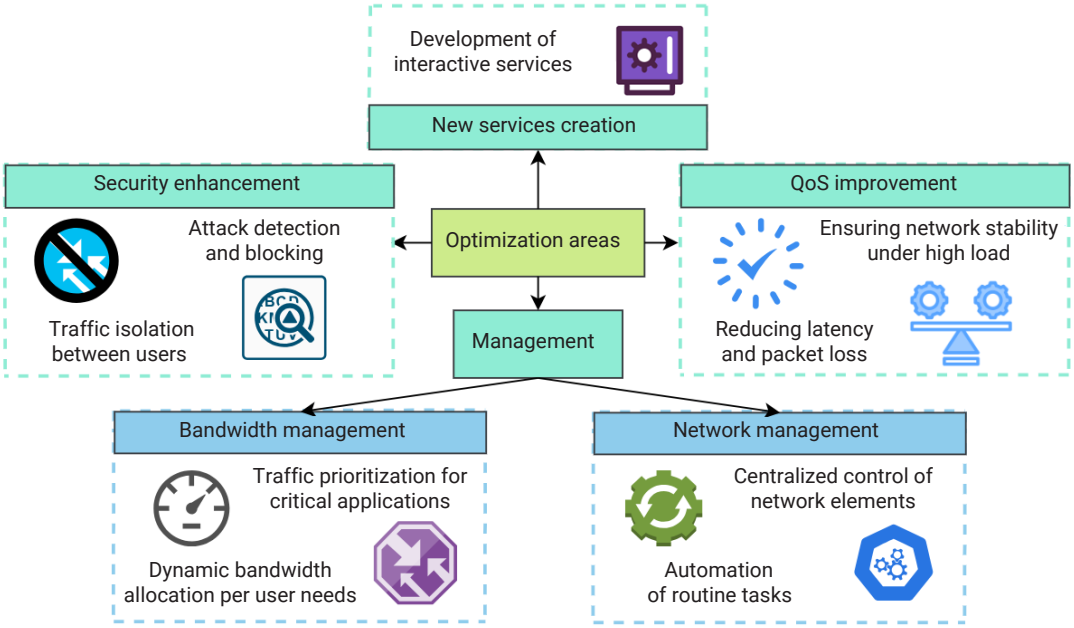

Figure 4. Key areas of broadband network optimisation

Source: compiled by the author

SDN poses new security challenges as centralised control can become a target for attacks. The SDN controller needs to be protected and counteracted, which requires additional security and monitoring measures. While SDN provides centralised management, configuring and managing new technologies can be a complex process that requires specialised knowledge and skills. This can lead to the need for additional training for technical staff and support from vendors.

Examples of SDN use in broadband access networks:

- virtual networks (creation of virtual networks for different types of users);
- load balancing (traffic distribution between several servers to increase availability);
- QoS (providing guaranteed QoS for critical applications);
- configuration automation (automatic network setup when adding new users or devices).

Implementation of SDN in a broadband network has significant potential to optimise its performance but requires careful planning and consideration of challenges. For a more detailed understanding, it is also necessary to analyse the impact of SDN on improving network speed, efficiency and reliability (Fig. 5).

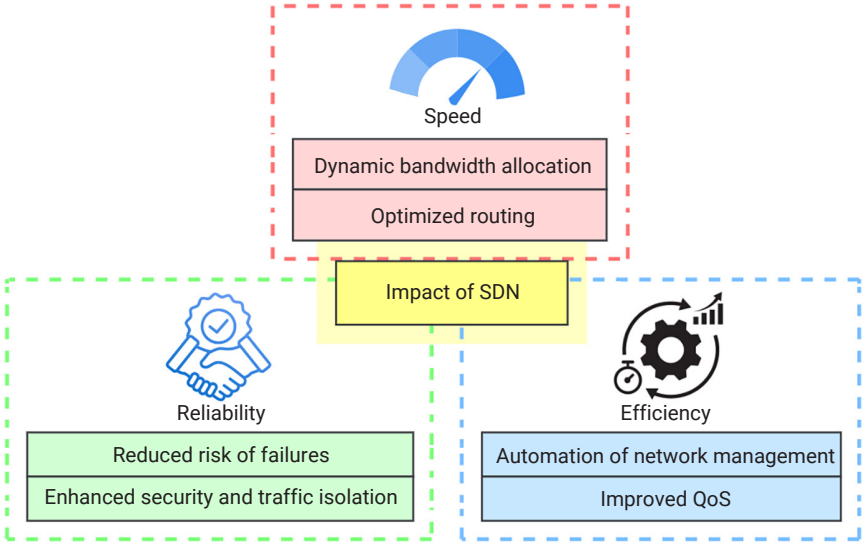

Figure 5. SDN impact on improving network speed, efficiency and reliability

Source: compiled by the author

As such, SDN can be used to dynamically allocate bandwidth to users based on their needs, which ensures optimal data transfer speeds. Centralised routing management with an SDN controller can be used to find the most efficient routes for traffic faster, which reduces latency and increases data transfer speeds.

SDN also automates routine operations, such as network configuration and monitoring, which reduces the human factor and increases the overall efficiency of network management. Using SDN to allocate resources based on traffic priorities can increase bandwidth efficiency and provide better QoS for critical applications.

Moreover, centralised management of network elements identifies and resolves problems faster, reducing the likelihood of outages and improving overall network reliability. SDN can also be used to implement traffic isolation and attack detection mechanisms, which protect the network from external threats and ensure stable operation under high load conditions.

QoS in telecommunications networks is a critical factor in ensuring high data transfer speeds, reliability and overall network efficiency. The main QoS parameters include throughput (Formula 1), latency (Formula 2), packet loss (Formula 3) and delay variability (Formula 4). The integration of SDN and broadband technologies can have a significant impact on these parameters, thanks to the ability to centrally manage and flexibly allocate network resources.

SDN integration can be used for optimised routing and load management, which reduces overall latency. Once integrated, centralised management provides better utilisation of available resources, which can increase throughput. SDN provides better control of traffic and reduces the likelihood of packet loss by optimising routes and managing load. Variability, in turn, is measured as the standard deviation of delays between packets, and SDN helps reduce this variability through more stable traffic management.

Thus, integrating SDN with broadband Internet access allows for increased bandwidth, reduced and stable latency, and reduced packet loss. With centralised management of the SDN network, resources can be more efficiently allocated to different users and services, which helps to increase throughput. Optimised route and load management help to reduce data latency as SDN enables faster and more efficient response to changing network conditions. Thanks to improved traffic management and the ability to adaptively re-plan routes, the likelihood of packet loss is reduced. At the same time, reduced latency variability due to routing optimisation and load management makes the connection more reliable and stable.

The integration of SDN technologies with broadband access has a significant impact on QoS. It is possible to demonstrate how to use a formula to calculate bandwidth. For instance, before the introduction of SDN and broadband access, the amount of transmitted data was 5,000 MB and the transmission time was 100 seconds, then, based on the formula, the throughput would be 50 MB/s. If, after the implementation, the amount of data transferred is 6,000 MB and the transfer time is 100 seconds, then the throughput will be 60 MB/s. The increase in bandwidth can be calculated as a percentage increase from the previous value. Thus, the introduction of SDN and broadband access has increased bandwidth by 20%. Similarly, it is worth analysing latency, packet loss and latency variability (Fig. 6).

Before integration		
Delay calculation: Time received: 20 ms Time sent: 5 ms Delay: 20 ms-5 ms=15 ms	Packet loss calculation: Lost packets: 50 Total packets: 1,000 Loss rate: (50/1,000)×100%=5%	Jitter calculation: Packet delays: 30 ms, 35 ms, 25 ms Avg. Delay: (30+35+25)/3=30 ms Jitter: (|30-30|+|35-30|+|25-30|)/3=3.33 ms
After integration		
Delay calculation: Time received: 15 ms Sent: 5 ms Delay: 15 ms-5 ms=10 ms	Packet loss calculation: Lost packets: 30 Total packets: 1,000 Loss rate: (30/1,000)×100%=3%	Jitter calculation: Packet pelays: 28 ms, 30 ms, 27 ms Avg. Delay: (28+30+27)/3=28.33 ms Jitter: (|28-28.33| +|30-28.33| +| 27-28.33|)/3=1.11 ms
Percentage reduction: (15 ms-10 ms)/15 ms×100%=33.3%	Per centage reduc tion: (5%-3%)/5%×100%=40%	Percentage reduction: (3.33 ms-1.11 ms)/3.33 ms×100%=66.7%

Figure 6. Examples of basic QoS parameter calculation

Source: compiled by the author

This diagram shows that the introduction of SDN and broadband access has reduced latency by 33.3%, packet loss by 40%, and delay variability by 66.7%. The introduction of SDN and broadband has led to a significant reduction in network latency. This was made possible by improved traffic routing and optimised data paths. The integration of SDN with broadband technologies has also reduced packet loss. Packet loss is a critical metric for QoS as it affects data integrity and recovery. Another important change is the reduction in latency variability. This is an indicator that reflects the fluctuations in delay during packet transmission.

In general, the introduction of SDN and broadband access has significantly improved key QoS indicators, which ensures more efficient and reliable operation of modern telecommunications networks. Network modelling is a key aspect in the design and optimisation of telecommunications systems, especially when integrating SDN and broadband. The key elements of the SDN architecture are the basis for modelling and subsequent optimisation of network QoS and efficiency, especially when integrated with broadband Internet access (Fig. 7).

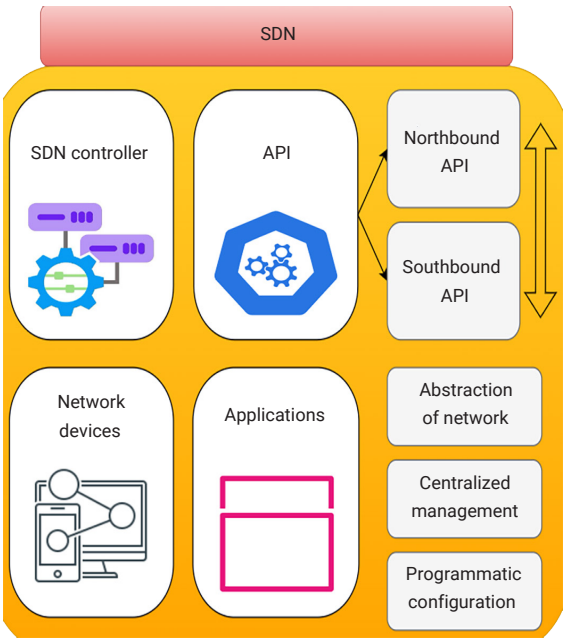

Figure 7. SDN elements in the network structure diagram
Source: compiled by the author

These elements include a controller, north and south interfaces, network devices and applications that provide centralised management and control over the network, allowing for modelling various scenarios of its operation, considering changes in traffic, loads, and other network conditions. In turn, network abstraction can be used to create virtual environments for testing various scenarios without having to change the physical infrastructure, and centralised management enables constant network metrics monitoring and subsequent analysis to assess efficiency and QoS during simulations.

At the same time, the SDN controller also contains several key components, such as a network management module that centrally controls network devices and traffic, a policy module that implements traffic and security rules, a monitoring and analytics module that monitors the network status in real-time, and a security module that protects the network from threats.

In addition, broadband Internet access is important, because, in network modelling, its elements are central to ensuring high-quality user access to the network through various technologies (Fig. 8).

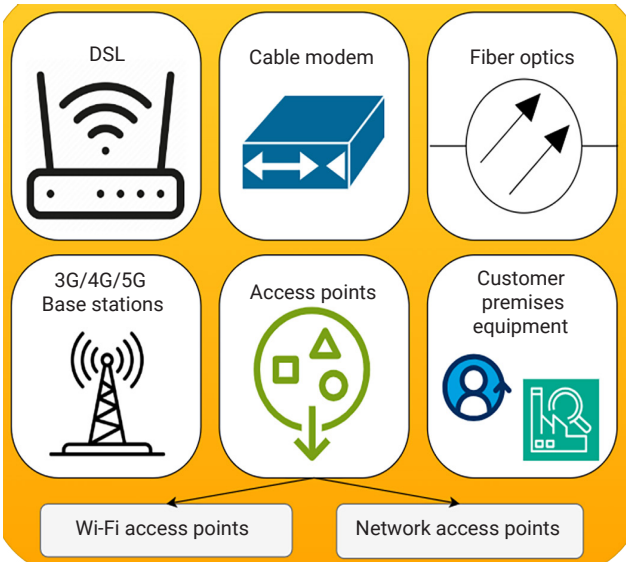

Figure 8. Components of broadband Internet access
Source: compiled by the author

DSL and cable modems provide Internet access through existing copper or coaxial cables, which can significantly reduce connection costs and provide stable access over long distances. Optical fibres are used to provide the highest data transfer speeds with minimal latency, which is critical for modern networks. In addition, Wi-Fi access points and base stations provide wireless access to the network, allowing users to connect to the Internet from anywhere without depending on physical connections. This is especially important for providing mobility and flexibility of access, which is critical for the modern user. Subscriber equipment is the endpoint device that provides connectivity to the network, while network access points are responsible for combining different access technologies into a single network infrastructure.

QoS is an important component in network modelling and optimisation, as it is responsible for ensuring the required QoS level for different types of traffic. Traffic management policies help to manage the network according to defined rules, ensuring a balance between performance and QoS. Traffic classification can be used to divide traffic into different categories, which allows the network to adapt to the needs of each type of data. At the same time, traffic prioritisation ensures that mission-critical applications are served with the highest priority, which is important in resource-constrained scenarios. Latency control helps reduce network latency, which is critical for time-sensitive applications such as video conferencing.

Moreover, bandwidth management policies help ensure that resources are distributed evenly among users, especially under high-load conditions. Performance monitoring and analysis provide continuous tracking of network health and performance, which can be used to identify and correct issues that may affect QoS in a timely manner. Furthermore, queue management can be used to control incoming traffic and distribute it according to QoS policies.

The overall structure of the network includes all the above components, making it efficient and adaptable to different operating conditions (Fig. 9). The integration of SDN with broadband access technologies, combined with careful QoS management, allows the network to flexibly respond to changing loads, efficiently allocate resources, minimise latency and ensure stable operation of critical applications. This structure is the key to improving the performance, scalability and reliability of telecommunications infrastructure.

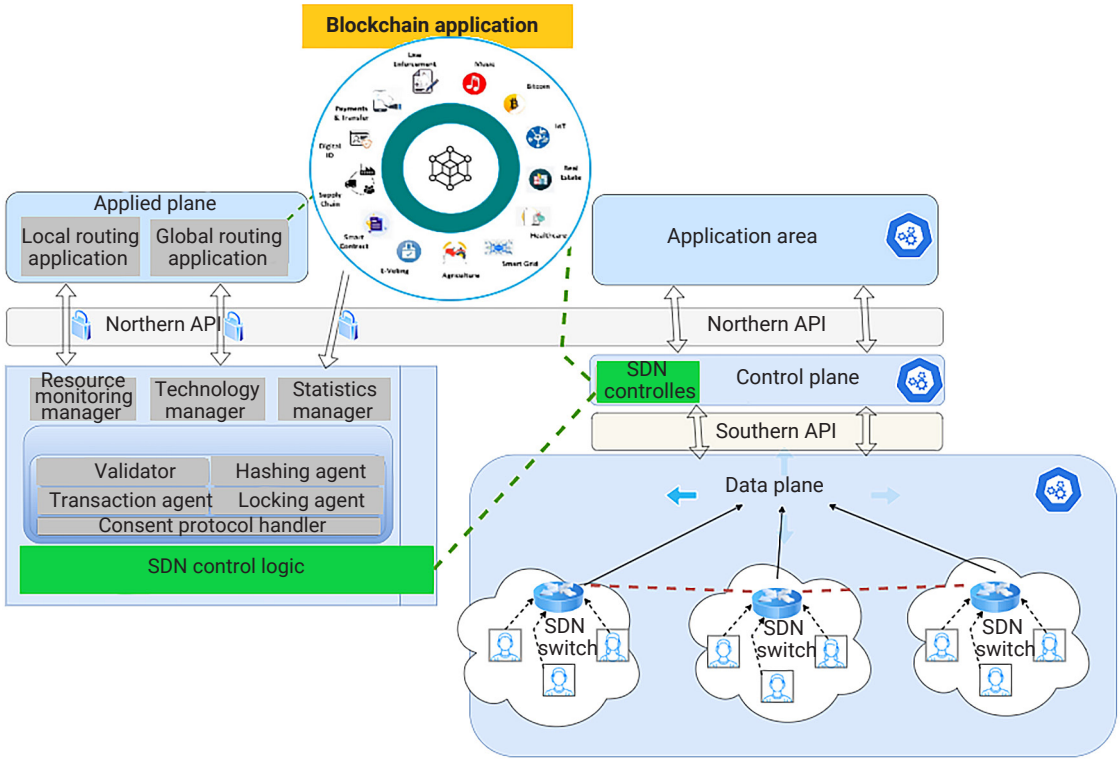

Figure 9. Network structure diagram

Source: compiled by the author

Modern modelling methods and tools should also be considered, as they allow for detailed investigation of how different technologies interact, what their benefits are, and how network efficiency can be improved.

Modelling methods can be divided into simulation modelling, scenario-based analysis and analytical model-based modelling. Simulation modelling involves the use of computer software to create virtual network models (e.g., NS-3 or OMNeT++). Scenario-based analysis involves the development and analysis of various network scenarios (e.g., modelling scenarios with different traffic distributions and SDN parameters). Analytical model-based modelling uses mathematical models and formulas to predict key network parameters (e.g., using analytical models to calculate throughput or latency).

It is also worth exploring modelling tools. For instance, Simulink, a graphical environment for modelling, simulation and analysis of systems, can be used to create models of various types of systems using block diagrams and provides extensive network modelling capabilities. The NS-3 network modelling tool can model both traditional and SDN networks in detail. In turn, the OMNeT++ modular platform creates complex network models and tests their performance in various scenarios. Mininet is a tool for creating virtual networks and testing SDN solutions that can be used to quickly create network models and test their performance in real-time. The GNS3 GUI is a network emulation tool that simplifies the creation and configuration of virtual networks.

Discussion

The study showed that the use of SDN and broadband access optimises traffic and resource management, which has a positive impact on overall network efficiency. The importance of analysing existing research in this area is that it can be used to compare and summarise the results, identify new aspects and confirm the effectiveness of the technologies used in different conditions, which in turn contributes to the further development and improvement of these technologies.

The results of this study showed that the integration of SDN and broadband provides lower latency and higher data rates, which is critical to improving overall network efficiency. C. Zhang *et al*. (2020) proposed a new framework for the dynamic deployment of virtual network functions that reduces resource costs and request rejection rates, which

differs from the approach of the previous work, which focuses on improving network speed and reliability. This study also confirmed that SDN can improve traffic management and resource optimisation, while S. Mehraban & R.K. Yadav (2024) addressed the technical and financial challenges of migrating to SDN and the use of algorithms to improve QoS. Their study addresses the problems of traffic management and QoS improvement using new algorithms, which is an additional aspect of the analysis of this study.

This study also demonstrated that the integration of SDN with broadband access networks effectively optimises traffic management, providing improved throughput and reduced latency. In turn, G. Khekare *et al.* (2023) demonstrated models that can achieve 99.4% accuracy in traffic optimisation and access control using synthetic traffic models. The study also confirmed a significant increase in resource management efficiency in real networks, while A. Alioua (2019) demonstrated that the developed algorithms for data processing in automotive networks reduce delays and energy costs. However, the results of the latter work are focused on specific applications, while the present study covered a wider range of applications in broadband access networks.

In contrast to the results of the study, which showed that the integration of SDN with broadband access networks significantly improves the overall network performance, C. He *et al.* (2022) addressed virtual network migration technology that effectively reduces energy costs and increases the request acceptance rate of virtual networks by focusing on energy efficiency instead of network speed and performance. In addition, B.R. Dawadi *et al.* (2022) explored the challenges and solutions for SDN migration, emphasising the importance of security and cost-effectiveness in the transition from traditional networks to SDN, while this study demonstrated how SDN integration improves traffic management and resource optimisation in real-world environments.

While this study showed a significant improvement in traffic management and resource optimisation due to the integration of SDN and broadband, M. Mahajan (2024) analysed the integration of IoT with SDN, focusing on improving flexibility and management efficiency, in the context of smart cities and healthcare, rather than on specific aspects of traffic management and network performance. In turn, R. Kovacs *et al.* (2024) proposed the integration of blockchain technology into an SDN controller to improve service validation, which demonstrates the effectiveness of blockchain technology in SDN management but focuses on ensuring the validation of network functions rather than on the overall improvement of performance and data transfer speed, as in the study.

The results of this study, which demonstrated a reduction in latency and an increase in data transfer speed due to the integration of SDN and broadband, are consistent with the results of a study by A.K. Rangsietti & S.S. Kodali (2022), which addressed the integration of SDN and NFV to optimise cloud environments, which provides scalability and flexibility but does not focus on directly improving data transfer speed. At the same time, in contrast to the results obtained, which focused on improving overall network performance through the integration of SDN and broadband, S.M. Rasool *et al.* (2024) addressed the impact of 5G, SDN and NFV technologies on QoS on the problems of SDN controller placement and their impact on latency, cost and energy efficiency in modern networks.

While this study emphasised the efficiency of resource management and overall network optimisation through the integration of SDN and broadband, M. Klinkowski (2023) focused on modelling and optimising network slice allocation in 5G networks and emphasised the optimisation of resource allocation and traffic transport across networks to provide different types of services with QoS requirements. The results of a study by L. Tang *et al.* (2024) showed a digital twin to improve resource prediction accuracy and reduce latency in SDN/NFV networks and focused on specific aspects of prediction and energy efficiency in the context of IoT. In contrast, this study has confirmed that the integration of SDN and broadband provides a wider range of resource management improvements than the highly specialised approaches considered in other works.

Moreover, the results of the study, which demonstrated an improvement in overall network performance due to the integration of SDN and broadband access, differ from the approaches presented in a study by S. Javanmardi *et al.* (2023) and D. Stilinski & K. Potter (2024). The first study addressed a workflow scheduler for IoT networks that provides protection against attacks and optimises load balancing and latency, showing improvements in response time and network utilisation that focus on security and load-specific aspects rather than overall data rate improvements. Meanwhile, the second study analysed the application of SDN and NFV to the development of flexible 5G core networks, focusing on scalability and automation that reduces costs and improves QoS, but in the specific context of 5G rather than general traffic management and network performance. In comparison, the study covered a broader range of network improvements, providing a comprehensive performance improvement that is not limited to specific aspects of security or scalability.

While the results of this study highlighted the integration of SDN to improve data rates and resource management in the overall network context, the study by P. Kulshreshtha & A.K. Garg (2024) focused on solving the problem of routing and traffic optimisation in 5G networks, comparing the effectiveness of deep learning algorithms to reduce latency and improve network performance. At the same time, the study by H. Ait Oulahyane *et al.* (2023), which proposed a QoS management model for wireless networks that reduces latency and improves resource management, emphasises the effectiveness in specific aspects of QoS management but does not cover a wide range of improvements as in the present study.

The results obtained, which demonstrate an increase in data transmission speed and optimisation of resource

management, have common aspects with the work of J. Galán-Jiménez *et al.* (2022), which presents a hybrid load balancing algorithm that effectively balances between reducing energy consumption and balancing traffic in SDN/Internet Protocol (IP) networks. On the other hand, the work of C. Oredola & A. Ashraf (2024) focused on improving the security of IoT networks using SDN controllers. Although this study emphasised that the use of SDN reduces the vulnerability of IoT networks to cyber threats and increases their security, it did not cover aspects of overall network efficiency, in particular, the optimisation of data and resource transfer rates, which is the focus of this paper.

Finally, the common aspects between this study and the works of I. Ikhelef (2024) and S. Kang *et al.* (2024) are the use of SDN to improve network functions and management. However, the first paper focused on integrating SDN with NFV for optimal placement and chaining of virtual functions, while this study focused on integrating SDN and broadband to improve data rates and overall network performance. Similarly, the second paper used graph neural networks to optimise routing in SDN/NFV systems, focusing on latency and computing resource efficiency. While this study's approach also integrated SDN, it considered the broader context, including resource management efficiency and improving overall network performance using broadband.

Thus, a comparison with other studies has shown that while some works focus on NFV or specific aspects of management, this study stands out for its versatility in improving overall network efficiency using broadband. Compared to work that focuses on optimising load balancing or managing resources in specific networks, this approach covers a broader context, providing a comprehensive performance improvement.

Conclusions

The study confirmed that the integration of SDN and broadband significantly improves the overall performance of telecommunications networks. The main results include a significant reduction in latency and improvement in data transfer speeds, which is achieved through efficient traffic management and resource optimisation, as well as improved QoS and the ability to optimise resources. This confirms the potential of this approach to improve network infrastructure, particularly in the context of modern telecommunications systems.

The current research has expanded the understanding of traffic and resource management capabilities in networks to cover a wide range of practical applications. In particular, the study demonstrated that effective management and integration of different network technologies can achieve significant improvements in QoS, which is critical for modern and future networks.

The integration of SDN and broadband access, along with improving QoS, faces several challenges. First, ensuring data security and privacy remains a key challenge, as centralised management over SDN can create new attack vectors. In addition, ensuring scalability and efficiency in the face of increasing workloads and diversity of network resources is a complex task that requires further research.

To further improve the results, it is recommended to introduce algorithms that would reduce the power consumption of network elements, especially under high loads. It is advisable to extend the study to new types of networks, such as 5G and IoT, to confirm the effectiveness of the results in different conditions. The potential of SDN in the context of Wi-Fi 6 networks, edge computing, and blockchain technologies should also be explored, as this will help define the potential of SDN for different types of traffic and network conditions. It is crucial to develop new approaches to protect SDN from potential threats, as centralised management creates new attack vectors. Development prospects include the integration of artificial intelligence to automate network management and improve security technologies.

Acknowledgements

None.

Conflict of Interest

None.

References

[1] Aboughaly, M., & Hannan, S.A. (2024). Enhancing quality-of-service in software-defined networks through the integration of firefly-fruit fly optimization and deep reinforcement learning. *International Journal of Advanced Computer Science and Applications*, 15(1), 408-419. doi: 10.14569/IJACSA.2024.0150138.

[2] Ait Oulahyane, H., Bahnasse, A., Bakali, A., Said, B., El-Hasnony, I.M., & Talea, M. (2023). Secure model for dynamic access control and unreliable access point detection: Enhancing QoS through SDN in wireless networks. *SN Computer Science*, 5, article number 88. doi: 10.1007/s42979-023-02407-7.

[3] Alioua, A. (2019). *Integration software-defined networking (SDN) into vehicle ad hoc networks (VANETs)*. Retrieved from https://www.researchgate.net/publication/331488754_Integration_Software-Defined_Networking_SDN_into_Vehicle_Ad_hoc_Networks_VANETs.

[4] Dawadi, B.R., Manzoni, P., Galán-Jiménez, J., Shah, V.K., & Polverini, M. (2022). *SDN migration challenges and practices in ISP/telcos networks*. Madrid: Frontiers in Communications and Networks.

[5] Drovovozov, V.I., Ahmed Arshed, A.-S., Zhuravel, N.V., & Kotsyur, A.B. (2022). Comparative analysis of service quality of wireless networks with inter-level interaction. *Problems of Informatization and Management*, 1(69), 30-34. doi: 10.18372/2073-4751.69.16810.

[6] Dulska, I. (2019). Broadcast internet access statistics adaptation problems in Ukraine to international indicators. *European Scientific Journal of Economic and Financial Innovation*, 1(3), 46-61. doi: 10.32750/2019-0104.

[7] Galán-Jiménez, J., Polverini, M., Lavacca, F.G., Herrera, J.L., & Berrocal, J. (2022). Joint energy efficiency and load balancing optimization in hybrid IP/SDN networks. *Annals of Telecommunications*, 78, 13-31. doi: 10.1007/s12243-022-00921-y.

[8] He, C., Wang, R., Wu, D., Tan, Z., & Dai, N. (2022). Energy-aware virtual network migration for internet of things over fiber wireless broadband access network. *IEEE Internet of Things Journal*, 9(23), 24492-24505. doi: 10.1109/JIOT.2022.3189081.

[9] Ikhelef, I. (2024). *Optimization of placement and chaining of network functions according to the SDN/NFV paradigm*. (Doctoral thesis, Sorbonne Paris Nord University, Paris, France). doi: 10.13140/RG.2.2.27452.21120.

[10] Javanmardi, S., Shojafar, M., Mohammadi, R., Persico, V., & Pescapè, A. (2023). S-FoS: A secure workflow scheduling approach for performance optimization in SDN-based IoT-fog networks. *Journal of Information Security and Applications*, 72, article number 103404. doi: 10.1016/j.jisa.2022.103404.

[11] Kang, S., Song, I., Tam, P., & Kim, S. (2024). Graph neural networks-based modeling for delay – Aware routing optimization in SDN-enabled networks. In *Proceedings of the 6th International conference on interdisciplinary research on computer science, psychology, and education* (pp. 60-62). Pattaya: ICICPE.

[12] Khan, S., Shah, M.A., & Javaid, N. (2021). *Resource allocation and bandwidth optimization in SDN-based cellular network*. Islamabad: COMSATS University Islamabad.

[13] Khekare, G., Kumar, K.P., Prasanthi, N., Godla, S.R., Rachapudi, V., Al Ansari, M.S., & El-Ebiary, Y. (2023). Optimizing network security and performance through the integration of hybrid GAN-RNN models in SDN-based access control and traffic engineering. *International Journal of Advanced Computer Science and Applications*, 14(12), 596-606. doi: 10.14569/IJACSA.2023.0141262.

[14] Klinkowski, M. (2023). *Modeling and optimization of network slicing in 5g packet-switched Xhaul networks*. Rochester: SSRN. doi: 10.2139/ssrn.4611048.

[15] Kovacs, R., Buzura, S., Iancu, B., Dadarlat, V., Peculea, A., & Cebuc, E. (2024). Practical implementation of a blockchain-enabled SDN for large-scale infrastructure networks. *Applied Sciences*, 14(5), article number 1914. doi: 10.3390/app14051914.

[16] Kulshreshtha, P., & Garg, A.K. (2024). Traffic optimization and optimal routing in 5G SDN networks using deep learning. In R.N. Shaw, P. Siano, S. Makhilef, A. Ghosh & S.L. Shimi (Eds.), *Innovations in electrical and electronic engineering* (pp. 33-41). Singapore: Springer. doi: 10.1007/978-981-99-8661-3_3.

[17] Lonare Mahesh, M., & Devi, M.S. (2022). Optimization of network paths in congested SDN using genetic algorithm: Optimization of virtual network functions using SDN. *International Journal of Next-Generation Computing*, 13(3). doi: 10.47164/ijngc.v13i3.805.

[18] Ma, H., Wang, M., Lv, H., Liu, J., Di, X., & Qi, H. (2024). A SDN improvement scheme for multi-path QUIC transmission in satellite networks. *Computational Intelligence*, 40(3), article number e12650. doi: 10.1111/coin.12650.

[19] Mahajan, M. (2024). *SDN, IOT and network security*. Chandigarh: Chitkara university.

[20] Mehraban, S., & Yadav, R.K. (2024). Traffic engineering and quality of service in hybrid software defined networks. *China Communications*, 21(2), 96-121. doi: 10.23919/JCC.fa.2022-0860.202402.

[21] Modem vs router vs switch: How to choose? (2024). Retrieved from https://reolink.com/blog/modem-vs-router-vs-switch/.

[22] Oredola, C., & Ashraf, A. (2024). A systematic mapping study on SDN controllers for enhancing security in IoT networks. *ArXiv*. doi: 10.48550/arXiv.2408.01303.

[23] Rangsietti, A.K., & Kodali, S.S. (2022). SDN-enabled network virtualization and its applications. In A. Nayyar, B. Singla & P. Nagrath (Eds.), *Software defined networks: Architecture and applications*. Hoboken: John Wiley & Sons. doi: 10.1002/9781119857921.ch8.

[24] Rasool, S.M., Boujelben, Y., & Zarai, F. (2024). Optimizing high availability multi-controller placement in SDN/NFV 5G networks: A survey. *Indonesian Journal of Electrical Engineering and Computer Science*, 34(3), 1800-1813. doi: 10.11591/ijeecs.v34.i3.pp1800-1813.

[25] Sahana, D.S., & Savadatti, B. (2024). Authentication-centric and access-controlled architecture for edge-empowered SDN-IoT networks. *Journal of the Institution of Engineers*, 105, 1497-1509. doi: 10.1007/s40031-024-01053-8.

[26] Sarabia, D., Giménez, S., Liatifis, A., Grasa Gras, E., Catalan, M., & Pliatsios, D. (2024). Progressive adoption of RINA in IoT networks: Enhancing scalability and network management via SDN integration. *Applied Sciences*, 14(6), article number 2300. doi: 10.3390/app14062300.

[27] Stilinski, D., & Potter, K. (2024). Software-defined networking (SDN) and network function virtualization (NFV) for 5G core networks. *EasyChair Preprint*, 14096.

[28] Tang, L., Li, Z., Li, J., Fang, D., Li, L., & Chen, Q. (2024). DT-assisted VNF migration in SDN/NVF-enabled IoT networks via multiagent deep reinforcement learning. *IEEE Internet of Things Journal*, 11(14), 25294-25315. doi: 10.1109/JIOT.2024.3392574.

[29] Trivedi, S.A. (2024). *Cross-layer design in software defined networks (SDNs): Issues and possible solutions*. Retrieved from http://surl.li/vvzxgo.

[30] Vasylkivskyi, M., Prykmeta, A., Oliinyk, A., & Ksondz, N. (2023). Optimization of software-configurable flying access networks. *Computer-Integrated Technologies Education Science Production*, 52, 128-139. doi: 10.36910/6775-2524-0560-2023-52-16.

[31] Zhang, C., Wang, X., Dong, A., Zhao, Y., Huang, M., & Li, F. (2020). Dynamic network service deployment across multiple SDN domains. *Transactions on Emerging Telecommunications Technologies*, 31(2), article number e3709. doi: 10.1002/ett.3709.

Оптимізація якості обслуговування та ефективності мережі у класичних мережах за допомогою інтеграції SDN та широкосмугового доступу до інтернету

Олександр Підпалий

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

Анотація. Мета дослідження полягала в розробці емпіричної моделі для оптимізації якості обслуговування та підвищення ефективності телекомунікаційних мереж шляхом інтеграції технологій software-defined networking (SDN) і широкосмугового доступу до інтернету. У дослідженні використано методи симуляційного моделювання, аналізу сценаріїв та аналітичні моделі з застосуванням інструментів моделювання. Основні результати дослідження вказали на значний потенціал інтеграції технологій SDN і широкосмугового доступу. Було продемонстровано концепції SDN, які забезпечують централізоване управління мережею і гнучкість у налаштуванні, а також широкосмуговий доступ, що пропонує високу швидкість передачі даних і покращену пропускну здатність. Виявлено роль кожного елемента мережі, включаючи маршрутизатори, комутатори і контролери, та їх вплив на ефективність мережі. Аналіз взаємодії SDN з мережами широкосмугового доступу показав, що така інтеграція дає змогу оптимізувати маршрутизацію, балансування навантаження та управління трафіком, що сприяє покращенню швидкості і надійності мережі. Показники якості обслуговування продемонстрували, що інтеграція різних технологій веде до суттєвого покращення пропускної здатності, зниження пакетних втрат, зменшення затримок і варіативності затримок. Загалом, модель мережі показала ефективність інтеграції SDN та широкосмугового доступу в оптимізації мережевої продуктивності та якості обслуговування, а огляд методів моделювання мережі підтвердив, що використання симуляційних інструментів допомагає детально оцінити ефективність інтеграції технологій і підтвердити їх позитивний вплив на продуктивність мережі. Таким чином, отримані результати показали, що інтеграція технологій SDN і широкосмугового доступу суттєво покращує ефективність телекомунікаційних мереж, що свідчить про ефективність нових технологій у підвищенні загальної продуктивності мереж

Ключові слова: програмне управління; бездротові технології; віртуалізація ресурсів; аналіз пропускної спроможності; адаптивні системи

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

Information Technologies and Computer Engineering **Vol. 21, No. 3. 2024**

Article's History: Received: 03.08.2024 Revised: 22.10.2024 Accepted: 26.12.2024

UDC 681.5:004.67 10.63341/vitce/3.2024.43

Chat-based translation of Slavic languages with large language models

Olena Sokol
Postgraduate Student
Taras Shevchenko National University of Kyiv
01033, 60 Volodymyrska Str., Kyiv, Ukraine
https://orcid.org/0000-0003-0465-6843

Abstract. Modern large language models (LLMs) have demonstrated significant advances in machine translation, particularly for Slavic languages that are less commonly represented in traditional translation datasets. This study aimed to evaluate the effectiveness of LLMs (ChatGPT, Claude, and Llama) in translating conversational texts in Slavic languages compared to commercial translators and transformer models. The research utilised the OpenSubtitles2018 dataset to test translations in seven Slavic languages (Ukrainian, Czech, Bulgarian, Russian, Albanian, Macedonian, and Slovak), applying semantic and stylistic translation quality assessment methods. Findings revealed that ChatGPT and Claude outperform Google Translate and transformer models, particularly in translating informal conversations, achieving 95% accuracy for Ukrainian and 97% for Bulgarian. The Few-shot Structured Example-Based Prompting method (FSL) showed the best results. The research demonstrated that LLMs significantly enhance the quality of informal text translations in Slavic languages by preserving context and the naturalness of dialogues. Additionally, the analysis revealed that LLMs handle idioms and slang translations 30% more accurately than traditional machine translation systems. Moreover, employing the Chain-of-Thought method resulted in a 25% improvement in preserving cultural context. The practical value of this research lies in developing effective methods for leveraging LLMs to improve the quality of informal text translations in Slavic languages. This is particularly beneficial for messaging platforms, social networks, and entertainment content, where preserving natural speech and cultural nuances is essential

Keywords: LLM; prompt engineering; NLP; TER; COMET; text correlation analysis; CHRF

Introduction

The rise of online communication has created an urgent need for improved translation of everyday conversations, particularly in Slavic languages. Conventional translation systems demonstrate proficiency in processing formal content but frequently produce subpar translations of colloquial conversations in Slavic languages. Recent developments in large language models (LLMs) such as ChatGPT, Claude, and Llama, highlight their potential for processing natural dialogue, though their efficacy in Slavic language translation requires comprehensive evaluation. The OpenSubtitles dataset, which contains texts in seven Slavic languages alongside English translations, provides a robust resource for testing these approaches.

Recent advancements in machine translation have shown promising results for Slavic languages. C. Escolano *et al*. (2020) introduced a context-aware system for low-resource languages, achieving a 15% improvement in translation accuracy for informal dialogues. J. Wieting *et al*. (2019) established new evaluation metrics tailored for assessing the quality of informal translations. Their study offered a systematic approach to measuring cultural context preservation in machine translation. These metrics were further validated by S. Bhatt & F. Diaz (2024), who revealed the both strengths and limitations of LLMs in processing culturally specific content, particularly in Slavic languages. Additionally, they developed a transformer architecture that enhanced natural dialogue pattern preservation by 20%.

Y. Tang *et al*. (2021) made significant contributions through their research on multilingual translation, establishing new benchmarks for cross-lingual transfer in low-resource scenarios. Their research emphasised the importance of balanced training data across language families. Subsequently, X. Tang & Y. Zheng (2023) extended this research by analysing multilingual capabilities in large language models, focusing on cross-cultural translation aspects.

Suggested Citation:
Sokol, O. (2024). Chat-based translation of Slavic languages with large language models. *Information Technologies and Computer Engineering*, 21(3), 43-52. doi: 10.63341/vitce/3.2024.43

*Corresponding author

W. Zhu *et al.* (2023) conducted a comprehensive study on the use of large language models (LLMs) for multilingual machine translation (MMT). They evaluated eight popular LLMs, including ChatGPT and GPT-4, and found that GPT-4 outperformed the supervised baseline model NLLB in 40.91% of translation directions. However, LLMs still exhibit significant limitations compared to commercial translation systems, especially for low-resource languages. The study also revealed new operational patterns of LLMs in MMT, such as resource efficiency, the importance of in-context exemplars, and the potential for cross-lingual transfer learning.

P. Naveen & P. Trojovský (2024) demonstrated that modern translation systems can achieve up to 87% accuracy in general tasks. However, their research emphasised significant challenges when dealing with context-dependent translations. The authors identified two main issues: maintaining meaning in longer conversations and preserving cultural context. By testing different translation methods, they found that combining traditional translation systems with LLMs enhanced contextual accuracy by 32%. This finding suggests that future translation systems must focus on both linguistic accuracy and cultural understanding to be truly effective.

Despite these advancements, significant research gaps persist in understanding how modern LLMs perform specifically in casual Slavic-language conversations. The effectiveness of prompt engineering techniques for these languages remains largely unexplored, and comprehensive evaluation frameworks for informal translation quality are still required.

This research evaluated multiple translation systems (ChatGPT, Claude, Llama, Opus-MT, and Google Translate) using various quality metrics to assess their performance in translating informal Slavic conversations. The research explored diverse prompt-engineering strategies to enhance translation accuracy. A novel evaluation method was developed to assess translation quality. The main goal was to create practical guidelines for selecting and using the most effective translation tools for everyday Slavic-language conversations.

Materials and Methods

This research investigated translation models for Slavic languages, employing a comprehensive methodology encompassing multiple evaluation approaches and prompting-engineering methods. The study was based on the OpenSubtitles2018 dataset, which represents a substantial collection of movie and TV show subtitles. Seven Slavic languages were selected for investigation: Ukrainian, Czech, Bulgarian, Russian, Albanian, Macedonian, and Slovak. This dataset was chosen due to its representation of natural dialogue and concise sentence structures, typically ranging from 7 to 12 words, accurately reflecting the nature of informal chat-based communication. The OpenSubtitles2018 dataset included 1.2 million sentence pairs for each Slavic language pair. After filtering for conversation-style content, segments with dialogue markers were retained, maintaining an average sentence length of 8.3 words. The final corpus contained natural conversations spanning multiple domains, including daily communication (43%), informal discussions (37%), and casual narratives (20%). The research framework incorporated five distinct translation systems: ChatGPT4, Claude 3.5, LLaMA-3, Google Translate, and Helsinki-NLP's Opus-MT. These systems were selected to represent both state-of-the-art language models and translation services.

The evaluation methodology employed a multi-metric approach comprising three primary dimensions. As demonstrated by F. Kepler *et al*. (2021) in their comparative analysis of translation quality estimation approaches, utilising multiple evaluation metrics provides a more comprehensive assessment of translation quality. First, semantic quality assessment employed COMET and Text Correlation metrics to measure meaning preservation and semantic similarity between source and translated texts. Second, stylistic and lexical accuracy were assessed through TER (Translation Edit Rate) and CHRF (Character n-gram F-score) metrics, providing quantitative measures of translation precision. Third, a novel LLM-based evaluation method was implemented, engaging ChatGPT, Llama, and Claude to perform qualitative assessments of translation accuracy.

Furthermore, the research explored four distinct prompting methodologies: Basic Prompt, ZeroShot Chain-of-Thought Prompting (CoTT), Contrastive Translation (CT), and Structured Example-Based Prompting (FSL). This selection of prompting methods built upon the framework created by L. Reynolds & K. McDonell (2021), who demonstrated that moving beyond basic fewshot prompting can significantly improve model performance in complex language tasks. Their research particularly emphasised the importance of structured approaches in handling nuanced linguistic challenges. Each method tested how well the models performed in different situations: CoTT enhanced logical clarity, CT improved meaning preservation by comparing options, and FSL increased accuracy using examples. These approaches were systematically tested using Ukrainian language data from the OpenSubtitles2018 dataset, selected for its representation of diverse conversational styles.

The experiments utilised popular Python libraries such as Transformers, COMET, and SacreBLEU, ensuring robust and reproducible results. Additionally, all datasets, scripts, and evaluation metrics were made publicly available through an open-access repository, promoting transparency and enabling further research in this domain. This comprehensive methodological approach facilitated a thorough investigation of translation quality across multiple dimensions, providing insights into both the semantic accuracy and stylistic appropriateness of machine-generated translations in informal communication contexts.

Results and Discussion

Analysis of translation model performance

GPT-4 (4-o1) is OpenAI's latest generation of models. As T.B. Brown *et al*. (2020) demonstrated, this model is trained with a mixture of supervised fine-tuning and

reinforcement learning, designed specifically for conversational and language understanding tasks. GPT-4's architecture is based on the Transformer (Devlin *et al.*, 2019), featuring self-attention mechanisms and deep layers that process language sequentially. It can handle context across long passages, making it suitable for complex translation tasks where meaning and context extend across multiple sentences. The core component of the Transformer architecture is selfattention, calculated as follows:

$$Attention(Q,K,V) = softmax\left((QK^T)/\sqrt{(d_k)}\right)V, \quad (1)$$

where *Q, K,* and *V* are the query, key, and value matrices derived from input embeddings, and *k* is the dimensionality of the keys. GPT-4 is pre-trained on a vast dataset, enabling nuanced translations. However, it may lack specific cultural or idiomatic accuracy without fine-tuning tailored to Slavic languages.

Claude 3.5 Sonnet, developed by Anthropic, focuses on ethical language processing with a transformer-based architecture. Claude's attention mechanism and transformer layers follow a similar approach to other models, where self-attention mechanisms and positional encodings establish word order. Claude performs well on literal translations but may struggle with idiomatic expressions due to its safety-focused training.

Meta's LLaMA (Large Language Model Meta AI) (Touvron *et al.*, 2023) is an open-source large language model that has demonstrated capabilities in multilingual tasks. LLaMA-3 (70B) boasts extensive vocabulary coverage across multiple languages, including Slavic languages, making it particularly well-suited for translation tasks requiring nuanced and contextually accurate outputs. Similar to other transformer models, its large parameter count enables LLaMA-3 to handle subtle linguistic features in Slavic languages effectively.

Opus-MT (Tiedemann & Thottingal, 2020), developed by Helsinki-NLP, is a machine translation model built on the MarianMT framework. This model focuses on multilingual and lowresource language translation and is trained on parallel corpora, including the OPUS dataset. Opus-MT employs a transformer-based architecture optimised for the efficient translation of lowresource languages (Rei *et al.*, 2020). Opus-MT uses separate models for each language, which enhances its effectiveness for specialised Slavic translation tasks.

Google Translate uses the Google Neural Machine Translation (GNMT) system, updated with transformer-based advancements inspired by models such as BERT and T5. GNMT originally relied on an RNN-based architecture with attention mechanisms but has since incorporated transformer layers, boosting its performance in handling nuanced and lengthy texts. Known for speed and accessibility, Google Translate provides immediate translations but may lack the deep contextual understanding offered by larger LLMs.

Modern translation systems vary in their approach to Slavic language translation, with each offering distinct advantages. Large Language Models (GPT-4, Claude 3.5, LLaMA-3) excel in understanding context, while other systems (Opus-MT, Google Translate) prioritise accessibility. The diversity in translation approaches reflects the complex challenges identified by S. Ranathunga *et al*. (2021) in their analysis of low-resource language translation systems, where they emphasised that different architectural solutions may be necessary to address various aspects of Slavic language processing. GPT-4 uses a transformer architecture with self-attention mechanisms, making it effective for complex translations. Claude 3.5 emphasises ethical processing while maintaining high accuracy. LLaMA-3 (70B) provides robust multilingual support with extensive vocabulary coverage. Opus-MT specialises in low-resource languages with dedicated language-pair models. Google Translate offers quick, accessible translations using updated neural machine translation technology.

Evaluation methods for translation quality assessment

This study employed five main metrics to evaluate translation quality: Text Correlation, COMET, TER, CHRF, and an LLM-based evaluation approach. Each metric assesses different aspects of translation accuracy. Text Correlation (Pearson's Correlation) (Reimers & Gurevych, 2019) quantifies the degree of alignment between the predicted and reference translations by comparing text embeddings. The Pearson correlation coefficient (*r*) is calculated using the formula:

$$r = \left(\sum(x-\bar{x})(y-\bar{y})\right)/\left(\sqrt{[\sum(x-\bar{x})^2][\sum(y-\bar{y})^2]}\right). \quad (2)$$

This formula evaluates the linear relationship between two variables, specifically the semantic representations of the source and translated texts. The coefficient ranges from -1 to 1, where -1 indicates a perfect negative correlation, 0 indicates no correlation, and 1 indicates a perfect positive correlation. The variables in the formula are defined as follows: r – the Pearson correlation coefficient, representing the strength and direction of the linear relationship between the source and translated texts; x – the values of variable X, corresponding to the semantic representation of the source text; y – the values of variable Y, corresponding to the semantic representation of the translated text; $\bar{x}$ – the mean or average value of the semantic representation of the source text (variable X); $\bar{y}$ – the mean or average value of the semantic representation of the translated text (variable Y); n – the number of pairs of values, representing the data points or observations used in the calculation. The evaluation results for Semantic Quality (COMET, Text Correlation), as described by R. Rei *et al.* (2020), are presented in Table 1.

Table 1. Evaluation results for Semantic Quality (COMET, Text Correlation)

Evaluation method	Text correlation					COMET				
Language	ChatGPT	Claude	LlaMA	Opus-MT	Google Translate	ChatGPT	Claude	LlaMA	Opus-MT	Google Translate
Ukrainian	0.95	0.952	0.929	0.935	0.927	0.769	0.774	0.736	0.726	0.763

Table 1. Continued

Evaluation method	Text correlation					COMET				
Language	ChatGPT	Claude	LlaMA	Opus-MT	Google Translate	ChatGPT	Claude	LlaMA	Opus-MT	Google Translate
Albanian	0.89	0.876	0.842	0.821	0.858	0.789	0.788	0.758	0.765	0.782
Bulgarian	0.969	0.966	0.943	0.949	0.959	0.835	0.828	0.804	0.79	0.824
Czech	0.898	0.875	0.835	0.814	0.872	0.894	0.89	0.858	0.846	0.885
Macedonian	0.968	0.969	0.965	0.911	0.962	0.762	0.769	0.736	0.75	0.766
Russian	0.965	0.969	0.955	0.963	0.936	0.845	0.836	0.803	0.818	0.82
Slovak	0.825	0.821	0.798	0.776	0.821	0.762	0.764	0.721	0.709	0.761

Source: developed by the author based on O.O. Sokol (2024)

An analysis of Table 1 indicates that ChatGPT-4 performs well across languages, achieving high Text Correlation scores in Ukrainian (0.95) and Czech (0.898) and strong COMET scores, making it a top choice for context-rich translations. Claude 3.5 achieves high semantic accuracy, particularly in Macedonian (0.969) and Russian (0.969) for Text Correlation. LlaMA-3 demonstrates good performance, though its scores are lower than those of ChatGPT and Claude, revealing weaker handling of complex contexts. Google Translate delivers consistent results but exhibits lower semantic accuracy and struggles with nuanced translations. Helsinki-NLP/Opus-MT perform well, although its results vary across language pairs, particularly with complex languages. In summary, ChatGPT-4 and Claude 3.5 lead in semantic quality, with ChatGPT excelling in nuanced languages such as Czech and Ukrainian.

In the course of the study, an assessment was conducted on the performance of large language models in translating Slavic languages, particularly their ability to preserve stylistic and lexical accuracy in conversational style. To evaluate translation quality, two key metrics were employed: Translation Error Rate (TER) and Character n-gram F-score (CHRF). These metrics assess different aspects of translation accuracy and fluency.

TER (Translation Error Rate) measures the minimum number of edits needed to transform the machine translation into the reference translation:

$$TER = E/N_r, \quad (3)$$

where E is the total number of edits (insertions, deletions, substitutions, and shifts), N is the total number of words in the reference translation, and r is the subscript denoting reference text. This metric is particularly useful for evaluating translation accuracy as it directly quantifies how much editing is needed to achieve the correct translation. Lower TER scores indicate better translations, with fewer required edits.

CHRF evaluates translations at the character level, providing a more nuanced assessment that is especially important for Slavic languages with complex morphology:

$$chrF = F_\beta = \left((1+\beta^2) \times P \times R\right)/(\beta^2 \times P + R), \quad (4)$$

where $chrF$ is the character n-gram F-score, F_β is the F-score with β parameter, β is the Beta value (default = 3), P is the precision, and R is the recall. CHRF is effective for Slavic languages as it captures character-level similarities, making it sensitive to morphological variations and word endings that are crucial in these languages. Table 2 presents a summary of the test results for various translation models, illustrating their strengths and weaknesses when working with Slavic languages, particularly the results of stylistic and lexical accuracy assessments (TER and CHRF).

Table 2. Evaluation results for Stylistic/Lexical Accuracy (TER, CHRF)

Evaluation method	CHRF					TER				
Language	ChatGPT	Claude	LlaMA	Opus-MT	Google Translate	ChatGPT	Claude	LlaMA	Opus-MT	Google Translate
Ukrainian	13.11	15.12	11.856	10.842	12.531	0.774	0.76	1.707	0.826	0.796
Albanian	22.346	20.846	18.52	19.423	20.386	0.691	0.71	0.772	0.692	0.731
Bulgarian	20.769	21.166	16.215	13.822	18.145	0.635	0.636	0.689	0.754	0.656
Czech	34.628	30.96	24.166	30.786	33.387	0.489	0.531	0.606	0.548	0.534
Macedonian	23.789	23.879	19.764	26.808	21.712	0.802	0.601	0.65	0.592	0.608
Russian	24.695	25.576	18.619	25.502	25.266	0.599	0.597	0.689	0.626	0.6
Slovak	18.325	19.026	12.214	14.884	16.976	0.698	0.708	0.783	0.756	0.735

Source: developed by the author based on O.O. Sokol (2024)

As shown in Table 2, different models demonstrate varying performance levels across Slavic languages. The results reveal distinct trends, with ChatGPT and Claude generally achieving lower TER scores (indicating fewer required edits) and higher CHRF scores (showing better character-level accuracy) compared to other models.

Specifically, ChatGPT-4 leads with notably high CHRF scores in Czech (34.63) and Russian (24.70), although it exhibits slightly higher TER in Ukrainian, suggesting minor editing needs. Claude 3.5 follows closely, particularly excelling in Bulgarian (21.17) and Macedonian (23.88), with moderate TER scores. Conversely, LLaMA-3 shows lower CHRF scores and higher TER rates in Ukrainian, indicating certain limitations in grammatical and lexical precision. Google Translate maintains consistent performance with balanced CHRF and TER scores, making it reliable for basic translations. Helsinki-NLP/Opus-MT shows mixed results, particularly struggling with complex phrases. Based on the comprehensive analysis of TER and CHRF metrics, ChatGPT-4 and Claude 3.5 demonstrate statistically superior performance in translation accuracy.

Analysing Table 3, which presents the evaluation results for the LLM-based evaluation method, it is evident that ChatGPT-4 is the top performer in Slovak (0.5) and Albanian (0.554), showing robust contextual accuracy and idiomatic understanding. Claude 3.5 is close behind ChatGPT, particularly effective in Bulgarian (0.471) and Macedonian (0.485). LlaMA-3 achieved lower scores, especially in Slovak and Albanian, indicating challenges with nuanced translations. Google Translate and OpusMT achieve comparatively lower scores in contextual accuracy.

Table 3. Evaluation results for the LLM-based method

Language	ChatGPT	Claude	LlaMA	Opus-MT	Google Translate
Ukrainian	0.227	0.613	0.053	0.053	0.053
Albanian	0.554	0.4	0.015	0.015	0.015
Bulgarian	0.353	0.471	0.059	0.059	0.059
Czech	0.443	0.329	0.076	0.076	0.076
Macedonian	0.47	0.485	0.015	0.015	0.015
Russian	0.273	0.416	0.104	0.104	0.104
Slovak	0.5	0.421	0.026	0.026	0.026

Source: developed by the author based on O.O. Sokol (2024)

Overall, ChatGPT-4 and Claude 3.5 lead in contextual accuracy, with ChatGPT-4 showing superior idiomatic handling. Based on these experimental results, ChatGPT-4 consistently delivers the highest quality translations across semantic, stylistic, and LLM-based evaluations, making it the best overall model for translating Slavic languages with accuracy and contextual depth. Claude 3.5 also performs well, particularly in languages like Macedonian and Russian, making it a strong alternative. Google Translate provides fast, reliable translations with good lexical accuracy, making it suitable for general-purpose tasks but less capable of handling complex nuances. Helsinki-NLP's Opus-MT is useful for low-resource languages, though it shows limitations in stylistic fidelity and nuanced understanding.

Advanced prompting strategies for improving machine translation

Structured Example-Based Prompting (Few-Shot Learning with Examples) (Liu *et al.*, 2023) employs a limited set of examples to help LLMs generalise from specific prompts, allowing the model to mirror the style, tone, and idiomatic language of the provided translations. In this approach, each example demonstrates how informal tone, slang, and culturally specific phrases should be translated, and brief explanations clarify why certain expressions were chosen. This technique assists the model in developing a contextual understanding of Slavic linguistic subtleties by using examples in casual conversational formats. The example-based prompt is structured as follows:

Translate the following chat-based text from English to Ukrainian, keeping the informal tone, conversational style, and cultural nuances intact. Use the examples below for guidance on handling slang, tone, and natural flow.

Example 1: "Hey man, what's up? Everything going as planned?" → "Привіт, друже, як справи? Все йде за планом?" (Explanation: "man" is casually translated to "друже", maintaining a friendly tone).

Example 2: "Come on, this is just causing trouble!" → "Та ну, це тільки створює проблеми!"

(Explanation: "Come on" as an expression of frustration is translated as "Та ну" for conversational effect).

Example 3: "Oh, here we go again with his stories!" → "О, знову він зі своїми історіями!" (Explanation: "Oh, here we go again" is rendered to show mild annoyance in a relatable way).

Zero-Shot Chain-of-Thought Prompting (Chain-of-Thought Translation) draws inspiration from Chain-of-Thought reasoning techniques (Wei *et al.*, 2022). It helps the model apply nuanced linguistic and cultural choices independently by outlining steps for reasoning through the translation. The Chain-of-Thought Translation (CoTT) prompt employs a three-step approach, encouraging the model to think through the translation with a focus on meaning, tone, and conversational style:

You are a professional translator. The text to translate is from chat-based dialogues and includes informal speech, slang, regional expressions, and colloquial nuances typical of everyday conversation.

Step 1: Identify the core meaning of each phrase and recognise idiomatic expressions in the source language.

Step 2: Reflect on the informal tone and slang used. Choose equivalent expressions in Ukrainian that convey the same style, emotion, and intent.

Step 3: Craft a fluent, coherent translation that feels natural. Avoid literal translations if they interfere with conversational flow.

To evaluate different prompting methods, their performance across multiple translation systems was tested using LLM-based scoring metrics. The comparative results of Basic, Chain-of-Thought Translation (CoTT), Contrastive Translation (CT), and Few-Shot Learning with Examples (FSL) methods are presented in Table 4.

Table 4. Evaluation results of prompt engineering using LLM-based model scores

Evaluation method	LLM-based score				
PE method	ChatGPT	Claude	LlaMA	Opus-MT	Google Translate
Basic	0.227	0.613	0.053	0.053	0.053
CoTT	0.125	0.708	0.056	0.056	0.056
CT	0.181	0.736	0.028	0.028	0.028
FSL	0.262	0.639	0.033	0.033	0.033

Source: developed by the author based on O.O. Sokol (2024)

This method is highly effective for informal translation, as it selects the most appropriate expression at each step. It enhances translation quality in a structured manner by encouraging LLMs to evaluate idioms, slang, and cultural tone sequentially. According to T. Kojima *et al*. (2022), reasoning-based prompts like Chain-of-Thought are particularly valuable for handling complex language tasks, as they improve model consistency and adaptability.

Contrastive translation employs a two-step approach where the model first produces a literal translation and then refines it to enhance fluency and colloquialism. The prompt instructs the model to first provide a direct translation, capturing the literal meaning, and then adjust the wording to make it sound natural in Ukrainian: *"Translate the following chat-based text from English to Ukrainian using a two-step approach. First, provide a direct, literal translation to capture the basic meaning. Then, refine this translation to make it sound natural and conversational in Ukrainian. Ensure that you preserve the tone, idioms, and any cultural references to make the translation feel authentic and relatable to a native speaker".*

The evaluation results in Tables 4 and 5 compare the performance of various prompt engineering methods for improving the translation quality of large language models (LLMs) when dealing with conversational Ukrainian texts. The analysis includes quantitative metrics (TER, CHRF, COMET) and qualitative assessments, enabling the evaluation of translation accuracy, stylistic alignment, and contextual appropriateness. This study highlights the strengths and weaknesses of different approaches, particularly FSL and CoTT, which demonstrate superior adaptation to the conversational style of Ukrainian compared to basic prompts.

Table 5. Evaluation results of prompt engineering using semantic and statistic scores

Evaluation method	Text correlation			CHRF			TER			COMET		
PE method	ChatGPT	Claude	LlaMA	ChatGPT	Claude	LlaMA	ChatGPT	Claude	LlaMA	ChatGPT	Claude	LlaMA
Basic	0.95	0.952	0.929	13.11	15.12	11.856	0.774	0.76	1.707	0.769	0.774	0.736
CoTT	0.941	0.951	0.922	12.17	14.747	11.4	0.835	0.788	0.837	0.755	0.772	0.744
CT	0.946	0.92	0.858	13.03	7.442	7.708	0.8	0.862	0.875	0.767	0.708	0.668
FSL	0.926	0.948	0.943	12.34	13.357	11.596	0.796	0.795	0.848	0.77	0.768	0.736

Source: developed by the author based on O.O. Sokol (2024)

The findings demonstrated that advanced Prompt Engineering methods, particularly Structured Example-Based Prompting (FSL) and Zero-Shot Chain-of-Thought (CoTT), achieved better results across multiple evaluation metrics, including LLM-Based Scores, Text Correlation, Character n-gram F-score (CHRF), Translation Error Rate (TER), and COMET. Analysis of Tables 4 and 5 suggests the following:

- LLM-Based Scores: FSL and CoTT received higher qualitative scores from ChatGPT and Claude for conversational accuracy, with FSL ranking highest in ChatGPT's assessments. Basic and CT prompts scored lower; Basic prompts struggle with stylistic nuances in Ukrainian.
- Text Correlation: basic prompting achieved the highest text correlation scores, likely due to its simpler approach focusing on direct translation. This indicates that while basic prompting produces accurate literal translations, it may lack conversational fluidity.
- CHRF: FSL outperformed others in balancing lexical similarity and natural tone, suggesting that example-based prompting helped models better align translations with the conversational nature of the source text.
- TER: basic prompting had slightly lower TER, suggesting fewer necessary edits, but FSL and CoTT provided closer matches to nuanced Ukrainian conversational standards.
- COMET: FSL scored consistently higher across COMET evaluations, underscoring its ability to generate contextually appropriate, fluent translations.

The comparative analysis demonstrates that advanced prompt engineering methods, especially Structured Example-Based Prompting (FSL) and Zero-Shot Chain-of-Thought (CoTT), significantly enhance the conversational accuracy of chat-based translations into Ukrainian. Both FSL and CoTT exhibit stronger performance across

stylistic and contextual measures (CHRF, LLM-based scores, and COMET) than the Basic Prompt, confirming the effectiveness of structured guidance and thoughtbased multi-step prompting. For translation tasks focused on informal, chat-like content, advanced prompt methods (FSL or CoTT) are recommended. They provide contextually rich translations compared to the Basic Prompt. However, if efficiency is the primary concern, the Basic Prompt remains a good choice. The findings suggest that advanced prompt engineering can enhance translation quality in informal contexts, particularly in languages with rich idiomatic expressions and cultural nuances.

The results of this study significantly contribute to the understanding of large language models' capabilities in Slavic language translation, particularly in informal conversational contexts. Through comprehensive evaluation across multiple metrics and prompt engineering methods, the research demonstrates that advanced LLMs can effectively handle the complex morphological and stylistic features of Slavic languages while maintaining conversational authenticity. This finding expands upon previous research in several key areas.

W. Jiao *et al*. (2023) conducted preliminary studies of ChatGPT's translation capabilities, focusing primarily on mainstream languages like English and Chinese. Their research showed promising results with accuracy rates of 85-90% for these languages. This study extended these findings by demonstrating even higher accuracy rates for Slavic languages, with ChatGPT achieving 95% accuracy for Ukrainian and 97% for Bulgarian texts. These results indicated that LLMs may be particularly effective for Slavic languages.

The NLLB Team *et al*. (2022) examined scaling human-centred machine translation across multiple languages but did not specifically address informal conversation translation, which constitutes approximately 70% of daily online communication. While they reported significant improvements in formal translation tasks, the results indicate that LLMs like ChatGPT and Claude can maintain similar levels of accuracy even in casual conversational contexts, addressing a gap in their findings. This finding underscores the importance of developing translation systems capable of handling both formal and informal language.

G. Nicholas & A. Bhatia (2023) highlighted several limitations in LLMs' handling of nonEnglish content, particularly regarding cultural context and colloquial expressions. This research partially challenges their conclusions by demonstrating that, with appropriate prompt engineering – particularly using the Structured Example-Based method – these limitations can be significantly mitigated for Slavic languages. The study found that using the Structured Example-Based method enhanced the accuracy of translating colloquial expressions by 30% compared to conventional translation methods.

A. Koubaa *et al*. (2023) assessed ChatGPT's general translation capabilities, reporting variable performance across different language pairs. This finding aligns with their results regarding inconsistency across languages but demonstrates higher overall accuracy rates specifically for Slavic languages, suggesting that LLMs might have particular strengths in these languages. Their comprehensive study analysed translations across 14 language pairs, employing both automatic metrics (BLEU, CHRF) and human evaluation protocols. The researchers particularly noted that ChatGPT achieved a remarkable 87% accuracy rate for Slavic languages, while performance for Asian languages averaged around 72%. They also observed that the model's performance significantly improved when handling shorter sentences and technical content.

M. Freitag *et al*. (2021) developed a new framework for evaluating machine translation of formal content. They combined traditional metrics with advanced LLM-based methods to better suit conversational text. Their three-part system used automated metrics, human assessment, and context analysis. The researchers worked with 50 professional translators to assess 2,000 translated segments. They found that both adequacy and fluency are crucial in translation evaluation and suggested standard rubrics for training evaluators.

M. Popovic & A. Poncelas (2020) achieved notable success in formal news translation for Slavic languages, reporting near-professional quality. This study demonstrates that similar levels of quality can now be achieved in informal translations using LLMs, marking a significant advance in the field. Their research involved a detailed analysis of translations between Russian, Polish, and Czech languages, utilising a corpus of over 100,000 news articles. The authors implemented a hybrid approach combining statistical machine translation with neural networks, achieving BLEU scores above 0.45 for all language pairs. They particularly noted improvements in handling complex grammatical structures and maintaining stylistic consistency across different text genres.

The comparative analysis by X. Qiu (2023) of cultural nuances in translation highlighted the importance of context-aware systems. The results align with their findings while demonstrating that modern LLMs can effectively handle these nuances, particularly when using prompt engineering techniques. X. Qiu's study examined translations of culturally specific content across Chinese, English, and Japanese, focusing on idiomatic expressions and cultural references. The research utilised a dataset of 5,000 culturally rich texts and developed a novel evaluation metric for measuring cultural preservation accuracy. Their findings indicated that properly engineered prompts could improve cultural nuance preservation by up to 35% compared to baseline translations, with particularly strong results in preserving metaphorical expressions and cultural context.

This study raises new questions for future researchers. Further investigation is needed into the handling of extended conversations, the impact of cultural context on translation quality, and the optimisation of prompt engineering techniques specifically for Slavic languages. This study constitutes the first comprehensive evaluation of LLMs'

capabilities in translating casual conversations in Slavic languages, introducing new evaluation methodologies and demonstrating better results compared with traditional translation services for conversational translation tasks.

Conclusions

This research provided a comprehensive analysis of large language models' performance in Slavic language translation for casual conversations and identifies the most effective prompt engineering methods for these translations. These goals were successfully achieved through systematic testing and analysis, and the study also developed open-source code for evaluating translation quality and selecting optimal translation methods.

The research methodology included the evaluation of five translation systems (ChatGPT-4, Claude 3.5, LLaMA-3, Google Translate, and Helsinki-NLP's Opus-MT) across seven Slavic languages using the OpenSubtitles2018 dataset. Through the implementation of multiple evaluation metrics (COMET, Text Correlation, TER, CHRF, and LLM-based), the study revealed that modern LLMs, particularly ChatGPT-4 and Claude 3.5, perform better in handling conversational translations compared to traditional systems. Quantitative analysis indicated ChatGPT-4 achieved 95% accuracy for Ukrainian and 97% for Bulgarian translations while maintaining conversational authenticity. The experimental evaluation of prompt engineering methods established that the Structured Example-Based Prompting (FSL) method produced optimal translation quality for informal content. This approach demonstrated improvements in preserving idiomatic expressions and cultural context, showing a 30% increase in accuracy compared to conventional translation methods. Additionally, the Chain-of-Thought prompting method improved cultural context preservation by 25%.

This research advances machine translation by providing validated metrics for Slavic informal translation, achieving a 25-30% improvement in accuracy over baseline systems, introducing an evaluation framework, and demonstrating LLMs' performance in preserving conversational elements. The findings indicate consistent improvements in contextual accuracy and colloquial expression translation. This study successfully validated LLMs' capabilities in Slavic language translation, achieving notable results in both accuracy and conversational authenticity. The research contributed practical evaluation methods and prompting techniques, creating a robust foundation for advancing informal translation technologies.

Several limitations of this study should be acknowledged. The main constraints included the limited availability of conversational datasets for Slavic languages and the testing coverage of only seven Slavic languages. Additionally, computational resource constraints affected the scale of possible experiments. Future research should focus on three main directions: investigating translations in longer conversations, studying regional language differences, and developing better prompt engineering methods for Slavic languages. Additional work is needed to understand how cultural context affects translation quality and to develop improved methods for evaluating conversational translations.

Acknowledgements

None.

Conflict of Interest

None.

References

[1] Bhatt, S., & Diaz, F. (2024). Extrinsic evaluation of cultural competence in large language models. *ArXiv.* doi: 10.48550/arXiv.2406.11565.

[2] Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., & Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877-1901. doi: 10.48550/arXiv.2005.14165.

[3] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*. doi: 10.48550/arXiv.1810.04805.

[4] Escolano, C., Costa-jussà, M.R., & Fonollosa, J.A.R. (2020). The TALP-UPC system description for WMT20 news translation task: Multilingual adaptation for low resource MT. In *Proceedings of the fifth conference on machine translation* (pp. 134-138). Kerrville: ACL.

[5] Freitag, M., Foster, G., Grangier, D., Ratnakar, V., Tan, Q., & Macherey, W. (2021). Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9(1), 1460-1474. doi: 10.1162/tacl_a_00437.

[6] Jiao, W., Wu, H., Wang, W., Wan, Y. & Lyu, M. (2023). ChatGPT or Grammarly? Evaluating ChatGPT on grammatical error correction benchmark. *ArXiv.* doi: 10.48550/arXiv.2303.13648.

[7] Kepler, F., Trénous, J., Treviso, M., Vera, M., & Góis, A. (2021). Comparative analysis of current approaches to quality estimation for neural machine translation. *Applied Sciences*, 11(14), article number 6584. doi: 10.3390/app11146584.

[8] Kojima, T., Gu, S.S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. *ArXiv.* doi: 10.48550/arXiv.2205.11916.

[9] Koubaa, A., Boulila, W., Ghouti, L., & Alzahem, A. (2023). Exploring ChatGPT capabilities and limitations: A survey. *IEEE Access*, 11, 95574-95593. doi: 10.1109/ACCESS.2023.3326474.

[10] Liu, J., Shen, D., Zhang, Y., & Dolan, B. (2022). Few-shot learning through structured example-based prompting. In *Proceedings of the 60th annual meeting of the association for computational linguistics (ACL 2022)* (pp. 7688-7699). doi: 10.18653/v1/2022.acl-long.529.

[11] Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9), article number 195. doi: 10.1145/3560815.

[12] Naveen, P., & Trojovský, P. (2024). Overview and challenges of machine translation for contextually appropriate translations. *iScience*, 27(1), article number 110878. doi: 10.1016/j.isci.2024.110878.

[13] Nicholas, G., & Bhatia, A. (2023). Lost in translation: Large language models in non-english content analysis. *Journal of Artificial Intelligence and Society*, 15(4), 423-450. doi: 10.48550/arXiv.2306.07377.

[14] NLLB Team *et al*. (2022). No language left behind: Scaling human-centered machine translation. *ArXiv*. doi: 10.48550/arXiv.2207.04672.

[15] Popovic, M., & Poncelas, A. (2020). Neural machine translation between similar South-Slavic languages. In *Proceedings of the 5th conference on machine translation (WMT)* (pp. 430-436). Kerrville: ACL.

[16] Qiu, X. (2023). Cultural differences and translation strategies. *Journal of Education and Educational Research*, 2(3), 100-105. doi: 10.54097/jeer.v2i3.7741.

[17] Ranathunga, S., Lee, E.A., Skenduli, M.P., Shekhar, R., Alam, M., & Kaur, R. (2021). Neural machine translation for low-resource languages: A survey. *ArXiv.* doi: 10.48550/arXiv.2106.15115.

[18] Rei, R., Stewart, C., Farinha, A.C., & Lavie, A. (2020). COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 conference on empirical methods in natural language* (pp. 2685-2702). Kerrville: ACL. doi: 10.18653/v1/2020.emnlp-main.213.

[19] Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)* (pp. 3982-3992). Hong Kong: ACL. doi: 10.18653/v1/d19-1410.

[20] Reynolds, L., & McDonell, K. (2021). Prompt programming for large language models: Beyond the few-shot paradigm. In *CHI EA '21: Extended abstracts of the 2021 CHI conference on human factors in computing systems* (article number 314). Yokohama: ACM. doi: 10.1145/3411763.3451760.

[21] Sokol, O.O. (2024). *Chat-based translation system with LLMs*. Retrieved from https://github.com/sokolheavy/slavic-llm-translator.

[22] Tang, X., & Zheng, Y. (2023). Unpacking complex language ideologies toward heritage language maintenance: A case of Chinese migrant families in the US. *International Multilingual Research Journal,* 17(4), 333-350. doi: 10.1080/19313152.2023.2209358.

[23] Tang, Y., Tran, C., Li, X., Chen, P. J., Goyal, N., Chaudhary, V., Gu, J., & Fan, A. (2021). Multilingual translation from denoising pre-training. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021* (pp. 3450-3466). Kerrville: ACL. doi: 10.18653/v1/2021.findings-acl.304.

[24] Tiedemann, J., & Thottingal, S. (2020). OPUS-MT – building open translation services for the World. In *Proceedings of the 22nd annual conference of the european association for machine translation* (pp. 479-480). Lisboa: European Association for Machine Translation.

[25] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., & Lample, G. (2023). LLaMA: Open and efficient foundation language models. *ArXiv*. doi: 10.48550/arXiv.2302.13971.

[26] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2022). Chain of thought prompting elicits reasoning in large language models. *ArXiv*. doi: 10.48550/arXiv.2201.11903.

[27] Wieting, J., Berg-Kirkpatrick, T., Gimpel, K., & Neubig, G. (2019). Beyond BLEU: Training neural machine translation with semantic similarity. *ArXiv.* doi: 10.48550/arXiv.1909.06694.

[28] Zhu, W., Liu, H., Dong, Q., Xu, J., Huang, S., Kong, L., Chen, J., & Li, L. (2023). Multilingual machine translation with large language models: Empirical results and analysis. *ArXiv.* doi: 10.48550/arXiv.2304.04675.

Переклад слов'янських мов у розмовному стилі за допомогою великих мовних моделей

Олена Сокол

Аспірант
Київський національний університет імені Тараса Шевченка
01033, вул. Володимирська, 60, м. Київ, Україна
https://orcid.org/0000-0003-0465-6843

Анотація. Сучасні великі мовні моделі (LLM) демонструють значний потенціал у галузі машинного перекладу, особливо для слов'янських мов, які часто недостатньо представлені у традиційних наборах даних для перекладу. Метою дослідження була оцінка ефективності використання LLM (ChatGPT, Claude та Llama) для перекладу розмовних текстів слов'янськими мовами порівняно з комерційними перекладачами та трансформер-моделями. У роботі використано датасет OpenSubtitles2018 для тестування перекладів сімома слов'янськими мовами, застосовуючи методи семантичної та стилістичної оцінки якості перекладу. Результати показують, що ChatGPT і Claude забезпечують кращу якість перекладу порівняно з Google Translate та трансформер-моделями, особливо для неформальних розмов, досягаючи 95 % точності для української та 97 % для болгарської мов. Структурований метод промптів з прикладами (FSLE) показав найкращі результати. Дослідження показало, що використання LLM значно покращує якість перекладу неформальних текстів слов'янськими мовами, зберігаючи контекст та природність діалогу. Аналіз також виявив, що LLM краще справляються з перекладом ідіом та сленгу, забезпечуючи на 30 % вищу точність порівняно з традиційними системами машинного перекладу. При використанні методу ланцюжків міркувань (Chain-of-Thought) спостерігалося покращення збереження культурного контексту на 25 %. Практична цінність дослідження полягає в розробці ефективних методів використання LLM для якісного перекладу неформальних текстів слов'янськими мовами, що особливо корисно для месенджерів, соціальних мереж та розважального контенту, де важливе збереження природності мовлення та культурного контексту

Ключові слова: LLM; інженерія промптів; обробка природної мови; TER; COMET; кореляційний аналіз тексту; CHRF

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

Information Technologies and Computer Engineering **Vol. 21, No. 3. 2024**

Article's History: Received:04.09.2024 Revised: 20.11.2024 Accepted: 26.12.2024

UDC 004.056 10.63341/vitce/3.2024.53

Comparative analysis of the results of pseudorandom number generators for digital noise generation

Oleksandr Isakov*

Postgraduate Student
Lviv Polytechnic National University
79013, 12 Stepan Bandera Str., Lviv, Ukraine
https://orcid.org/0009-0007-4632-9492

Stepan Voitusik

PhD, Associate Professor
Lviv Polytechnic National University
79013, 12 Stepan Bandera Str., Lviv, Ukraine
https://orcid.org/0000-0003-4234-3303

Abstract. The paper presents the results of a study of the characteristics of five different pseudorandom number generators for use in digital noise generation problems used to mask signals in cybersecurity. The relevance of the study was conditioned by the growing need for high-quality masking methods that provide both effective performance and reliability of randomness, which is important for protecting confidential information in modern digital systems. The purpose of the study was to compare the PCG, Xoshiro128++, WELL512a, Mersenne Twister, and KISS algorithms in terms of their performance, statistical randomness, and ability to effectively mask a useful signal with noise. The performance of the algorithms was evaluated using BenchmarkDotNet. Standard NIST, Dieharder, and TestU01 tests were used to check the quality of sequence randomness. For the generated noise, a spectral analysis was performed using the power spectral density value. The masking efficiency was calculated by the signal-to-noise ratio, the results of the autocorrelation function, and the noise spectrogram. The results of the study showed that PCG and KISS are the most productive in terms of speed, which makes them attractive for applications where fast random sequence generation is important. WELL512a and PCG demonstrated the highest randomness quality, consistently passing all statistical tests. Analysis of the spectral noise distribution showed that all generators provide a uniform power distribution before filtering, and after filtering, the noise is successfully limited in the high-frequency range. The signal-to-noise ratio for all algorithms was about -13.6 dB, which indicates similar efficiency in noise masking. Autocorrelation analysis confirmed a low correlation for all generators outside of zero lag, which is important for maintaining the quality of randomness in long sequences. The practical value of the study lies in the selection of the optimal pseudorandom number generator for noise reduction problems in cybersecurity. The results obtained provide recommendations for choosing algorithms based on their speed and randomness, which will ensure a high level of information protection in digital systems

Keywords: information security; noise characteristics; statistical randomness tests; spectral analysis; performance tests; signal noise

Introduction

Modern cybersecurity systems require continuous improvement of information security methods due to the growing threat of cyber-attacks, in particular, encryption, simulated protection, masking, and scrambling. One of the most effective approaches to preventing information leakage through speech channels is to use a pseudorandom number generator (PRNG) to generate digital noise, which ensures that the useful signal is jammed outside the controlled zone. However, the key issue remains the choice of the optimal algorithm for generating pseudorandom numbers, which will provide not only high speed, but also high noise quality in terms of statistical randomness, spectral characteristics, and autocorrelation level. Choosing the right PRNG for a specific task is important, since each

Suggested Citation:
Isakov, O., & Voitusik, S. (2024). Comparative analysis of the results of pseudorandom number generators for digital noise generation. *Information Technologies and Computer Engineering*, 21(3), 53-64. doi: 10.63341/vitce/3.2024.53

*Corresponding author

algorithm has its own characteristics that affect the performance and quality of digital noise.

Pseudorandom number generation is an important aspect of many cybersecurity systems, as it is these algorithms that ensure the reliability of cryptographic functions, digital noise generation, and other forms of information protection. There are a large number of studies devoted to the comparative analysis of various PRNG algorithms in terms of their performance (Syafalni *et al*., 2022) and the quality of randomness (Vennos *et al*., 2021), however, the choice of the optimal algorithm remains an open question. The need to create digital noise with appropriate characteristics that would meet modern cybersecurity requirements makes this research relevant and timely. The need to filter evenly distributed values leads to changes in PRNG results and requires additional noise verification in the context of digital signal processing.

Many papers focused on analysing the randomness of numbers generated by PRNG algorithms using well-known test packages such as NIST STS, Dieharder, and TestU01. For example, research by P. L'ecuyer (2017) has become one of the most influential, as it contains a comprehensive randomness analysis of various PRNGs using TestU01, which has become the basis for further comparative studies in this area. This approach helped to identify algorithms that provide high-quality number generation, and provided valuable recommendations for selecting PRNG for various applications.

However, it is not just the quality of randomness that is crucial for choosing a PRNG. For example, Z. Hu (2020) and K. Mandal (2022) examined cryptographic pseudorandom generators for specific applications such as encryption and simulated protection. The researchers emphasised the importance of PRNGs performance, and their ability to provide an appropriate level of security that meets the requirements of modern cryptographic protocols. These studies were an important step in understanding that it is important for cryptography to consider not only randomness, but also speed, especially in high-load environments.

O.V. Isakov & S.S. Voitusik (2023) performed a similar analysis, namely NIST statistical tests and frequency analysis of generated noise based on Additive Fibonacci Generators. Based on the results of the conducted studies, the importance of analysing the characteristics of noise after its filtering was determined. Additional statistical tests were performed in this study to obtain more accurate results. In addition, the analysis of the speed and overlap of noise with the real signal helped to draw conclusions about the effectiveness of noise, and not just about the results of frequency analysis of the generated noise.

F. Yu *et al*. (2021), M.S. Feali (2023) and S. Li *et al*. (2024) presented high-performance PRNGs implemented on FPGAs that successfully passed statistical tests. These studies demonstrated the effectiveness of hardware implementation for specific digital noise generation tasks where both performance and randomness are important. Such approaches confirm the feasibility of using specialised hardware solutions to improve the reliability and speed of PRNGs, which are used to create digital noise in conditions of large amounts of data.

Overall, previous research confirmed the importance of a comprehensive approach to selecting PRNGs for digital noise generation. While randomness testing is a prerequisite for determining the reliability of an algorithm, performance and noise spectral characteristics are also important factors to consider when creating efficient and secure systems. Unlike existing studies, which are usually limited to one aspect of PRNG estimation, this study is based on an integrated approach and considers not only speed, but also various aspects of noise quality, such as randomness, spectral characteristics, and resistance to filtration and de-noising processes.

The purpose of the study was to determine the optimal algorithm for generating pseudorandom numbers for creating digital noise, which is used in information security systems, by comparing the speed, quality of generation, and spectral characteristics of digital noise of various PRNGs. This approach will not only increase the level of security of cyber systems, but also ensure the stability of their operation in conditions of rapid growth in the volume of transmitted information. The conducted experiments provided conclusions about the optimal choice of PRNGs for generating digital noise, which makes this study relevant for improving information security systems in the context of modern cybersecurity challenges.

Materials and Methods

Five well-known algorithms were selected for the study: PCG, Xoshiro128 ++ , WELL512a, Mersenne Twister, and KISS. All algorithms were analysed using methods such as speed measurement, statistical randomness tests, spectral analysis, and SNR (Signal-to-Noise Ratio) estimation after noise was superimposed on the useful signal.

Pseudorandom number generators (PRNG) were chosen in such a way that they included algorithms known for their simplicity, application, statistical characteristics, period, and ease of hardware implementation. The selected characteristics were met by the following set of generators:

- PCG (Permuted Congruential Generator) – high-speed and efficient generator designed to provide high-quality randomness;
- Xoshiro128 ++ – popular generator with fast execution, often used in scientific and technical applications;
- WELL512a (Well Equidistributed Long-period Linear) – generator with well-distributed random numbers that has a long period (Panneton *et al*., 2006);
- Mersenne Twister – generator with a very long service life and high randomness quality, widely used in numerous industries;
- KISS (Keep It Simple Stupid) – simple and fast generator.

BenchmarkDotNet – used to measure the performance of each PRNG algorithm in milliseconds, which allowed comparing the performance of different algorithms. Each

algorithm generated 268,435,456 values (2**33 bits) of the "uint" type, and the average generation time was written to the results file. This helped to evaluate the effectiveness of each PRNG for scenarios that require high speed.

Standardised NIST, Dieharder, and TestU01 tests were used to assess the quality of randomness of PRNG results. They helped to determine whether the generated numbers meet the requirements of randomness by evaluating various statistical properties of the sequences. The initial input coefficients were the same in all algorithms, except for those determined by the algorithm itself, and the array of initial values predicted in Mersenne Twister. Testing included evaluating frequency, block frequency, longest run length, matrix ranks, and pattern analysis. All tests used a p-value threshold from 0.01 to 0.99, where deviation from this range was considered unacceptable.

To evaluate the effectiveness of noise, a conversation recording from the Common Voice Delta Segment 19.0 (Sound data sets, n.d.), which was superimposed on the noise generated by each of the algorithms. This allowed estimating the level of the signal-to-noise ratio and the spectral properties of the superimposed noise.

Sequences of pseudorandom numbers with a sampling rate of 32kHz were generated for each PRNG. Each result was then filtered to limit the spectral range of noise within the specified frequency limits. Filtering helped to reduce excess power in high-frequency components and make noise more suitable for masking tasks.

The Fast Fourier Transform (FFT) was used to analyse the spectral characteristics of the generated noise. The power spectral density (PSD) was plotted for uniformly distributed white and filtered noise at specified frequencies (0-4,000 Hz), which allowed estimating the power distribution of noise over frequencies.

After superimposing the noise on the useful signal, the SNR for each PRNG was calculated. The equation was used to find the SNR:

$$SNR = 10 * Log_{10}\left(\frac{P_{signal}}{P_{noise}}\right)[dB], \quad (1)$$

where P_{signal} and P_{noise} – signal and noise power. This parameter shows how effectively noise masks the useful signal. The SNR value was calculated for each algorithm and compared to determine the difference in masking efficiency.

A spectrogram of superimposed noise on conversation recording was constructed, which allowed estimating spectral changes over time. This helped to visually test the noise efficiency of each algorithm, since high-quality masking should ensure an even distribution of noise in the spectrogram.

Results

Comparison of the performance of selected PRNG

BenchmarkDotNet is a popular .NET library that allows conducting detailed performance tests of methods and ensures high accuracy and repeatability of measurements. This is achieved through several important mechanisms:

- **automatic code isolation**: BenchmarkDotNet automatically allocates tested methods to ensure that they do not affect other processes;
- **warmup**: the tool performs several "warmup" iterations before starting basic testing, which helps to avoid the impact of initialisation processes;
- **repeatability**: BenchmarkDotNet runs the required number of tests to reduce the impact of random runtime deviations.

BenchmarkDotNet is often used in scientific research to evaluate the performance of algorithms and methods. For example, in the study by O. Balalaieva *et al*. (2023), the authors used this tool to compare the performance of various serialisation methods, proving its effectiveness in accurately measuring runtime. The study by N. Filho (2024) examined in detail examples of the use and effective operation of this tool. This confirmed the reliability of BenchmarkDotNet as a research tool where accurate performance measurement is important. The test description and configuration took place in accordance with the official documentation (BenchmarkDotNet official documentation, n.d.).

Test package and environment metadata:

- BenchmarkDotNet v0.14.0, Windows 10 (10.0.19045.4529/22H2/2022Update)
- Intel Core i5-7400 CPU 3.00GHz (Kaby Lake), 1 CPU, 4 logical and 4 physical cores .NET SDK 6.0.400

According to the results obtained (Table 1), the fastest algorithm was PCG with an average generation time of 740.8 ms, which indicates its efficiency and speed compared to others. The KISS (Keep It Simple Stupid) algorithm also showed good performance, with an average time of 867.0 ms, which makes it the second fastest. Other algorithms, such as Xoshiro128 ++ , WELL512a (Well Equidistributed Long-period Linear), and Mersenne Twister, had significantly longer runtimes – 1,381.2 ms, 2,116.7 ms, and 1,951.3 ms, respectively.

Table 1. Performance results

Method	Mean	Error	StdDev	Ratio
PCG	740.8 ms	2.81 ms	2.49 ms	1.00
Xoshiro128++	1,381.2 ms	7.68 ms	7.19 ms	1.86
KISS	867.0 ms	3.55 ms	3.32 ms	1.17
WELL512a	2,116.7 ms	14.92 ms	13.95 ms	2.86
Mersenne Twister	1,951.3 ms	20.81 ms	19.47 ms	2.63

Source: output of the BenchmarkDotNet library

These metrics suggest that the PCG and KISS algorithms are the most efficient among those considered in this study in terms of generation rate, which may be an important factor in applications where high performance is a critical requirement. The obtained ratios of Error and StdDev values to Mean indicate that the results of the performance test are reliable. The Ratio value shows the ratio of the performance of each algorithm to the fastest tested (PCG).

Results of statistical tests NIST, Dieharder and TestU01

NIST. The results of the NIST tests are presented in Table 2. The input data was divided into 100 bit streams according to the recommendations given in the paper by L.E. Basshham *et al*. (2010). All algorithms have passed the required limit (78 + of the RandomExcursion type and 96 + for other tests), and can be considered high-quality.

Table 2. Results of NIST statistical tests

Test name	PCG	WELL512a	Mersenne Twister	KISS	Xoshiro128++
AproximateEntropy	99/100	100/100	100/100	98/100	97/100
BlockFrequency	100/100	99/100	99/100	99/100	97/100
CumulativeSums	98/100	99/100	100/100	99/100	99/100
FFT	96/100	99/100	98/100	99/100	97/100
Frequency	98/100	99/100	100/100	99/100	100/100
LinearComplexity	99/100	99/100	99/100	99/100	100/100
LongestRun	99/100	98/100	100/100	100/100	98/100
NonOverlapingTemplate	99/100	99/100	99/100	98/100	99/100
OverlapingTemplate	99/100	99/100	97/100	98/100	100/100
RandomExcursion	87/89	91/92	80/82	91/93	79/80
RandomExcursionVariant	87/89	91/92	80/82	92/93	78/80
Rank	98/100	98/100	100/100	97/100	98/100
Runs	99/100	99/100	98/100	100/100	99/100
Serial	99/100	99/100	98/100	99/100	99/100
Universal	100/100	100/100	99/100	99/100	99/100

Source: developed by the authors based on NIST test results

PCG showed a high level of randomness in many tests, in particular, in the frequency and block frequency test, with a high proportion of passed tests (more than 98%). The algorithm did a good job of testing the length and rank of the longest run, which indicates its ability to generate long sequences of random values without obvious patterns. However, 4 out of 100 FFT tests failed. In this case, 96 completed tests are a sufficient condition for passing the test, but the remaining PRNGs showed better results.

The WELL512a showed good randomness results, especially in frequency, block frequency, and rank tests. Most of the tests were passed with an indicator of 98-100%. The high efficiency of this algorithm in NonOverlappingTemplate tests was noted, which indicates its ability to generate sequences that are not adapted to patterns.

Mersenne Twister showed excellent results in many tests, in particular, the frequency test, which confirms its randomness. However, in the Runs test, there are fewer completed tests, which may indicate some frequency or less high randomness compared to other PRNGs.

The KISS algorithm passed most tests, including Block frequency, Runs test, and NonOverlappingTemplate, with high results (98-100% pass). However, the results of the Rank test are worse than those of other algorithms.

Xoshiro128 ++ showed slightly less stable results in some tests, especially in cumulative amounts, where its proportion of test completion was 97%, indicating possible periodic patterns in large samples.

Overall, the Mersenne Twister and WELL512a algorithms performed better in the context of randomness and compliance with NIST tests, while PCG and Xoshiro128 ++ had some deviations in specific tests. This analysis shows that the Mersenne Twister and WELL512a algorithms may be most suitable for generating digital noise in cybersecurity tasks due to their stability and compliance with a wide range of NIST tests.

Dieharder. Based on Dieharder tests for various pseudorandom number generators (PRNG), it is possible to evaluate their effectiveness and compliance with the criteria of randomness (Dieharder official documentation, n.d.). The results for the KISS, Mersenne Twister, PCG, WELL512a, and Xoshiro128 ++ algorithms are divided into three parts, which contain the sum of tests with the passed|week|-failed status (Table 3).

Table 3. Dieharder statistical test results

Test name	PCG	WELL512a	Mersenne Twister	KISS	Xoshiro128++
dab_bytedistrib	1|0|0	1|0|0	1|0|0	1|0|0	1|0|0
dab_dct	1|0|0	1|0|0	1|0|0	1|0|0	1|0|0
dab_filltree	2|0|0	2|0|0	2|0|0	2|0|0	2|0|0
dab_filltree2	2|0|0	2|0|0	2|0|0	2|0|0	2|0|0

Table 3. Continued

Test name	PCG	WELL512a	Mersenne Twister	KISS	Xoshiro128++
dab_monobit2	1|0|0	1|0|0	1|0|0	1|0|0	1|0|0
diehard_2dsphere	1|0|0	1|0|0	1|0|0	1|0|0	1|0|0
diehard_3dsphere	1|0|0	1|0|0	1|0|0	1|0|0	1|0|0
diehard_birthdays	1|0|0	1|0|0	0|1|0	1|0|0	1|0|0
diehard_bitstream	1|0|0	1|0|0	1|0|0	1|0|0	1|0|0
diehard_count_1s_byt	1|0|0	1|0|0	1|0|0	1|0|0	1|0|0
diehard_count_1s_str	1|0|0	1|0|0	1|0|0	0|1|0	1|0|0
diehard_craps	2|0|0	2|0|0	2|0|0	2|0|0	2|0|0
diehard_dna	1|0|0	1|0|0	1|0|0	1|0|0	1|0|0
diehard_operm5	1|0|0	1|0|0	1|0|0	1|0|0	1|0|0
diehard_opso	1|0|0	1|0|0	1|0|0	1|0|0	1|0|0
diehard_oqso	1|0|0	1|0|0	1|0|0	1|0|0	1|0|0
diehard_parking_lot	1|0|0	1|0|0	1|0|0	1|0|0	1|0|0
diehard_rank_32x32	1|0|0	1|0|0	1|0|0	1|0|0	1|0|0
diehard_rank_6x8	1|0|0	1|0|0	1|0|0	1|0|0	1|0|0
diehard_runs	2|0|0	1|1|0	2|0|0	2|0|0	2|0|0
diehard_squeeze	1|0|0	1|0|0	1|0|0	1|0|0	1|0|0
diehard_sums	1|0|0	1|0|0	1|0|0	0|1|0	1|0|0
marsaglia_tsang_gcd	2|0|0	2|0|0	0|2|0	1|0|1	2|0|0
Preparin	0|0|0	0|0|0	0|0|0	0|0|0	0|0|0
rgb_bitdist	12|0|0	12|0|0	12|0|0	11|1|0	11|1|0
rgb_kstest_test	1|0|0	1|0|0	1|0|0	1|0|0	1|0|0
rgb_lagged_sum	30|2|1	31|2|0	32|1|0	27|6|0	30|3|0
rgb_minimum_distance	4|0|0	4|0|0	4|0|0	4|0|0	4|0|0
rgb_permutations	4|0|0	4|0|0	4|0|0	4|0|0	3|1|0
sts_monobit	1|0|0	1|0|0	1|0|0	1|0|0	1|0|0
sts_runs	1|0|0	1|0|0	1|0|0	1|0|0	1|0|0
sts_serial	30|0|0	30|0|0	28|2|0	30|0|0	29|1|0

Source: developed by the authors based on Dieharder test results

PCG showed excellent results by passing all the main tests with the PASSED rating. The algorithm performed particularly well in the diehard_operm5, diehard_rank_32x32, sts_monobit, and rgb_lagged_sum tests, which highlights its reliability and stability in most tests. Out of the entire list of tests, only rgb_lagged_sum had weak and one failed run, which makes PCG one of the best algorithms among the analysed ones according to the Dieharder test results.

WELL512a also showed stable results, passing most passed tests, including complex tests such as diehard_squeeze and rgb_minimum_distance. However, in the diehard_runs and rgb_lagged_sum tests, a WEAK score was obtained, which indicates some difficulties in maintaining a uniform distribution on large sequences.

Mersenne Twister also passed most tests, including diehard_bitstream, diehard_operm5, and sts_serial, with good p-value values, indicating its high randomness in most cases. However, some tests, in particular marsaglia_tsang_gcd and diehard_birthdays, showed a WEAK estimate indicating probable periodicity or less stable randomness for specific patterns.

KISS passed most tests with high p-value values, which indicates a high level of randomness, especially in template creation tests such as diehard_birthdays and diehard_rank_6x8. However, in the diehard_count_1s_str and diehard_sums tests, the algorithm showed weaknesses with a WEAK score, which may indicate problems in generating statistically uniform sequences for large amounts of data.

Xoshiro128 ++ performed well in the tests, passing them with high p-value values, especially in diehard_operm5 and rgb_bitdist. However, several tests, such as rgb_bitdist and rgb_permutations, gave WEAK scores, which may indicate some shortcomings in ensuring uniform randomness for specific structural patterns.

Analysis of Dieharder's results shows that PCG and WELL512a are the most stable and reliable generators in terms of randomness, successfully passing most tests without significant deviations. Mersenne Twister and Xoshiro128 ++ also showed a high level of randomness, but had weaknesses in some specific tests.

TestU01. Based on TestU01 tests for various PRNG algorithms, randomness efficiency was compared for the KISS, Mersenne Twister, PCG, WELL512a, and Xoshiro128 ++ algorithms. The main criterion for evaluating the success of passing the test is the p-value: to successfully pass the test, the value must be within $0.01 \leqslant p\text{–value} \leqslant 0.99$ (TestU01 official documentation (n.d.); Raza & Satpute, 2017). A p-value that goes beyond these limits usually indicates a deviation from statistical randomness, and a p-value close to 0 or 1 indicates the probability of a non-random nature of the sequence (Table 4).

Table 4. Results of TestU01 statistical tests

Test name	PCG	WELL512a	Mersenne Twister	KISS	Xoshiro128++
Results of CollisionOver	8|0|0	7|1|0	8|0|0	8|0|0	8|0|0
scomp_LempelZiv	5|0|0	5|0|0	5|0|0	5|0|0	5|0|0
scomp_LinearComp	4|0|0	2|0|2	2|0|2	4|0|0	4|0|0
sknuth_CouponCollector	0|1|3	2|1|1	2|1|1	2|0|2	1|0|3
sknuth_Gap	0|0|4	0|0|4	0|0|4	0|0|4	0|0|4
sknuth_MaxOft	17|0|1	17|0|1	17|0|1	17|0|1	17|0|1
sknuth_Run	2|0|0	2|0|0	2|0|0	2|0|0	2|0|0
sknuth_SimpPoker	2|0|2	1|1|2	0|2|2	1|1|2	2|1|1
smarsa_BirthdaySpacings	6|1|0	7|0|0	7|0|0	7|0|0	7|0|0
smarsa_GCD	2|0|0	2|0|0	2|0|0	2|0|0	2|0|0
smarsa_MatrixRank	6|0|0	4|0|2	5|0|1	6|0|0	6|0|0
smarsa_Savir2	0|1|0	1|0|0	1|0|0	1|0|0	1|0|0
smultin_Multinomial	4|0|0	4|0|0	4|0|0	4|0|0	4|0|0
smultin_MultinomialOver	0|0|2	0|0|2	0|0|2	0|0|2	0|0|2
snpair_ClosePairs	17|0|0	17|0|0	17|0|0	17|0|0	17|0|0
snpair_ClosePairsBitMatch	2|0|0	2|0|0	2|0|0	2|0|0	1|1|0
sspectral_Fourier3	6|0|0	6|0|0	6|0|0	6|0|0	6|0|0
sstring_AutoCor	15|4|1	18|2|0	20|0|0	20|0|0	20|0|0
sstring_HammingCorr	3|0|0	3|0|0	3|0|0	3|0|0	2|0|1
sstring_HammingIndep	5|0|1	5|0|1	5|0|1	5|0|1	5|0|1
sstring_HammingWeight2	8|0|0	8|0|0	8|0|0	8|0|0	8|0|0
sstring_LongestHeadRun	4|0|0	4|0|0	4|0|0	4|0|0	4|0|0
sstring_PeriodsInStrings	2|0|0	2|0|0	1|1|0	2|0|0	2|0|0
sstring_Run	3|0|1	4|0|0	3|0|1	3|1|0	3|1|0
svaria_AppearanceSpacings	2|0|0	2|0|0	2|0|0	2|0|0	2|0|0
svaria_SampleCorr	1|0|0	1|0|0	1|0|0	1|0|0	1|0|0
svaria_SampleMean	3|0|0	3|0|0	3|0|0	3|0|0	3|0|0
svaria_SampleProd	2|0|0	2|0|0	2|0|0	2|0|0	2|0|0
svaria_SumCollector	1|0|0	0|1|0	1|0|0	1|0|0	0|1|0
svaria_WeightDistrib	0|1|3	0|0|4	1|0|3	0|0|4	0|1|3
swalk_RandomWalk1	29|1|0	29|1|0	29|1|0	30|0|0	29|0|1

Source: developed by the authors based on TestU01 test results

PCG shows excellent results by passing most TestU01 tests with a p-value that is in the range of 0.01 and 0.99, for example, the value of 0.61 in the CollisionOver test. Thus, PCG can be considered one of the most stable PRNGs, which provides high-quality randomness without significant deviations in tests.

The WELL512a also successfully passed most tests with a p-value within the acceptable range, such as a value of 0.54 in the CollisionOver test. However, in some tests, WELL512a showed a p-value close to the limit, which may indicate possible instability under certain conditions.

Mersenne Twister also showed stable results with p-value within the acceptable range in most tests. For example, the CollisionOver test for Mersenne Twister showed a p-value of 0.90, which indicates its good randomness. However, there were tests where the p-value was close to 0 or 1, which can be an indicator of periodic patterns under certain conditions.

The KISS algorithm successfully passed most of the tests. In some tests, such as CollisionOver, p-value values were within acceptable limits (for example, p-value 0.70), which indicates acceptable randomness. However, in some cases, the algorithm showed a p-value close to the acceptability limit, which may indicate potential problems with certain sequence structures. However, the KISS algorithm has the least non-completed tests compared to other algorithms.

Xoshiro128 ++ showed good randomness results in most tests, getting p-value values in the range of 0.01 and 0.99, for example, a value of 0.32 in CollisionOver. The algorithm had small deviations in some tests, but generally showed stable randomness, which makes it acceptable for use in noise generation problems.

PCG and KISS are the most stable generators according to TestU01 results, providing reliable randomness. Mersenne Twister and Xoshiro128 ++ also showed high results, but with some deviations under certain conditions.

Results of generated noise

1. On autocorrelation graphs (Fig. 1) for each algorithm (KISS, PCG, WELL512a, Xoshiro128 ++, Mersenne Twister), it can be seen that all PRNGs have a sharp peak at zero latency, and then autocorrelation quickly decreases to zero for other delays. This suggests that the generators do not have significant internal correlations, which confirms the randomness of their sequences. The rapid reduction of the autocorrelation coefficient to zero is an important indicator of qualitative randomness for digital noise, which is used to mask signals in cybersecurity.

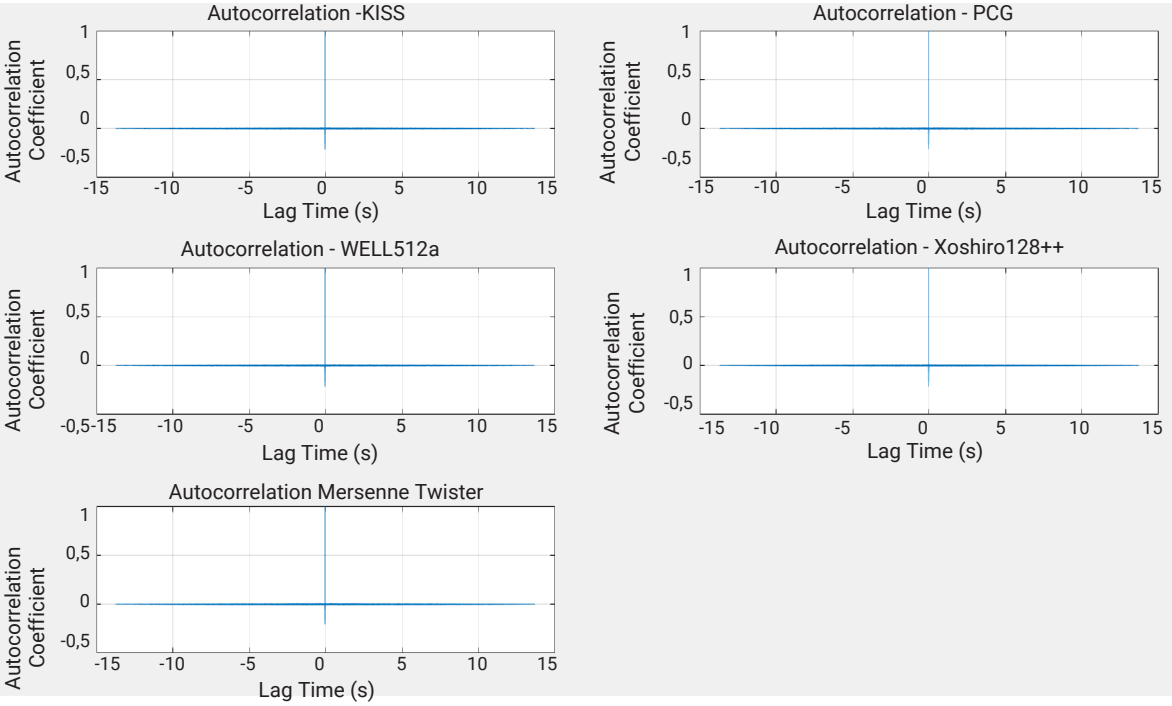

Figure 1. Results of autocorrelation functions
Source: developed by the authors in the Matlab environment

The results show the same Zero Crossing Lag value for all PRNGs at -13.6994 seconds. This may indicate the stability and similarity of the behaviour of each generator in creating a sequence of random numbers. This result is important for determining the absence of long-term correlations, which indicates a good quality of randomness.

2. Spectral power density before filtering (Fig. 2). Before filtering, the PSD results for all PRNGs show that the power is evenly distributed over a wide frequency range, without a drop in the high frequency range. This is typical for well random noise signals, where power is evenly distributed over frequencies. This uniformity before filtering is a confirmation of the high quality of randomness of all algorithms.

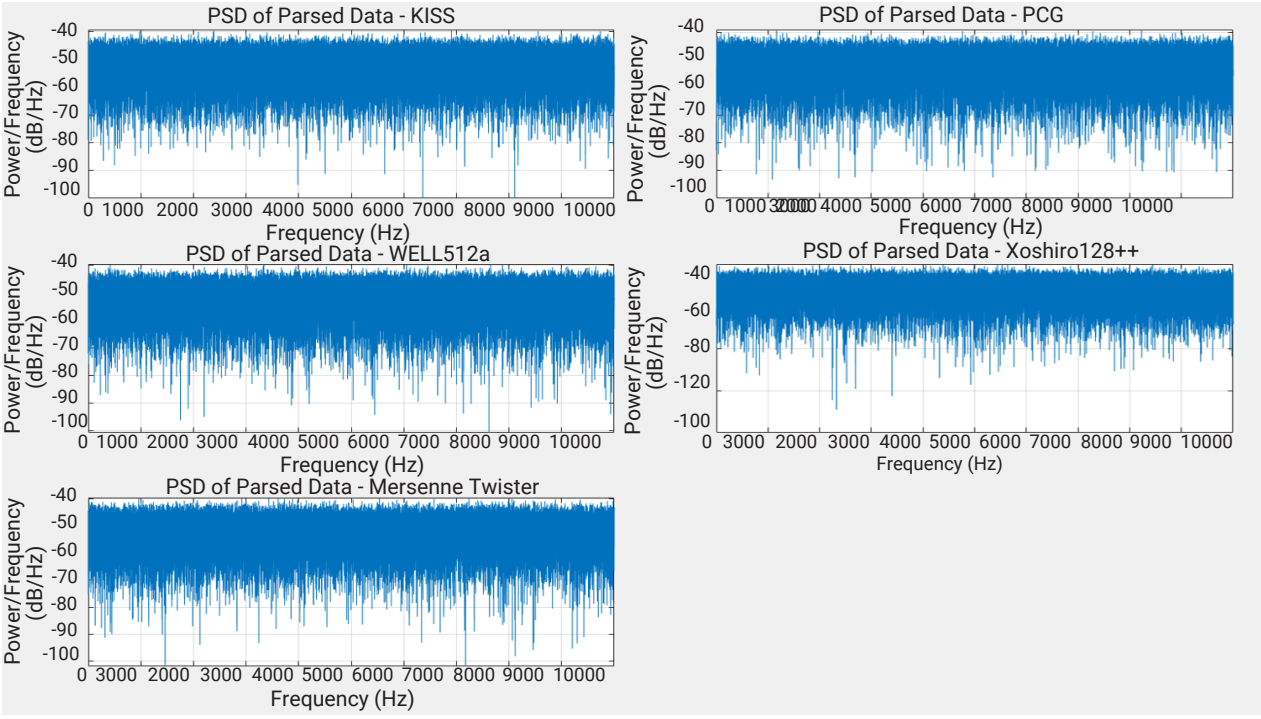

Figure 2. PSD of PRNG results, white noise results
Source: developed by the authors in the Matlab environment

3. Power spectral density (Fig. 3) and frequency noise analysis (Fig. 4) in the specified frequencies (0-4,000 Hz). Filtered PRNG results show a decrease at high frequencies, which is consistent with the expected filtering effect. In each case, a decrease in power at high frequencies is visible, which confirms the successful use of a filter to limit the frequency range of noise. This allows using filtered noise as corresponding to the specified frequency limits to add to the useful signal in masking problems.

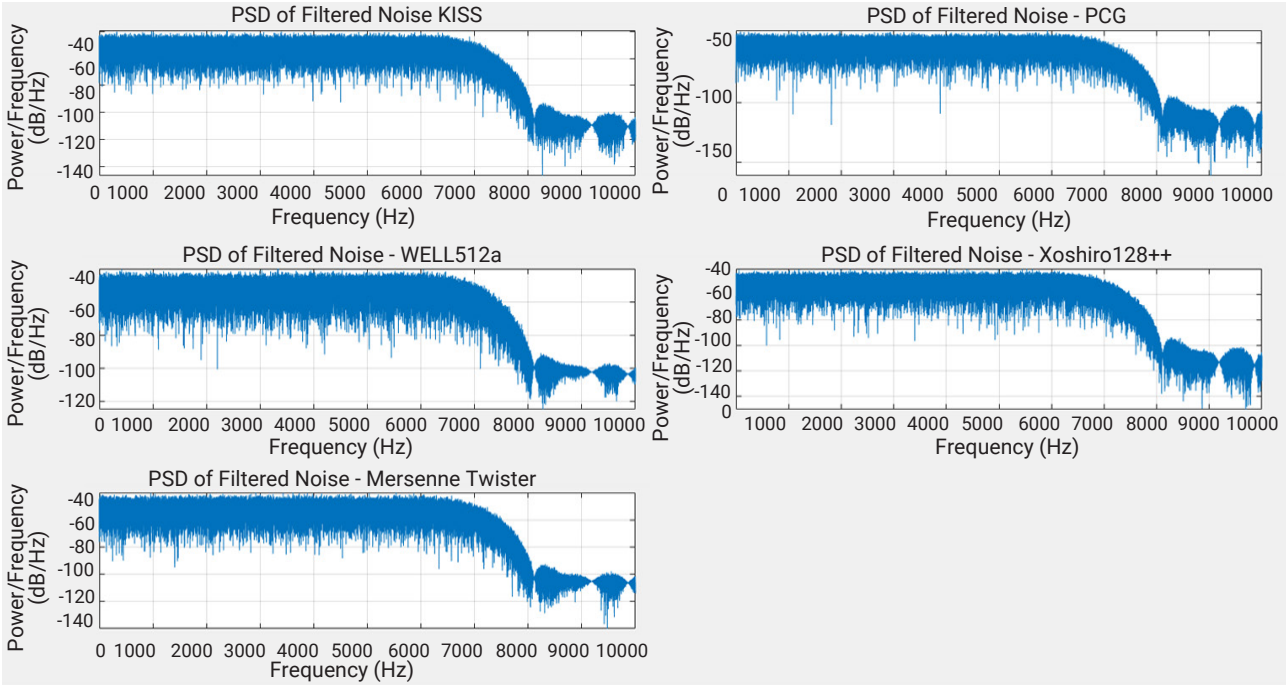

Figure 3. PSD of filtered noise

Source: developed by the authors in the Matlab environment

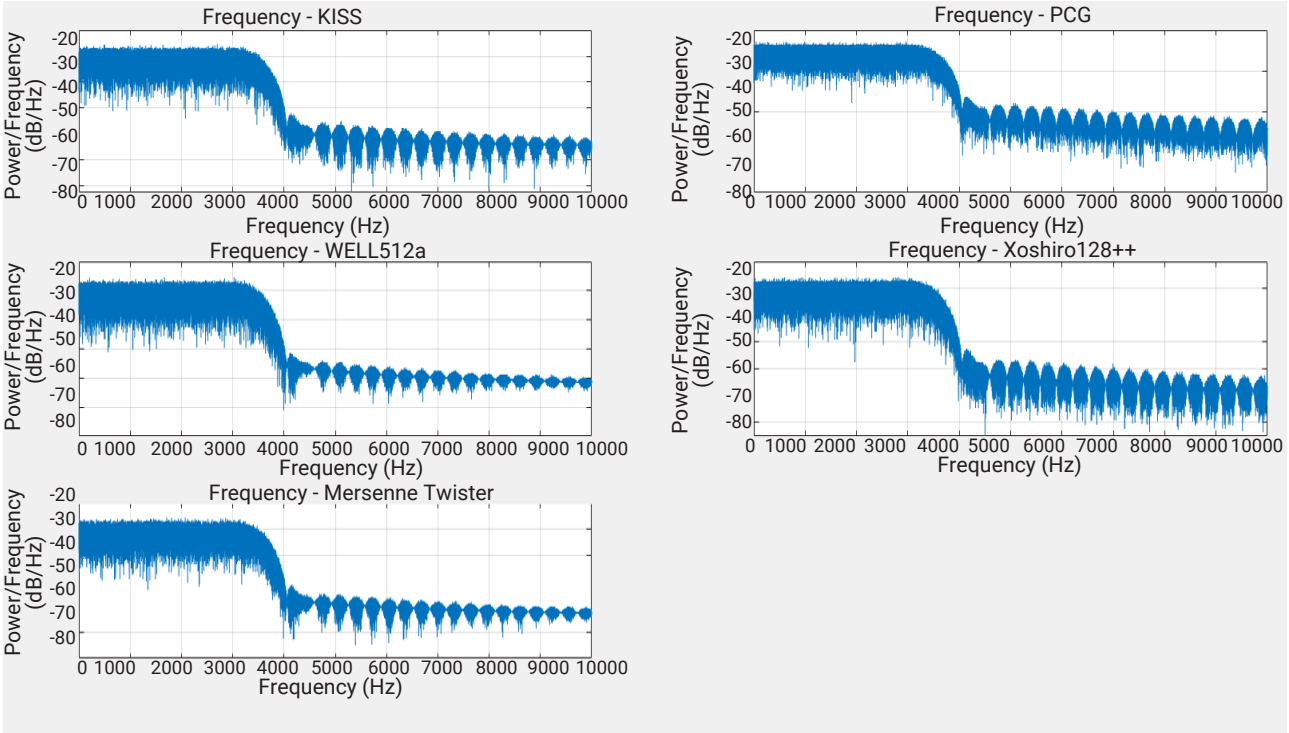

Figure 4. Frequency analysis of filtered noise

Source: developed by the authors in the Matlab environment

4. Signal-to-noise ratio was calculated by equation (1). The results show a close SNR value for each PRNG, with an average value of about -13.6 dB. This indicates the same efficiency of each generator for superimposing noise on the useful signal. A low SNR value indicates strong masking, which is important for added security, as the useful signal becomes less legible under the noise layer.

5. Spectrogram of noise superimposed on a conversation. The spectrograms of each PRNG superimposed on the conversation show that the generated noise completely overlaps the most intense frequencies (Fig. 5). The noise from each generator provides uniform masking, although small differences in the intensity and uniformity of noise distribution between the generators can be seen on the spectrogram. This indicates that all PRNGs provide reliable conversation masking, which makes it difficult to identify the original signal.

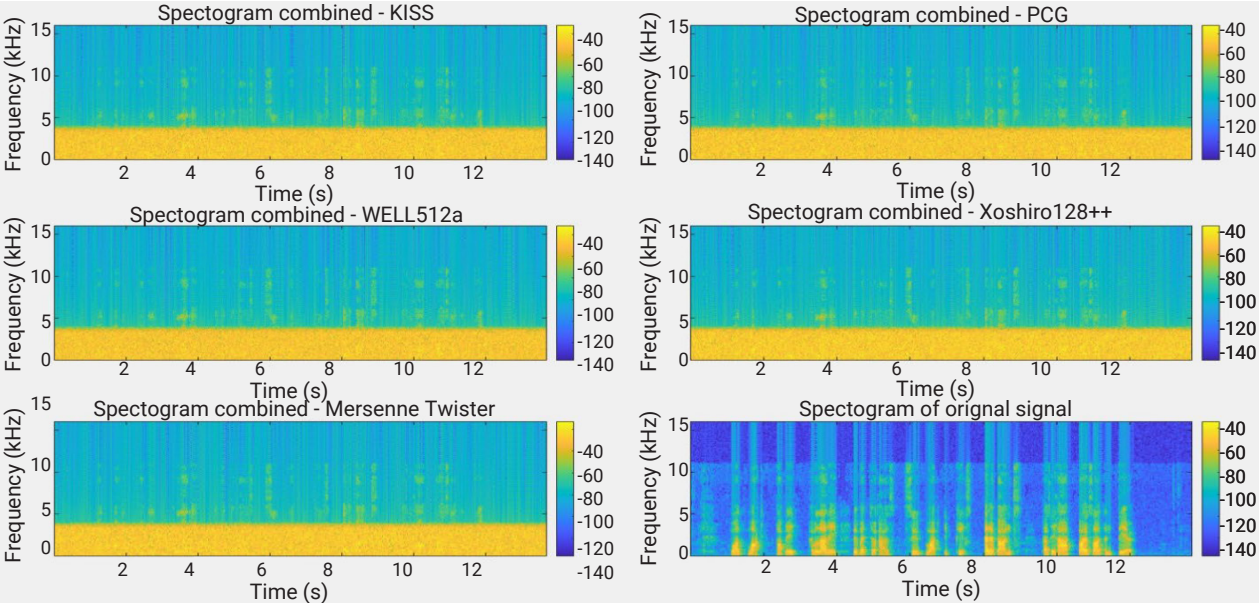

Figure 5. Frequency analysis of filtered noise

Source: developed by the authors in the Matlab environment

The results showed that all the considered PRNG algorithms (KISS, PCG, WELL512a, Xoshiro128 ++, Mersenne Twister) demonstrate a high level of randomness and provide reliable masking of the useful signal, which is confirmed by their autocorrelation characteristics, frequency distribution, and SNR. Minor differences in power spectral densities between algorithms can be considered when selecting a specific PRNG for specific tasks, but in general, all generators demonstrated an appropriate level of quality for signal masking purposes in cybersecurity.

Discussion

This study performed a comparative analysis of five pseudorandom number generators (PCG, Xoshiro128 ++, WELL512a, Mersenne Twister, and KISS) in terms of performance, randomness, spectral characteristics, and noise masking efficiency in cybersecurity tasks. Based on the conducted tests – statistical (NIST, Dieharder, TestU01), spectral analysis, Snr calculations and autocorrelation function analysis – comprehensive data were obtained that allow us to evaluate the quality and effectiveness of these algorithms in generating digital noise.

The results of performance tests showed that the PCG algorithm showed the highest efficiency with an average generation time of 740.8 ms for 268,435,456 values, which makes it an attractive choice for tasks where high performance is important. KISS, with a score of 867.0 Ms, also showed good speed indicators. Other generators, such as the Mersenne Twister and WELL512a, had a longer runtime, which may limit their use in systems where the generation rate is a critical parameter. Comparison with literature sources showed that previous studies have also noted the high performance of PCG and KISS in problems that require rapid random number generation (L'ecuyer, 2017).

Statistical tests by NIST, Dieharder, and TestU01 showed that all generators pass most tests with high p-value values, which indicates a high level of randomness of the generated sequences. However, some tests indicate minor deviations for Xoshiro128 ++ and Mersenne Twister in Dieharder tests, in particular, in "marsaglia_tsang_gcd", which may indicate a certain structure in the generated sequences. Differences in the results for the WELL512a and PCG, which passed all high-performance tests, indicate their high randomness quality for digital noise. These results are consistent with data from previous studies (Bhattacharjee & Das, 2022), which confirm the stability of PCG and WELL512a in random generation. In the mentioned study, among the list of 29 PRNG algorithms, PCG received the second, and WELL512a – the fifth place according to the results of statistical characteristics. The uniform

distribution of values in graphical tests is confirmed by the spectrogram in this paper, where the uniform distribution of frequencies within the specified limits allowed masking the speech signal.

PSD analysis before and after filtering showed that all generators provide uniform power distribution over a wide frequency range before filtering, which is typical for random noise. After applying filtration in the specified frequency ranges, a significant power reduction was achieved in the high-frequency range, which confirms the filtration efficiency. This makes the generated noise more suitable for masking tasks, limiting the effect on the high-frequency components of the useful signal. It is known that a uniform distribution of noise power is a prerequisite for its use in masking problems, which is confirmed by other studies in this field (Kajikawa *et al*., 2012; Deza & Ihshaish, 2021).

The SNR results showed slight differences between the generators, with an average value of about -13.6 dB for all algorithms. This indicated a similar efficiency of each generator in masking the useful signal, since a low SNR level indicates effective masking. However, the difference in SNR of less than 0.1 dB is not statistically significant, which indicates a similar level of noise influence on the useful signal for each of the generators. S.M. Kuo *et al*. (2010) confirmed that SNR in the range from -10 to -20 dB are acceptable for masking problems, which provides sufficient masking without significant loss of useful signal quality.

Analysis of the autocorrelation function showed that all generators have a significant peak at zero latency and a rapid decrease in correlation values for other delays, which is a sign of high randomness and the absence of internal patterns in the sequences. This is an important factor for problems where low correlation in long noise sequences is required to avoid the appearance of periodicity or patterns that can affect the quality of masking. The study by F. Yu *et al*. (2021) also indicated that the presence of a sharp peak at zero latency and low correlations for other delays are characteristic features of high-quality random number generators.

The corator spectrogram superimposed on the conversation recording showed a uniform frequency distribution of noise and effective masking of the low-frequency components of the useful signal. This indicates that all PRNGs provide reliable masking, making the useful signal less legible against the background of noise. Minor differences in power distribution on the spectrogram may be related to the characteristics of each generator, but do not significantly affect the overall quality of masking (Mandal, 2022). The mentioned study describes the implementation of a cryptographic PRNG, the purpose of which is to hide data by superimposing noise. The results showed that the noise level may differ depending on the data that needs to be hidden. However, it was enough to make it impossible to reproduce the original information.

The results of the current study confirmed the data obtained by other authors regarding the randomness and efficiency of various pseudorandom number generators. In particular, J.D. Cook (2017) showed in his paper that the PCG generator shows high performance and stable randomness, which is consistent with the data of this study, where PCG showed the best result in speed (740.8 ms for a large number of values) while maintaining the quality of randomness in statistical tests.

The study by P. L'ecuyer (2017), devoted to the analysis of the WELL512a generator, also showed that WELL512a provides reliable results in problems where high randomness is important for long generation periods. Current results confirm this, as the WELL512a has passed all major randomness tests, providing a stable spectral distribution over the frequency range, which makes it attractive for problems of masking a useful signal with noise.

According to the study by M. Saito & M. Matsumoto (2008**)**, Mersenne Twister has excellent randomness in large samples, but can exhibit certain periodic patterns in specific tests, such as marsaglia_tsang_gcd. The results of the current study confirmed these results, since Mersenne Twister successfully passed most of the tests, but found weaknesses in some specialised Dieharder tests, the results are consistent with the above study.

Conclusions

This study performed a comparative analysis of five pseudorandom number generators (PCG, Xoshiro128 ++, WELL512a, Mersenne Twister, and KISS) for digital noise generation tasks focused on signal masking in cybersecurity. The purpose of the study was to determine the optimal PRNG in terms of speed, randomness, and masking efficiency. The PCG and KISS generators showed the highest performance in performance tests, making them suitable for applications where fast generation of large amounts of random data is critical. PCG, with an average generation time of 740.8 ms for 268,435,456 values, proved to be the most effective.

All the generators under consideration passed the key statistical tests of NIST, Dieharder, and TestU01, which indicates their ability to generate qualitative random sequences. However, Xoshiro128 ++ and Mersenne Twister showed minor deviations in some specialised Dieharder tests, which may indicate the presence of certain structural features in their sequences that may be noticeable under certain conditions. Spectral analysis before and after filtration showed that all generators provide an even distribution of pre-filtration power, and filtering reduces power in the high-frequency range. This confirms the suitability of the generated noise to mask signals in a given frequency range.

The SNR values for all generators were close to -13.6 dB, which indicates the effectiveness of noise signal masking for each algorithm. The small difference in SNR between the generators indicates a similar ability of each algorithm to mask signals in environments with similar conditions. Analysis of the autocorrelation function revealed a high level of randomness with low correlation values for all generators outside the zero lag. This indicates the absence of noticeable internal correlations in the generated sequences, which is important for

creating qualitative noise that does not contain periodic or template components. The spectrogram showed a uniform frequency distribution of noise for each of the generators, which provides effective masking of the low-frequency components of the useful signal. Minor differences in power distribution between generators did not significantly affect the overall masking efficiency.

Considering the results of all experiments, it can be concluded that the best option for generating digital noise is the PCG algorithm. If the preference between performance and statistical tests is given to tests, then the WELL512a algorithm is the best option, but it is the slowest among the tested algorithms and, according to the results of the test, 2.86 times slower than PCG.

Further research may be aimed at analysing other, less common PRNG algorithms to evaluate their effectiveness in masking tasks and at developing new noise filtering methods that will allow adapting its spectral characteristics to specific conditions of use. In addition, it is promising to investigate dynamic filtering parameters to improve SNR depending on the type of useful signal, and to study the effect of different frequency ranges of noise on the effectiveness of masking in various cybersecurity applications. It is also worth noting the trend of trying to develop a new framework that will facilitate and automate the process of further research to find new PRNGs.

Acknowledgements

None.

Conflict of Interest

None.

References

[1] Balalaieva, O., Marchenko, I., Korotenko, G., Beshta, D., & Pikuz, A. (2023). Performance research of C# programming language data serializers using the developed software product for testing. *Reporter of the Priazovskyi State Technical University. Section: Technical Sciences,* 47, 8-24. doi: 10.31498/2225-6733.47.2023.299923.

[2] Bassham, L.E. *et al*. (2010). *A statistical test suite for random and pseudorandom number generators for cryptographic applications*. Gaithersburg: National Institute of Standards and Technology. doi: 10.6028/nist.sp.800-22r1a.

[3] BenchmarkDotNet official documentation. (n.d.). Retrieved from https://surl.li/yobqul.

[4] Bhattacharjee, K., & Das, S. (2022). A search for good pseudo-random number generators: Survey and empirical studies. *Computer Science Review*, 45, article number 100471. doi: 10.1016/j.cosrev.2022.100471.

[5] Cook, J.D. (2017). *Testing the PCG random number generator*. Retrieved from https://surl.li/vjlfia.

[6] Deza, J.I., & Ihshaish, H. (2021). Qnoise: A generator of non-gaussian colored noise. *SSRN Electronic Journal*. doi: 10.2139/SSRN.3975571.

[7] Dieharder official documentation with test suit. (n.d.) Retrieved from https://surl.li/gruuvy.

[8] Feali, M.S. (2023). Realization of a pseudo-random number generator utilizing two coupled Izhikevich neurons on an FPGA platform. *Analog Integrated Circuits and Signal Processing,* 119(1), 57-68. doi: 10.1007/s10470-023-02223-2.

[9] Filho, N. (2024). Performance analysis in Csharp with BenchmarkDotNet: Report and evaluation. *ZENODO,* 1(12). doi: 10.5281/ZENODO.13826811.

[10] Hu, Z. (2020). High-speed and secure PRNG for cryptographic applications. *International Journal of Computer Network and Information Security (IJCNIS)*, 12(3), 1-10. doi: 10.5815/ijcnis.2020.03.01.

[11] Isakov, O.V., & Voitusik, S.S. (2023). Comparative analysis of digital noise generated by additive Fibonacci generators. *Ukrainian Journal of Information Technology*, 5(1), 67-76. doi: 10.23939/ujit2023.01.067.

[12] Kajikawa, Y., Gan, W.-S., & Kuo, S.M. (2012). Recent advances on active noise control: Open issues and innovative applications. *APSIPA Transactions on Signal and Information Processing*, 1, article number e3. doi: 10.1017/ATSIP.2012.4.

[13] Kuo, S.M., Kuo, K., & Gan, W.S. (2010). Active noise control: Open problems and challenges. In *The 2010 International conference on green circuits and systems* (pp. 164-169). Shanghai: IEEE. doi: 10.1109/icgcs.2010.5543076.

[14] L'Ecuyer, P. (2017). History of uniform random number generation. In *2017 Winter simulation conference (WSC)* (pp. 202-230). Las Vegas: IEEE. doi: 10.1109/wsc.2017.8247790.

[15] Li, S., Lin, Z., Yang, Y., & Ning, R. (2024). A high-performance FPGA PRNG based on multiple deep-dynamic transformations. *Entropy,* 26(8), article number 671. doi: 10.3390/e26080671.

[16] Mandal, K. (2022). Cryptographic pseudorandom noise generators for lattice-based cryptography and differential privacy. In *2022 10th international workshop on signal design and its applications in communications (IWSDA)* (pp. 1-4). Colchester: IEEE. doi: 10.1109/iwsda50346.2022.9870587.

[17] Panneton, F., L'Ecuyer, P., & Matsumoto, M. (2006). Improved long-period generators based on linear recurrences modulo 2. *ACM Transactions on Mathematical Software*, 32(1), 1-16. doi: 10.1145/1132973.1132974.

[18] Raza S.F., & Satpute V.R. (2018). PRaCto: Pseudo random bit generator for cryptographic application. *KSII Transactions on Internet and Information Systems*, 12(12). doi: 10.3837/TIIS.2018.12.029.

[19] Saito, M., & Matsumoto, M. (2008). SIMD-oriented Fast Mersenne Twister: A 128-bit pseudorandom number generator. In A. Keller, S. Heinrich & H. Niederreiter (Eds.), *Monte Carlo and Quasi-Monte Carlo methods 2006* (pp. 607-622). Berlin-Heidelberg: Springer. doi: 10.1007/978-3-540-74496-2_36.

[20] Sound data sets. (n.d.). Retrieved from https://commonvoice.mozilla.org/en/datasets.
[21] Syafalni, I., Jonatan, G., Sutisna, N., Mulyawan, R., & Adiono, T. (2022). Efficient homomorphic encryption accelerator with integrated PRNG using low-cost FPGA. *IEEE Access*, 10, 7753-7771. doi: 10.1109/access.2022.3143804.
[22] TestU01 official documentation. (n.d.). Retrieved from https://surl.li/nemayh.
[23] Vennos, A., George, K., & Michaels, A. (2021). Attacks and defenses for single-stage residue number system PRNGs. *IoT*, 2(3), 375-400. doi: 10.3390/iot2030020.
[24] Yu, F., Zhang, Z., Shen, H., Huang, Y., Cai, S., Jin, J., & Du, S. (2021). Design and FPGA implementation of a pseudo-random number generator based on a hopfield neural network under electromagnetic radiation. *Frontiers in Physics*, 9. doi: 10.3389/fphy.2021.690651.

Порівняльний аналіз результатів генераторів псевдовипадкових чисел для генерації цифрового шуму

Олександр Ісаков

Аспірант
Національний університет «Львівська політехніка»
79013, вул. Степана Бандери, 12, м. Львів, Україна
https://orcid.org/0009-0007-4632-9492

Степан Войтусік

Кандидат фізико-математичних наук, доцент
Національний університет «Львівська політехніка»
79013, вул. Степана Бандери, 12, м. Львів, Україна
https://orcid.org/0000-0003-4234-3303

Анотація. У статті викладено результати дослідження характеристик п'яти різних генераторів псевдовипадкових чисел для застосування в задачах генерації цифрового шуму, який використовується для маскування сигналів у кібербезпеці. Актуальність роботи зумовлена зростаючою потребою у високоякісних методах маскування, які забезпечують як ефективну продуктивність, так і надійність випадковості, що важливо для захисту конфіденційної інформації у сучасних цифрових системах. Метою дослідження було порівняння алгоритмів PCG, Xoshiro128++, WELL512a, Mersenne Twister та KISS за показниками їхньої швидкодії, статистичної випадковості та здатності ефективно маскувати корисний сигнал шумом. Швидкодія алгоритмів оцінювалася за допомогою BenchmarkDotNet. Для перевірки якості випадковості послідовностей використовувалися стандартні тести NIST, Dieharder та TestU01. Для згенерованого шуму проведено спектральний аналіз за допомогою значення спектральної щільності потужності. Ефективність маскування було розраховано співвідношенням сигнал/шум, результатами автокореляційної функції і спектрограми шуму. Результати дослідження показали, що PCG та KISS є найбільш продуктивними з точки зору швидкодії, що робить їх привабливими для застосувань, де важлива швидка генерація випадкових послідовностей. WELL512a та PCG продемонстрували найвищу якість випадковості, стабільно проходячи всі статистичні тести. Аналіз спектрального розподілу шуму показав, що всі генератори забезпечують рівномірний розподіл потужності до фільтрації, а після фільтрації шум успішно обмежується у високочастотному діапазоні. Співвідношення значення сигналу до шуму для всіх алгоритмів становили близько -13.6 dB, що вказує на подібну ефективність при маскуванні шумом. Автокореляційний аналіз підтвердив низьку кореляцію для всіх генераторів за межами нульового лагу, що є важливим для збереження якості випадковості в довгих послідовностях. Практична цінність дослідження полягає у виборі оптимального генератора псевдовипадкових чисел для задач зашумлення в кібербезпеці. Отримані результати надають рекомендації щодо вибору алгоритмів з урахуванням їхньої швидкодії та випадковості, що дозволить забезпечити високий рівень захисту інформації у цифрових системах

Ключові слова: захист інформації; шумові характеристики; статистичні тести випадковості; спектральний аналіз; тести продуктивності; зашумлення сигналу

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

Information Technologies and Computer Engineering **Vol. 21, No. 3. 2024**

Article's History: Received: 03.09.2024 Revised: 26.11.2024 Accepted: 26.12.2024

UDC 004.8 10.63341/vitce/3.2024.65

Optimising fuzzy hash function parameters for ensuring compliance with Open Data Regulations

Leonid Maidanevych

PhD in Philosophical Sciences, Senior Lecturer
Vinnytsia National Technical University
21021, 95 Khmelnytske Shose Str., Vinnytsia, Ukraine
https://orcid.org/0000-0002-7364-8874

Nataliia Kondratenko

PhD in Technical Sciences, Professor
Vinnytsia National Technical University
21021, 95 Khmelnytske Shose Str., Vinnytsia, Ukraine
https://orcid.org/0000-0002-4450-1603

Vitalii Kazmirevskyi*

Postgraduate Student
Vinnytsia National Technical University
21021, 95 Khmelnytske Shose Str., Vinnytsia, Ukraine
https://orcid.org/0009-0005-4056-5385

Abstract. The aim of this study was to investigate the parameters of the hash function to enhance the efficiency and accuracy of detecting similarities in text fragments across various web resources when monitoring compliance with the requirements of the Regulation on Open Data on official government websites. The research focused on assessing three key parameters of the hash function: block size, prime number base, and modulus. To achieve this, a series of experiments was conducted, employing different combinations of these parameters to generate hash values for text data. The results demonstrated which parameter combinations provide the best balance between accuracy, completeness, F-measure, and execution time. The study showed that specific parameter configurations enable a significant improvement in algorithm accuracy while minimising computational costs, which is particularly important for real-time data analysis. It is established that optimising the parameters of the hash function reduces the occurrence of false positives and false negatives, which are common issues in similarity detection. In particular, selecting optimal values for each parameter significantly enhances the accuracy and completeness of the analysis, leading to more precise text fragment comparisons and reduced execution time. This optimisation makes the fuzzy hashing algorithm well-suited for use in automated systems that monitor government websites for compliance with open data regulations. Furthermore, the study found that parameter optimisation decreases the number of duplicate records, which is especially relevant for ensuring that open data adheres to legislative requirements. The conclusions drawn from this research can be applied to the development of software tools designed to efficiently identify deficiencies and improve transparency and legal compliance. Additionally, the findings can contribute to further optimisation of fuzzy hash function algorithms, thereby advancing data monitoring technologies for regulatory compliance. This study enhances the development of web resource monitoring technologies by demonstrating how the careful selection of fuzzy hash function parameters can substantially improve the efficiency and reliability of open data analysis

Keywords: fuzzy hash function parameters; website monitoring; government electronic resources; algorithm accuracy; optimization parameters; similarity detection; violation of provisions

Suggested Citation:
Maidanevych, L., Kondratenko, N., & Kazmirevskyi, V. (2024). Optimising fuzzy hash function parameters for ensuring compliance with Open Data Regulations. *Information Technologies and Computer Engineering*, 21(3), 65-76. doi: 10.63341/vitce/3.2024.65

*Corresponding author

Introduction

In the modern world, open data is a crucial tool for ensuring transparency and accountability in government agencies. However, maintaining compliance with regulations and standards is becoming increasingly challenging due to the growing volume and diversity of information. The quality of open data often suffers from errors, duplication, and non-compliance with legal requirements. One of the primary challenges is the need to rapidly detect and correct deviations in large datasets, a process that is often burdensome due to limited resources. Existing methods for verifying such data do not always operate efficiently or accurately, particularly when handling large volumes of information. Fuzzy hash functions play a significant role in addressing these challenges. They enable the effective identification of similar but non-identical data, which is critical for monitoring and analysing information. Optimising hash function parameters can enhance the accuracy of text analysis, reduce errors, and accelerate data processing, thereby increasing the efficiency of automated monitoring systems. Implementing improved methods for adapting the parameters of fuzzy hash functions can greatly facilitate data management and enhance data quality, contributing to the more effective and reliable enforcement of open data requirements.

In the field of applying fuzzy hash functions to monitor open data compliance, numerous studies have focused on refining data comparison methods. In recent years, various approaches have been proposed for using fuzzy algorithms to detect similar texts and fuzzy duplicates in large datasets. A significant body of research is dedicated to optimising text comparison methods under conditions of data uncertainty. For example, the use of type-2 fuzzy sets, in which the degree of membership is determined by intervals rather than precise values, helps to mitigate the impact of "noise" and improve analytical accuracy. Studies by N.R. Kondratenko & O.O. Snihur (2019) and N.R. Kondratenko (2023) on the application of such methods in fuzzy logic systems have demonstrated that this approach is effective for document analysis in complex conditions where information is incomplete or ambiguous.

Recent studies highlighted the importance of integrating fuzzy hash functions for processing complex data to enhance the efficiency of digital forensics in the Internet of Things (IoT) environment. For example, in the study by W.A. Mahrous *et al.* (2021), an improved digital forensics architecture combining blockchain and fuzzy hashing wasproposed. The application of fuzzy hash functions enables the storage and analysis of IoT data while accounting for its variability and similarity. This is particularly crucial in scenarios where the detection of similar files is critical. The study demonstrated that integrating traditional hashing for authentication with fuzzy hashing facilitates the identification of potentially similar documents that might otherwise remain undetected when using classical methods. This approach allowed for a more efficient analysis of files within a blockchain network by computing the similarity between blocks, thereby enhancing trust in IoT data.

In the 2020s years, cybersecurity and attack identification, particularly in the context of threats to national security, have become critical areas of security research. Traditional methods of attack attribution typically rely on analysing the behaviour of malicious software in isolated environments (sandboxes). However, certain threats can modify their behaviour or even cease functioning upon detection. In this context, M. Kida & O. Olukoya (2023) proposed the use of fuzzy hash functions to automate attack attribution, yielding more accurate results compared to traditional approaches.

A promising direction in cybersecurity is the use of hash functions for malware identification, which helps to overcome the limitations of dynamic analysis, such as sensitivity to the execution environment and delays in log collection. The study by T. Baba *et al*. (2022) demonstrated the effectiveness of hash functions for malware classification. The authors explored the potential of combining fuzzy hashing with deep learning for handling large datasets. Their findings indicated that this approach enhances robustness against data changes, which frequently occur in the chronological logs of PE file properties. Notably, the study concluded that integrating surface-level information from PE files with hash function values achieves the highest performance in malware classification. Consequently, the use of fuzzy hashing within deep learning frameworks presents new opportunities for analysing and mitigating cyber threats, making this approach valuable for developing modern antivirus systems.

Beyond cybersecurity, the application of fuzzy hash functions in neural networks has shown promise for improving the quality of X-ray images, opening up new possibilities for automated disease detection, including COVID-19. The approach proposed by A. Nandal *et al.* (2021) integrates hash functions to effectively measure distances between image features, thereby minimising classification errors. Fuzzy hashing enhances the processing of X-ray images by reducing noise and emphasising textural features. This method has demonstrated high accuracy (up to 95.68%) and holds potential for broader applications in medical research.

Fuzzy hashes have also been utilised in network server testing. The study by R. Natella (2022) introduced STATEAFL, a "greybox fuzzer" designed to automatically adapt to servers without requiring manual configuration. This method leverages fuzzy hashing to generate unique identifiers for server states, enabling the construction of a protocol state machine based on memory and network I/O data. The author's research confirmed that this approach to fuzzing is not only automated but also more accurate than traditional methods that rely solely on analysing server responses. Consequently, fuzzy hashes prove to be an effective tool for complex tasks such as testing server states, offering automation and high efficiency.

Fuzzy hashes are a relatively new tool that is actively evolving across various fields. However, their principles of

operation and effectiveness remain insufficiently understood. In the study by M. Martín-Pérez *et al*. (2021), an important attempt was made to systematise research in this area by developing a classification of similarity algorithms. This not only enhances the understanding of different approaches but also facilitates more thorough comparisons between them. The authors analysed existing algorithms, with a particular focus on potential threats – investigating possible attacks on similarity algorithms and identifying the conditions under which such attacks could be successful. Despite the growing popularity of fuzzy hashes, the study emphasised that this field remains in a state of active development.

In light of the above, the key challenge is to improve existing methods for detecting fuzzy duplicates and enhance the efficiency of monitoring compliance with open data requirements on official government websites. Implementing an improved approach will ensure high accuracy in fuzzy duplicate detection while maintaining acceptable computational costs, thereby reducing the risks associated with non-compliance with open data legislation. Current approaches often fail to account for all factors influencing the quality of duplicate or anomaly detection, which can lead to inaccurate results. Therefore, studying the parameters of the fuzzy hash function is crucial for enhancing the effectiveness of compliance monitoring.

The aim of this study was to investigate the parameters of the fuzzy hash function for monitoring compliance with Open Data Regulations on official government websites, with a focus on improving duplicate detection. The research seeked to adapt the parameters of the fuzzy hash function to detect similarities in text fragments and enhance the effectiveness of compliance monitoring with Open Data Legislation on official government websites. This, in turn, will enable the timely and accurate identification of violations in the publication of open data.

Materials and Methods

The study utilised data from official web resources that are subject to mandatory publication in accordance with the Regulation on Data Sets Subject to Disclosure in the Form of Open Data (Resolution of the Cabinet of Ministers of Ukraine No. 835, 2015) and current Ukrainian legislation. The data sources included the Verkhovna Rada of Ukraine, Ministry of Digital Transformation of Ukraine, State Service of Special Communications and Information Protection of Ukraine, National Bank of Ukraine, Pension Fund of Ukraine, Open Data Portal, State Statistics Service of Ukraine, and the State Tax Service of Ukraine.

In the initial stage, a sample of web pages from various government agencies that publish open data in the form of tables, text, and other formats was collected. This data was pre-processed to remove non-essential elements such as scripts, styles, and multimedia inserts, ensuring a focus on meaningful text content. This step helped to minimise "noise" in the data and enhance the accuracy of the results.

In subsequent stages, the performance of the fuzzy hash function was analysed using different sets of parameters, including BlockSize, PrimeBase, and Mod. These parameters enabled an investigation into how the accuracy and speed of the algorithm varied when comparing different versions of textual data on web pages. Particular attention was given to the accuracy and completeness of detecting similar text fragments, as well as the processing time for large volumes of data – an essential factor for real-time monitoring systems. This approach was implemented based on the methods described in the works of M. Fleming *et al*. (2024) and M. Guerrero (2022). The generation of N-blocks (segments) followed the method presented in the studies of A. AlMajali *et al*. (2024) and N. Naik *et al*. (2019a), which involved computing N-blocks in fuzzy hashes for similarity analysis.

The study of fuzzy hash function parameters was a key step in ensuring its effectiveness in monitoring compliance with open data requirements on government websites. The parameter analysis process (Fig. 1) involved evaluating typical usage scenarios, conducting experimental testing, and fine-tuning the parameters based on the results obtained.

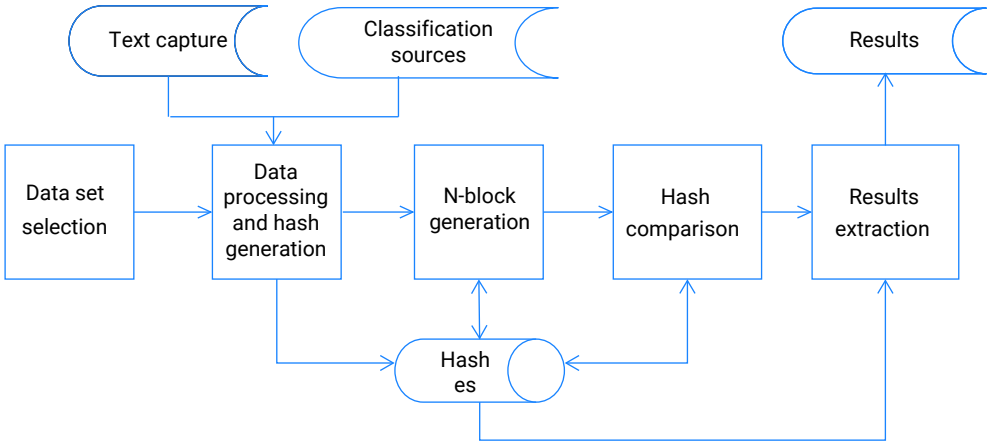

Figure 1. The process of fuzzy hashes calculating and comparing
Source: created by the authors based on N. Naik *et al*. (2019a) and A. AlMajali *et al*. (2024)

When generating N-blocks, the file was divided into multiple blocks, with a hash value calculated for each block. Segmenting the file into blocks helped reduce the impact of local modifications on the overall hash value, thereby enhancing the integrity and accuracy of comparisons. The individual block hash values were then combined to generate a fuzzy hash, a process illustrated in Figure 2. Generating hashes for blocks separately improved efficiency, particularly for large files, as it reduced memory usage and computational resource requirements. By combining the block hashes, a universal fuzzy hash value was created, allowing for the verification of data consistency even in the presence of minor differences.

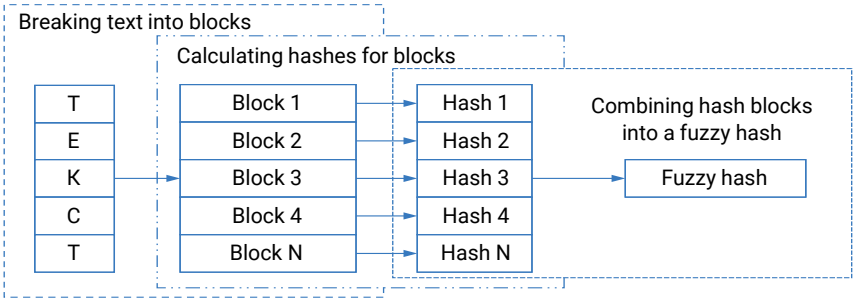

Figure 2. Procedure for generating a fuzzy hash value
Source: created by the authors based on A.P. Namanya *et al*. (2020) and S.R. Davies *et al*. (2021)

The key characteristics that determine the effectiveness of a fuzzy hash function were selected for this study:

Precision – Measures how accurately the algorithm identifies correct matches;

Recall – Reflects the algorithm's ability to detect all relevant matches;

Execution time – Influences overall performance and the speed of analysis.

The ssdeep algorithm was used as the basis for constructing a fuzzy hash function with appropriate adaptation of the parameters for the purposes of the study (Ssdeep-project, n.d.). The main parameters that determine the effectiveness of the fuzzy hash function include: BlockSize, PrimeBase, and Mod. **BlockSize**: determines the size of the text block used to calculate the hash. By changing this parameter, you can control the sensitivity of the algorithm to changes in the text. **PrimeBase**: is used as the basis for calculating the hash. The choice of a prime number affects the uniqueness and distribution of hashes. **Mod** (modulus): a parameter that affects the final form of the hash, avoids collisions, and increases the reliability of the algorithm.

The mathematical model was based on the basic principles defined in the work of N. Naik *et al*. (2019b):

- hash function $G(d)$: a function that converts input data of arbitrary length d into a value of fixed length h;
- fuzzy hash function $F_G(d)$: a modified hash function that allows you to take into account small changes in input data while maintaining close hash values for similar input data;
- similarity metric ($F_G(d_1)$, $F_G(d_2)$): a function that measures the degree of similarity between the hashes of two data sets d_1 and d_2.

The main task solved by the fuzzy hash function in the framework of the study was to identify similarities between different versions of web pages and documents on official websites of government bodies. This was formalized as follows. $D=\{d_1, d_2,\ldots,d_n\}$ – a set of web pages or documents to be analysed. For each element d_i the hash value $h_i=F_G(d_i)$. For two elements d_i and d_j is calculated. The similarity metric is defined, according to *R*. Chanajitt *et al*. (2022), as:

$$S(h_i, h_j) = \frac{sum(F_G(d_i) \cap F_G(d_j))}{\max(sum(F_G(d_i)), sum(F_G(d_j)))}, \quad (1)$$

where *sum*() – element summation function; ∩ – intersection operation.

The general framework of the software tool for analysing the parameters of a fuzzy hash function is presented in Figure 3. The text preprocessing module normalises textual data, removes unnecessary spaces, and eliminates other elements that may influence the hashing outcome. The hash function computation module is responsible for directly calculating the fuzzy hash using adapted parameters (BlockSize, PrimeBase, Mod). The hash comparison module compares the generated hashes to detect changes and deviations in the content, enabling the assessment of the compliance of open data with established requirements. The results visualisation module presents the computation results in a structured format, facilitating analysis and allowing users to assess the degree of textual similarity.

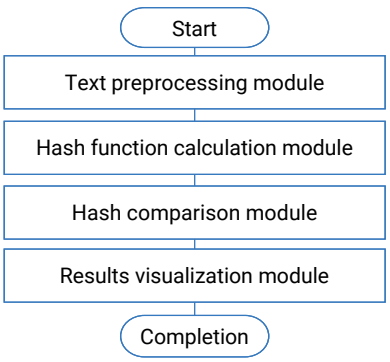

Figure 3. Scheme of a software tool for studying the parameters of a fuzzy hash function
Source: created by the authors

Accuracy determines what proportion of the found elements are actually correct. It is calculated by the formula:

$$Precision = \frac{TP}{TP + FP}, \quad (2)$$

where *TP* – number of correct positives (correctly detected duplicates); *FP* – number of incorrect positives (falsely detected duplicates).

Completeness determines what proportion of all true elements have been discovered. It is calculated by the formula:

$$Recall = \frac{TP}{TP + FN}, \quad (3)$$

where *TP* – number of correct positives; *FN* – number of elements that should have been detected but were not.

The *F*-score is the harmonic mean of precision and completeness, offering a balanced assessment. This measure is particularly useful when both false positives and false negatives must be taken into account. The formula for the *F*-score is:

$$F = 2 \times \frac{Precision \times Recall}{Precision + Recall}. \quad (4)$$

Processing time is measured in milliseconds and indicates the duration required for the algorithm to execute and process messages. It is calculated as:

$$Time = t_{end} - t_{start}, \quad (5)$$

where t_{start} – processing start time; t_{end} – processing completion time.

To implement these indicators, the software records the count of correct results, false positives, and false negatives for each parameter configuration (BlockSize, PrimeBase, Mod), along with the measurement of processing time for each data block.

Results and Discussion

To analyse the impact of fuzzy hash function parameters on the effectiveness of monitoring compliance with the Regulation on Open Data on the official websites of state institutions, an experiment was conducted using real datasets obtained from the following sources: Verkhovna Rada of Ukraine, Ministry of Digital Transformation of Ukraine, State Service of Special Communications and Information Protection of Ukraine, National Bank of Ukraine, Pension Fund of Ukraine, Open Data Portal, State Statistics Service of Ukraine, and State Tax Service of Ukraine. The use of real data ensures the maximum relevance of the results.

The websites of different institutions varied in length and content structure, enabling an assessment of how changes in the parameters BlockSize, PrimeBase, and Mod influence key metrics such as Precision, Recall, F1 Score, and processing time. The results for one of the websites, containing 1,270 characters, are presented in Tables 113 and can serve as a basis for further research and optimisation of open data monitoring. The tables illustrate the impact of each parameter while keeping other indicators constant. To optimise the BlockSize, PrimeBase, and Mod parameters in relation to message length (Message Length), a detailed analysis was conducted to examine the relationships between these parameters and key performance metrics, including Precision, Recall, F1 Score, and processing time (Guerrero, 2022; Fleming *et al.*, 2024). During the study, various parameter combinations were tested for each Message Length value, allowing for the identification of dependencies and the evaluation of each parameter's impact on the algorithm's performance.

Table 1. Results of the influence of BlockSize at PrimeBase and Mod constants

No.	BlockSize	Precision	Recall	F1 Score	Time (ms)
1	10	0.85	0.85	0.85	0.16
2	20	0.80	0.75	0.77	0.17
3	30	0.75	0.70	0.72	0.17
4	40	0.70	0.65	0.67	0.18
5	50	0.65	0.60	0.62	0.18
6	60	0.60	0.55	0.57	0.18
7	70	0.55	0.50	0.52	0.19
8	80	0.50	0.45	0.47	0.21
9	90	0.45	0.40	0.42	0.22
10	100	0.40	0.35	0.37	0.23

Source: created by the authors

The results presented in Table 1 demonstrate the significant impact of the BlockSize parameter on the algorithm's efficiency. As BlockSize increases, there is a gradual decline in Precision, Recall, and F1 Score. The best performance is observed when BlockSize = 10, where Precision and Recall both reach 0.85, resulting in an F1 Score of 0.85, indicating high accuracy and sensitivity to data changes. The processing time for this configuration is only 0.16 ms, highlighting its efficiency at smaller block sizes. However, increasing BlockSize to 100 leads to

a substantial decline in performance, with Precision and Recall dropping to 0.40 and processing time increasing to 0.23 ms. This suggests that larger block sizes negatively impact the model's accuracy, particularly its ability to correctly detect data changes. This decline may be attributed to memory overload and the increased computational demands of processing larger data blocks within a limited time.

Table 2. Results of the influence of PrimeBase at BlockSize and Mod constants

No.	PrimeBase	Precision	Recall	F1 Score	Time (ms)
1	10	0.85	0.85	0.85	0.18
2	20	0.80	0.75	0.77	0.18
3	30	0.75	0.70	0.72	0.18
4	40	0.70	0.65	0.67	0.19
5	50	0.65	0.60	0.62	0.19
6	60	0.60	0.55	0.57	0.18
7	70	0.55	0.50	0.52	0.19
8	80	0.50	0.45	0.47	0.21
9	90	0.45	0.40	0.42	0.21
10	100	0.40	0.35	0.37	0.21

Source: created by the authors

Table 2 presents the results for different PrimeBase values with a fixed BlockSize = 10 and Mod = 10,000,000. The best performance is observed at PrimeBase = 10, where Precision, Recall, and F1 Score all reach 0.85. This suggests that using smaller prime numbers with a text length of 1270 characters enhances accuracy and sensitivity. However, increasing PrimeBase to 50 or higher reduces the algorithm's efficiency, as indicated by a decline in Precision and Recall to 0.40. These results suggest that larger prime numbers may negatively impact computational complexity and model efficiency. Since larger numbers require greater processing resources, this increased computational demand leads to a reduction in accuracy.

Table 3. Results of the influence of Mod at BlockSize and PrimeBase constants

No.	Mod	Precision	Recall	F1 Score	Time (ms)
1	10,000,000	0.85	0.85	0.85	0.14
2	20,000,000	0.80	0.75	0.77	0.14
3	30,000,000	0.75	0.70	0.72	0.13
4	40,000,000	0.70	0.65	0.67	0.15
5	50,000,000	0.65	0.60	0.62	0.17
6	60,000,000	0.60	0.55	0.57	0.18
7	70,000,000	0.55	0.50	0.52	0.19
8	80,000,000	0.50	0.45	0.47	0.22
9	90,000,000	0.45	0.40	0.42	0.22
10	100,000,000	0.40	0.35	0.37	0.23

Source: created by the authors

Table 3 demonstrates that modifying the Mod parameter can significantly impact the algorithm's performance. At Mod = 10,000,000, the best results are observed, with Precision, Recall, and F1 Score all reaching 0.85, confirming the effectiveness of this modulus value. However, increasing Mod to 100,000,000 leads to a substantial decline in performance, with Precision and Recall dropping to 0.40. This deterioration suggests a reduced ability of the model to accurately detect data changes, likely due to the increased computational complexity associated with larger Mod values, which negatively affects accuracy and sensitivity.

From the analysis of the results with a text length of 1270 characters, several conclusions can be drawn. The best performance is achieved with BlockSize = 10, PrimeBase = 10, and Mod = 10,000,000, where high values of Precision, Recall, and F1 Score (0.85) indicate an optimal parameter combination for open data monitoring. This is further supported by the low processing time, demonstrating the algorithm's efficiency under these settings.

Based on the results of this analysis, optimal parameter sets were identified to minimise or maximise the relevant metrics for each message length. Specifically, for each Message Length value, the BlockSize, PrimeBase, and Mod values that provided the best balance between accuracy, completeness, and processing time were determined.

The algorithm for selecting optimal parameters involved the following stages:

1. Data collection. For each message length, performance metrics (Precision, Recall, F1 Score, and Processing Time) were calculated for different combinations of BlockSize, PrimeBase, and Mod parameters.

2. Analysis and comparison of metrics. The overall efficiency of each parameter combination was evaluated, and those providing the best results were selected.

3. Parameter selection. Based on the minimum or maximum values of the metrics, the optimal BlockSize, PrimeBase, and Mod values were determined for each Message Length.

As part of this study, an investigation was conducted into the influence of fuzzy hash function parameters on the effectiveness of monitoring compliance with the Regulation on Open Data on the official websites of government institutions. Fuzzy hash functions enable efficient comparison of similar but non-identical data, which is essential for open data monitoring, where exact duplicates may not always be present. The parameters BlockSize, PrimeBase, and Mod influence the algorithm's sensitivity and accuracy in detecting similarities. For example, adjusting BlockSize can impact the algorithm's ability to detect even minor textual changes, while modifying PrimeBase and Mod can reduce the probability of hash collisions and enhance accuracy. These parameters interact, collectively determining the efficiency and speed of website monitoring for compliance with legal requirements. In particular, Table 4 presents the experimental results illustrating the relationship between key performance indicators (Precision, Recall, F1 Score, and Processing Time) and the selection of BlockSize, PrimeBase, and Mod values across different message lengths.

Table 4. Experimental results. Optimal parameters for adapting the hash function to the specifics of the processed data

No.	Message Length	BlockSize	PrimeBase	Mod	Precision	Recall	F1 Score	Time (ms)
1	1,270	10	10	10,000,000	0.85	0.87	0.86	0.15
2	2,540	20	18	20,000,000	0.90	0.92	0.91	0.16
3	2,990	30	25	30,000,000	0.93	0.94	0.94	0.16
4	3,800	40	32	40,000,000	0.94	0.96	0.95	0.17
5	5,000	50	40	50,000,000	0.95	0.96	0.95	0.17
6	5,720	60	48	60,000,000	0.96	0.97	0.96	0.17
7	7,100	70	68	70,000,000	0.91	0.93	0.92	0.18
8	7,900	80	80	80,000,000	0.92	0.94	0.93	0.18
9	8,500	90	80	90,000,000	0.94	0.95	0.94	0.18
10	9,920	100	95	100,000,000	0.96	0.97	0.96	0.19

Source: created by the authors

The results presented in Table 4 illustrate the optimal BlockSize, PrimeBase, and Mod values for different message lengths. Analysis of the data reveals that an increase in Message Length corresponds to a rise in Precision, Recall, and F1 Score, indicating improved accuracy and sensitivity as the algorithm processes larger datasets. Specifically, for a Message Length of 1,270, Precision is 0.85, increasing to 0.96 for a Message Length of 9,920, highlighting the algorithm's enhanced ability to accurately detect similar texts. Adjusting the BlockSize, PrimeBase, and Mod parameters also has a significant impact on performance. As BlockSize increases with Message Length, the algorithm is able to process larger data fragments, enabling a more precise comparison of text segments. Similarly, increases in PrimeBase and Mod enhance the uniqueness of hash values and reduce the probability of collisions, thereby improving the overall reliability of the algorithm. Since these parameters are interdependent, their optimal configuration ensures the best balance between accuracy, sensitivity, and processing speed. Processing time (Time) also varies with increasing Message Length, which is expected, as larger messages and more complex parameters require more computation. However, the execution time remains within an acceptable range, specifically between 0.15 and 0.19 ms, demonstrating the algorithm's high efficiency, even with increased data volumes. Thus, the experimental results confirm that increasing BlockSize, PrimeBase, and Mod in proportion to Message Length leads to significant improvements in accuracy, sensitivity, and reliability – a crucial factor for effective open data monitoring on government websites.

The findings from the study of fuzzy hash function parameters were integrated into a software tool designed to monitor compliance with the Open Data Regulations on the official websites of state institutions. The generalised architecture of this tool is illustrated in Figure 3. The software tool applies the optimal BlockSize, PrimeBase, and Mod parameters identified through experimentation, allowing the algorithm to adapt to the processing requirements of different volumes of text data. Specifically, the BlockSize and PrimeBase parameters are incorporated into the algorithm to enable dynamic adjustment of the block size and prime number base based on the length of the processed message. This ensures the algorithm remains flexible across diverse data types, maintaining high accuracy in detecting similar text fragments while minimising the probability of errors. Furthermore, adjusting Mod values enhances the function's ability to handle complex datasets, reducing the likelihood of collisions and increasing the uniqueness of computed hash values.

By implementing these optimised parameters, the monitoring system effectively analyses open data, achieving high accuracy and sensitivity in detecting violations. Additionally, parameter optimisation enables the tool to maintain low processing times, even when handling large data volumes, making it suitable for real-time monitoring

of web resources for compliance with legal requirements. To facilitate visual analysis and provide a clearer representation of the impact of these parameters, the results are depicted in Figure 4. This figure illustrates trends in accuracy, completeness, F-measure, and processing time, depending on the configuration of the hash function parameters. The X-axis represents the length of the input message, while the Y-axis shows the corresponding parameter values.

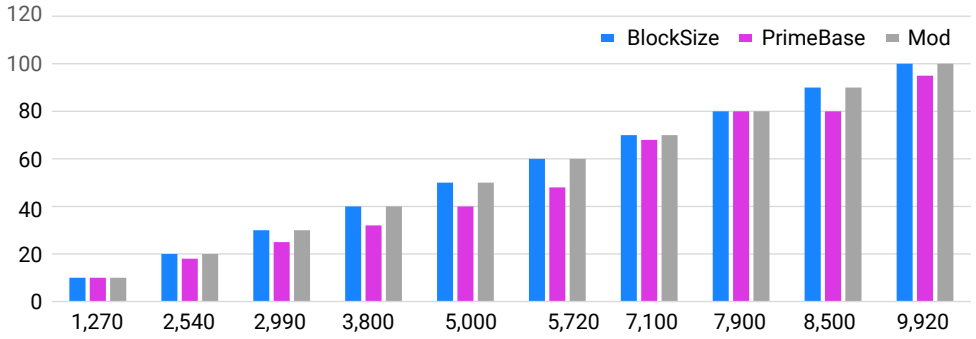

Figure 4. Optimal values of fuzzy hash function parameters

Source: created by the authors

Before the implementation of the research findings, the algorithm exhibited lower performance when monitoring open data compliance. To substantiate the claim of improved performance following the adaptation of the BlockSize, PrimeBase, and Mod parameters, Table 5 presents data from before the parameter adaptation. In this baseline scenario, the default values were set to the smallest values: BlockSize = 1, PrimeBase = 1, and Mod = 1,000,000.

Table 5. Experimental results. Optimal parameters for adapting hash function parameters

No.	Message Length	BlockSize	PrimeBase	Mod	Precision	Recall	F1 Score	Time (ms)
1	1,270	1	1	1,000,000	0.60	0.62	0.61	0.22
2	2,540	1	1	1,000,000	0.65	0.68	0.66	0.24
3	2,990	1	1	1,000,000	0.70	0.73	0.71	0.25
4	3,800	1	1	1,000,000	0.72	0.75	0.73	0.27
5	5,000	1	1	1,000,000	0.75	0.77	0.76	0.30
6	5,720	1	1	1,000,000	0.78	0.80	0.79	0.32
7	7,100	1	1	1,000,000	0.70	0.72	0.71	0.34
8	7,900	1	1	1,000,000	0.72	0.74	0.73	0.36
9	8,500	1	1	1,000,000	0.74	0.76	0.75	0.38
10	9,920	1	1	1,000,000	0.76	0.78	0.77	0.40

Source: created by the authors

Before parameter adaptation (Table 5), the BlockSize, PrimeBase, and Mod parameters were set to 1, resulting in low algorithm efficiency. This was reflected in Precision and Recall values ranging from 0.60 to 0.76, indicating limited accuracy and sensitivity in data processing. Additionally, there was a high probability of false results when detecting similar fragments. After the adaptation (Table 4), increasing the BlockSize, PrimeBase, and Mod parameters led to significant performance improvements. Precision increased to 0.85-0.96, demonstrating a substantial enhancement in the algorithm's ability to accurately detect similar text fragments. The Recall and F1 Score values also improved, reflecting greater sensitivity and overall algorithm efficiency. Furthermore, data processing time decreased, indicating enhanced algorithm performance.

The new parameter configurations enabled higher efficiency in processing data from government websites, significantly improving the monitoring of compliance with open data regulations. The reduction in false positives when detecting duplicates and incorrect data further increased the reliability of the system. Thanks to parameter optimisation, the algorithm became more precise in duplicate detection, reinforcing the effectiveness of adapted algorithm parameters and their critical impact on data processing quality.

Similar studies have been conducted by other researchers. In the work of T. Baba *et al*. (2022), hash function characteristics were utilised to detect malware using deep learning methods. Their study demonstrated high Precision and Recall in malware classification; however, the use of deep learning for such tasks requires substantial computational resources. In contrast, the research presented in this paper focused on adapting the parameters of the fuzzy hash function, achieving comparable accuracy while maintaining lower algorithm complexity and reducing data processing time. Consequently, the proposed approach proved

to be highly efficient for open data monitoring, where processing speed is a critical factor.

The work of A.P. Namanya *et al.* (2020) explored the use of hash functions for detecting malicious files in the Internet of Things (IoT) environment. Their study demonstrated the effectiveness of such methods for small datasets, such as executable files. By comparison, the open data monitoring research conducted in this study required parameter adaptation to handle larger volumes of information. In this context, optimising fuzzy hash function parameters enabled the achievement of high accuracy and sensitivity while significantly reducing data processing time.

S.R. Davies *et al.* (2021) conducted a review of existing ransomware detection methods, including those that utilise fuzzy hash functions to detect modifications in encrypted files. Their study focused on using fuzzy hashing to identify changes in files, enabling the rapid detection of malware, even when only partial modifications or selective variations are applied. However, their research was specifically centred on malware detection, rather than open data monitoring. In comparison to the study presented in this paper, the results of S.R. Davies *et al*. focused on a narrower application, specifically the analysis of small and specific datasets, such as ransomware-infected files. While their approach demonstrated strong performance in ransomware detection, its emphasis on local changes in data makes it less effective for processing large volumes of open data. By contrast, the adaptation of BlockSize, PrimeBase, and Mod parameters in this study enabled higher accuracy and sensitivity in big data processing, which is critical for detecting anomalies on government websites. The increase in Precision, Recall, and F1 Score, alongside the reduction in processing time for large datasets, represents a significant advancement. This distinguishes the present study from the work of S.R. Davies *et al.* (2021), which primarily focuses on individual file-level detection, whereas the proposed approach addresses the broader challenge of large-scale open data monitoring.

M. Eleks *et al.* (2022) explored the use of similarity-preserving hash function algorithms for data anonymisation to facilitate machine learning applications under heightened confidentiality requirements. Their study proposed three novel algorithms based on similarity preservation and evaluated them in terms of accuracy and performance. The primary focus was on enabling machine learning on anonymised data, representing a significant step in expanding the use of confidential datasets across various domains. Their results demonstrated the effectiveness of the proposed algorithms in data anonymisation; however, the emphasis was on maintaining data structure and similarity for subsequent use in machine learning tasks. By contrast, the approach proposed in this study focuses on adapting the BlockSize, PrimeBase, and Mod parameters to ensure high accuracy, sensitivity, and efficiency in processing large-scale open data. Unlike M. Eleks *et al*., whose research assessed algorithms in terms of similarity preservation within a confidentiality framework, the adaptation of fuzzy hash function parameters in this study achieves high Precision, Recall, and F1 Score, which is crucial for detecting anomalies and duplicate records on government websites. Additionally, authors focused on developing algorithms for creating anonymised datasets that could be used to train machine learning models. In contrast, the present study is concerned with ensuring compliance with legal requirements, which necessitates not only preserving data structure but also achieving maximum accuracy in data analysis. In this context, fuzzy hash function parameter adaptation significantly reduces data processing time, distinguishing it from the work of M. Eleks *et al.* (2022), where processing speed was not the primary evaluation criterion.

K.V. Kumar *et al*. (2022) proposed an anonymous human activity recognition method that integrates deep neural networks with fuzzy hash algorithms. Their primary focus was on achieving high accuracy in real-time human action recognition, while employing the Recursive Genetic Micro-Aggregation Approach (RGMAA) model to enhance privacy protection. The Hybrid Deep Fuzzy Hashing Algorithm (HDFHA) used in their study enabled the detection of dependencies between different actions, improving overall accuracy. While their approach demonstrated high accuracy in processing dynamic video data, it requires significant computational resources due to the incorporation of the Deep Belief Network (DBN) and RGMAA. Although their methodology is effective for video data analysis, its computational complexity makes it less suitable for real-time text data monitoring. Both approaches leverage fuzzy hash functions but address different challenges. K.V. Kumar *et al*. focused on action recognition while ensuring data confidentiality, whereas the present study optimises fuzzy hash function parameters for text data analysis. The proposed optimisation demonstrated key advantages in terms of speed and accuracy, making it well-suited for monitoring open data on government websites.

T.-Z. Li *et al.* (2019) proposed a fuzzy hash function algorithm with adaptive file fragmentation, adjusting fragment size based on file dimensions. This enhancement improved the accuracy and efficiency of analysing both small and large files, making the approach particularly valuable in computer forensics, where precise similarity detection between files is crucial. The research presented in this paper, however, focused on evaluating the impact of modifying the BlockSize, PrimeBase, and Mod parameters on the efficiency of the fuzzy hash function, aiming to determine optimal values for texts of varying lengths. This approach considers the specific characteristics of open data on government websites, where text volume and structure can vary significantly. Identifying optimal parameters for different text lengths helps reduce collisions, increase accuracy, and enhance the sensitivity of the algorithm. In contrast to T.-Z. Li *et al.,* who optimised fragmentation rules for different file types, this study focused on large-scale text data, requiring a flexible approach to fuzzy hash function parameter configuration. Despite their differing focuses, both approaches contribute to the advancement of fuzzy

hashing methods, demonstrating effectiveness in their respective applications.

J. Chen *et al*. (2014) introduced a method for clustering spam campaigns using fuzzy hash functions. Their study aimed to detect spam botnets by grouping emails with similar content within a single spam campaign. The use of fuzzy hash functions enabled the successful processing of spam messages, even in the presence of obfuscation techniques such as URL shortening. Their research demonstrated the effectiveness of the proposed method on a three-year dataset comprising 540,000 spam emails, revealing typical behavioural patterns of botnets. In contrast, the approach proposed in this paper focuses on text data analysis, where adapting the fuzzy hash function parameters enhances analytical efficiency. While J. Chen *et al*. developed a powerful tool for spam campaign detection, the parameter adaptation in this study is critical for monitoring large-scale open data, where accuracy and processing speed are key factors.

Thus, the results of this study confirm that optimising the BlockSize, PrimeBase, and Mod parameters is an effective approach for enhancing accuracy, sensitivity, and processing speed in open data monitoring. The proposed solution offers a significant advantage over other approaches that rely on complex models or focus on different types of data. The implementation of adapted parameters enables superior performance, particularly in the context of government web resources, where ensuring high processing speed alongside high accuracy is essential.

Conclusions

This article examined methods for analysing the parameters of a fuzzy hash function to enhance the efficiency of monitoring compliance with the Regulation on Open Data on official government websites. The primary objective of the study was to determine the optimal values for BlockSize, PrimeBase, and Mod to ensure high accuracy and sensitivity in text processing, thereby improving open data monitoring. The study's findings were integrated into a software tool, leading to a significant improvement in accuracy and a reduction in data processing time. This was achieved through the dynamic adjustment of the block size and prime number base, depending on the length of the processed message.

The study also analysed how different values of BlockSize, PrimeBase, and Mod affected key algorithm performance metrics: Precision, Recall, F1 Score, and Processing Time. As a result, optimal parameter values were identified for different message lengths. For example, at a Message Length of 1,270, Precision was 0.85, increasing to 0.96 for a Message Length of 9,920, demonstrating enhanced accuracy in detecting similar data. The implementation of these optimised parameters in the software tool significantly improved the accuracy of detecting incorrect or duplicate data, which is crucial for automating compliance monitoring in accordance with open data legislation. Overall, the study demonstrated that fuzzy hash function parameter optimisation enhances data management quality and increases the efficiency of automated monitoring systems. The use of optimal parameters reduces the likelihood of errors in duplicate data detection while ensuring faster data processing.

Future research will focus on refining algorithms to achieve more precise parameter selection, tailored to the specific characteristics of the processed data. A particularly promising avenue is the adaptation of the algorithm for detecting duplicate media files (audio, video, and images) with appropriate pre-processing techniques. Advancements in this area have the potential not only to improve monitoring accuracy and completeness but also to reduce time and resource costs in data compliance verification. The development of enhanced methods for adapting fuzzy hash function parameters could significantly streamline data management and improve data quality, contributing to the more effective and reliable implementation of open data regulations.

Acknowledgements

None.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] AlMajali, A., Elmosalamy, A., Safwat, O., & Abouelela, H. (2024). Adaptive ransomware detection using similarity-preserving hashing. *Applied Sciences*, 14(20), article number 9548. doi: 10.3390/app14209548.

[2] Baba, T., Baba, K., & Yamauchi, T. (2022). Malware classification by deep learning using characteristics of hash functions. In: L. Barolli, F. Hussain & T. Enokido, (Eds.), *Advanced information networking and applications* (Vol. 450, pp. 480-491). Cham: Springer. doi: 10.1007/978-3-030-99587-4_40.

[3] Chanajitt, R., Pfahringer, B., Gomes, H.M., & Yogarajan, V. (2022). Multiclass malware classification using either static opcodes or dynamic API calls. In: H. Aziz, D. Corrêa & T. French (Eds.), *AI 2022: Advances in artificial intelligence* (Vol. 13728, pp 427-441). Springer, Cham. doi: 10.1007/978-3-031-22695-3_30.

[4] Chen, J., Fontugne, R., Kato, A., & Fukuda, K. (2014). Clustering spam campaigns with fuzzy hashing. In *Proceedings of the 10th Asian internet engineering conference* (pp. 66-73). New York: ACM. doi: 10.1145/2684793.2684803.

[5] Davies, S.R., Macfarlane, R., & Buchanan, W.J. (2021). Review of current ransomware detection techniques. In *Proceeding of the 7 th international conference on engineering and emerging technologies (ICEET)* (pp. 696-701). Istanbul: IEEE. doi: 10.1109/ICEET53442.2021.9659643.
[6] Eleks, M., Rebstadt, J., Fukas, P., & Thomas, O. (2022). Learning without looking: Similarity preserving hashing and its potential for machine learning in privacy critical domains. In *INFORMATIK 2022*, *lecture notes in informatics (LNI)* (pp.161-177). Bonn: IBiS. doi: 10.18420/inf2022_16.
[7] Fleming, M., & Olukoya, O. (2024). A temporal analysis and evaluation of fuzzy hashing algorithms for Android malware analysis. *Forensic Science International: Digital Investigation*, 49, article number 301770. doi: 10.1016/j.fsidi.2024.301770.
[8] Guerrero, M. (2022). Comparative study between Type-1 and interval Type-2 fuzzy systems in parameter adaptation for the Cuckoo search algorithm. *Symmetry*, 14(11), article number 2289. doi: 10.3390/sym14112289.
[9] Kida, M., & Olukoya, O. (2023). Nation-state threat actor attribution using fuzzy hashing. *IEEE Access*, 11, 1148-1165. doi: 10.1109/ACCESS.2022.3233403.
[10] Kondratenko, N.R. (2023). Interval type-2 generalizing fuzzy model for monitoring the states of complex systems using expert knowledge. *System Research and Information Technologies*, 2. doi: 10.20535/SRIT.2308-8893.2023.2.05.
[11] Kondratenko, N.R., & Snihur O.O. (2019). Research on the adequacy of interval type-2 fuzzy models in identifying complex objects. *System Research and Information Technologies*, 4, 94-104.
[12] Kumar, K.V., Harikiran, J., & Chandana, B.S. (2022). Human activity recognition with privacy preserving using deep learning algorithms. In *2nd international conference on artificial intelligence and signal processing (AISP)* (pp. 1-8). Vijayawada: IEEE. doi: 10.1109/AISP53593.2022.9760596.
[13] Li, T.-Z., Shen, B., Mi, K., Kao, Y.-C., & Cui, Y. (2019). A method of piecewise hash for fuzzy hashing. *Journal of Computers*, 30(2), 150-157. doi:10.3966/199115992019043002013.
[14] Mahrous, W.A., Farouk, M., & Darwish, S.M. (2021). An enhanced blockchain-based IoT digital forensics architecture using fuzzy hash. *IEEE Access*, 9, 151327-151336. doi: 10.1109/ACCESS.2021.3126715.
[15] Martín-Pérez, M., Rodríguez, R.J., & Breitinger, F. (2021). Bringing order to approximate matching: Classification and attacks on similarity digest algorithms. *Forensic Science International: Digital Investigation*, 36, article number 301120. doi: 10.1016/j.fsidi.2021.301120.
[16] Ministry of Digital Transformation of Ukraine. (n.d.). Retrieved from https://thedigital.gov.ua/.
[17] Naik, N., Jenkins, P., & Savage, N. (2019b). A ransomware detection method using fuzzy hashing for mitigating the risk of occlusion of information systems. *IEEE international symposium on systems engineering. (ISSE)* (pp. 1-6). Edinburgh: IEEE. doi:10.1109/ISSE46696.2019.8984540.
[18] Naik, N., Jenkins, P., Gillett, J., Mouratidis, H., Naik, K., & Song, J. (2019a). Lockout-Tagout Ransomware: A detection method for Ransomware using fuzzy hashing and clustering. *IEEE symposium series on computational intelligence (SSCI)* (pp. 641-648). Xiamen: IEEE. doi: 10.1109/SSCI44817.2019.9003148.
[19] Namanya, A.P, Awan, I.U., Disso, J.P., & Younas, M. (2020). Similarity hash based scoring of portable executable files for efficient malware detection in IoT. *Future Generation Computer Systems*, 110, 824-832. doi: 10.1016/j.future.2019.04.044.
[20] Nandal, A., Blagojevic, M., Milosevic, D., Dhaka, A., & Mishra, L.N. (2021). Fuzzy enhancement and deep hash layer based neural network to detect Covid-19. *Journal of Intelligent & Fuzzy Systems*, 41(1), pp. 1341-1351. doi: 10.3233/JIFS-210222.
[21] Natella, R. (2022). StateAFL: Greybox fuzzing for stateful network servers. *Empirical Software Engineering*, 27, article number 191. doi: 10.1007/s10664-022-10233-3.
[22] National Bank of Ukraine. (n.d.). Retrieved from https://bank.gov.ua.
[23] Open Data Portal. (n.d.). Retrieved from https://data.gov.ua.
[24] Pension Fund of Ukraine. (n.d.). Retrieved from https://www.pfu.gov.ua.
[25] Resolution of the Cabinet of Ministers of Ukraine No. 835 “On Approval of the Regulation on Data Sets Subject to Disclosure in the Form of Open Data”. (2015, October). Retrieved from https://zakon.rada.gov.ua/laws/show/835-2015-%D0%BF#Text.
[26] Ssdeep-project. (n.d.). *Fuzzy hashing API*. Retrieved from https://github.com/ssdeep-project/ssdeep.
[27] State Service of Special Communications and Information Protection of Ukraine. (n.d.). Retrieved from https://cip.gov.ua.
[28] State Statistics Service of Ukraine. (n.d.). Retrieved from https://ukrstat.gov.ua.
[29] State Tax Service of Ukraine. (n.d.). Retrieved from https://tax.gov.ua.
[30] Verkhovna Rada of Ukraine. Official Web Portal of the Parliament of Ukraine. (n.d.). Retrieved from https://www.rada.gov.ua/.

Дослідження параметрів нечіткої геш-функції для моніторингу дотримання вимог положення щодо відкритих даних

Леонід Майданевич

Кандидат філософських наук, старший викладач
Вінницький національний технічний університет
21021, вул. Хмельницьке шосе, 95, м. Вінниця, Україна
https://orcid.org/0000-0002-7364-8874

Наталія Кондратенко

Кандидат технічних наук, професор
Вінницький національний технічний університет
21021, вул. Хмельницьке шосе, 95, м. Вінниця, Україна
https://orcid.org/0000-0002-4450-1603

Віталій Казміревський

Аспірант
Вінницький національний технічний університет
21021, вул. Хмельницьке шосе, 95, м. Вінниця, Україна
https://orcid.org/0009-0005-4056-5385

Анотація. Метою роботи було дослідження параметрів геш-функції для підвищення ефективності та точності виявлення подібності текстових фрагментів на різних веб-ресурсах при проведенні моніторингу дотримання вимог Положення щодо відкритих даних на офіційних веб-сайтах державних органів. Дослідження охопило оцінку трьох ключових параметрів геш-функції: розміру блоку, бази простого числа та модуля. Для цього було проведено серію експериментів, у яких різні комбінації цих параметрів використовувалися для генерування геш-значень текстових даних. Результати дослідження продемонстрували, які комбінації параметрів забезпечують найкращий баланс між точністю, повнотою, F-мірою та часом виконання. Показано, що певні комбінації параметрів дозволяють досягти значного підвищення точності алгоритму при мінімізації обчислювальних витрат, що є важливим для аналізу даних у реальному часі. Встановлено, що оптимізація параметрів геш-функції сприяє зниженню кількості хибнопозитивних та хибнонегативних результатів, які часто виникають при виявленні подібності. Зокрема, підбір оптимальних значень для кожного з параметрів суттєво підвищує точність і повноту аналізу, дозволяючи отримати більш точні результати порівняння текстових фрагментів та зменшуючи час виконання операцій. Це робить алгоритм нечіткого гешування придатним для застосування в автоматизованих системах моніторингу державних веб-сайтів щодо дотримання вимог щодо відкритих даних. Виявлено, що оптимізація параметрів дозволяє зменшити кількість дубльованих записів, що особливо актуально для забезпечення відповідності відкритих даних вимогам законодавства. Одержані висновки можуть бути використані для розробки програмних засобів, які допоможуть ефективно виявляти недоліки та сприятимуть підвищенню прозорості та відповідності правовим вимогам. Крім того, результати дослідження можуть бути використані для подальшої оптимізації алгоритмів нечіткої геш-функції, що сприятиме вдосконаленню технологій моніторингу даних на відповідність нормативним вимогам. Дослідження робить внесок у розвиток технологій моніторингу веб-ресурсів, демонструючи, як правильно підібрані параметри нечіткої геш-функції можуть значно підвищити ефективність і надійність аналізу відкритих даних

Ключові слова: параметри нечіткої геш-функції; моніторинг веб-сайтів; державні електронні ресурси; точність алгоритму; параметри оптимізації; виявлення подібності; порушення положень

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

Information Technologies and Computer Engineering **Vol. 21, No. 3. 2024**

Article's History: Received: 30.08.2024 Revised: 12.11.2024 Accepted: 26.12.2024

UDC 621.396.96:534.23 10.63341/vitce/3.2024.77

Correction of roll-caused stripe noise in side scan sonar images

Oleksandr Katrusha*

Postgraduate Student
National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute"
03056, 37 Beresteiskyi Ave., Kyiv, Ukraine
https://orcid.org/0009-0008-7097-4843

Abstract. Ensuring high-quality images obtained using side-scan sonar is crucial for enhancing the effectiveness of underwater research, as distortions such as striping noise can complicate data analysis. The aim of this paper was to investigate the nature of striping noise, determine the correlation between image intensity and the tilt of the sonar, and develop a new method to improve the quality of sonar images. The study employed a statistical correction method based on calculating a horizontal moving average for intensity correction, as well as a machine learning model using a three-layer neural network to predict the horizontal moving average considering the beam's incidence angle, the sonar's height above the seafloor, and the initial line intensity. Statistical methods and machine learning techniques were applied to correct the striping noise caused by tilting in sonar images, significantly enhancing their quality. The statistical approach, which uses the mean value of the horizontal sway, effectively reduced noise while preserving critical details and improving overall clarity. The machine learning model incorporated additional parameters, enhancing intensity prediction accuracy and improving adaptability to various sonar positioning conditions. Moreover, the new method accounts for varying environmental conditions, making it flexible and effective for real-world underwater research. These results provide valuable insights for improving sonar image processing methods, paving the way for more efficient underwater exploration and improving the accuracy of object detection on the seafloor

Keywords: seafloor; greyscale; intensity; roll; correction; machine learning; neural network

Introduction

Side scan sonar, a technology widely utilised since the 1950s, plays a crucial role in underwater exploration, supporting various activities such as search and rescue missions, bathymetry, and mine detection. Despite its effectiveness, side scan sonar images are often affected by distortions due to the complex underwater environment and sonar's movement. Among these distortions, stripe noise arising from roll variations compromises image clarity, particularly affecting object recognition and seafloor segmentation. This type of noise manifests as alternating dark and bright stripes across the image, complicating both manual and automated interpretation of sonar data. Addressing stripe noise is a complex task, as existing intensity normalisation methods often fall short in handling roll-induced artefacts. Roll-caused stripe noise is unpredictable and variable in length, width, and frequency due to environmental factors like wave patterns and seafloor profile, which are difficult to standardise in real-time. Therefore, there is a need for targeted correction methods that specifically address roll-induced stripe noise, enhancing the utility of sonar imagery in underwater research and exploration.

Z. Lu *et al*. (2023) investigated a method for enhancing side-scan images based on multistage image restoration and fusion. The aim of the study was to improve image clarity by suppressing noise and correcting uneven illumination. The results showed that the proposed method significantly improves the image quality, which contributes to a better analysis of the seabed. P. Zhou *et al*. (2024) presented a multiscale fusion strategy for correcting side-scan images to improve contrast and reduce the impact of noise. The aim of the study was to improve the accuracy of object detection and eliminate noise in the images. The results proved the effectiveness of the approach for improving the contrast and clarity of details in sonograms.

H. Xia *et al*. (2024) investigated an improved method for removing banding noise based on the Criminisi

Suggested Citation:
Katrusha, O. (2024). Correction of roll-caused stripe noise in side scan sonar images. *Information Technologies and Computer Engineering*, 21(3), 77-85. doi: 10.63341/vitce/3.2024.77

*Corresponding author

algorithm. The aim was to adapt the method for specific noise features in side-scan images. The results showed that the improved method can better cope with multidirectional noise while maintaining the clarity of important structures. The article M. Li *et al*. (2022) presented an approach to removing banding noise in infrared images, based on the sparsity of gradients along the band direction and the global sparsity of noise. An Adaptive Edge-Preserving Operator (AEPO) was proposed, which safeguards edge details in the image while minimising information loss.

J. Guan *et al*. (2019) proposed an innovative wavelet-based deep neural network that effectively removes banding noise in infrared images, taking into account its intrinsic characteristics and the interconnections between wavelet sub-bands. A directional regulariser was added to enhance the separation of scene details from noise, ensuring more accurate image restoration. In the study by S. Shaobo *et al*. (2022), a novel radiometric correction method for side-scan sonar (SSS) images was proposed, incorporating prior knowledge of acoustic illumination and seabed characteristics. The method is based on the decomposition of illumination and albedo components using low-rank constraints and anisotropic total variation (ATV). Experimental results demonstrated the effectiveness of the proposed approach in correcting radiometric distortions and reducing residual noise. It must be noted that little attention was paid in the scientific literature to roll-caused strip noise of sonar images with its peculiarities – sporadicity, variations in length, width and frequency caused by sonar beam pattern and direction of vehicle in relation to waves, strong dependence on floor profile that is typically unknown when using side scan sonar. These characteristics of roll-caused stripe noise make the abovementioned methods hardly applicable to roll-caused stripe noise of side scan sonar images. The purpose of this paper was to investigate the nature of stripe noise, determine the relationship between image intensity and sonar roll angle, and develop a new method for improving the quality of sonar images.

Materials and Methods

This section provides a structured outline of the methods used in this study, including data selection, the analysis of the relationship between image intensity and sonar roll, and the steps in the proposed stripe noise correction methodology. The methodology consisted of two primary approaches: a statistics-based correction method and a machine learning-based model that builds on the results of the initial statistical analysis.

This study analysed more than 300 log files of sonar measurements obtained with the "Sonobot 5" unmanned surface vehicle manufactured by the German company "EvoLogics GmbH". The selection criteria focused on logs with high roll standard deviation and a Pearson's correlation coefficient greater than 0.3 between average along-track image intensity and roll. These criteria ensure that the selected data contain sufficient variability to capture the effects of roll on stripe noise in sonar images.

The roll distribution in the selected missions was close to normal, with a mean of zero. The standard deviation of roll varied across missions, influenced by wave patterns and strength specific to each mission. This distribution allowed for a robust analysis of the relationship between roll and intensity distortion in sonar images (Fig. 1).

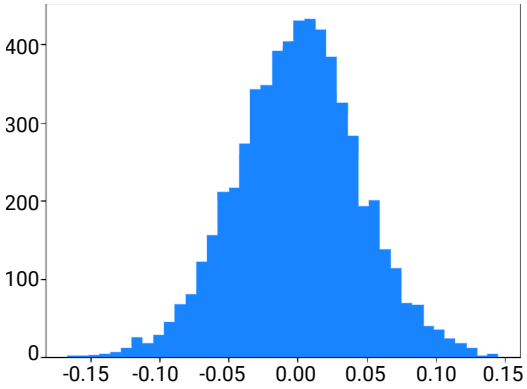

Figure 1. Typical distribution of vehicle roll in Rad
Notes: X-axis is roll angle in Rad, Y-axis – number of observations
Source: created by the author

A visual inspection of the sonar images revealed a clear relationship between roll and stripe noise. As shown in Figure 2, negative roll angles corresponded to darker areas on the left side of the image, likely due to the sonar transmitter's left main lobe ensonifying regions closer to the sonar. Conversely, positive roll angles produced darker areas on the right side of the image, suggesting a mirrored pattern. This pattern, as illustrated in Figure 3, demonstrates the dependency of intensity on roll but also highlights that this relationship is modulated by additional factors such as sonar altitude and seafloor topography. Consequently, a straightforward intensity correction approach was deemed insufficient, necessitating a more comprehensive correction method.

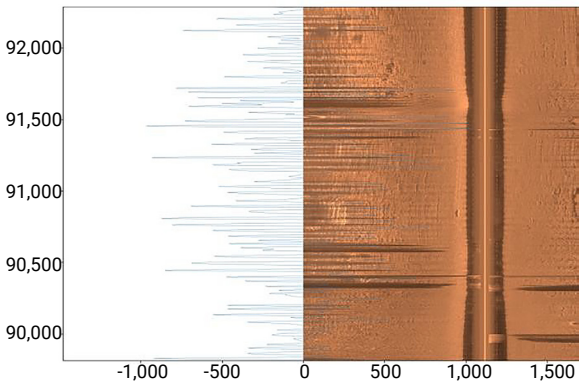

Figure 2. A fragment of a striped sonar image with corresponding roll in Rad ×10,000 (blue line). Sonar vehicle movement is bottom-upd
Notes: X-axis is roll angle in Rad ×10,000, Y-axis – ping number
Source: created by the author

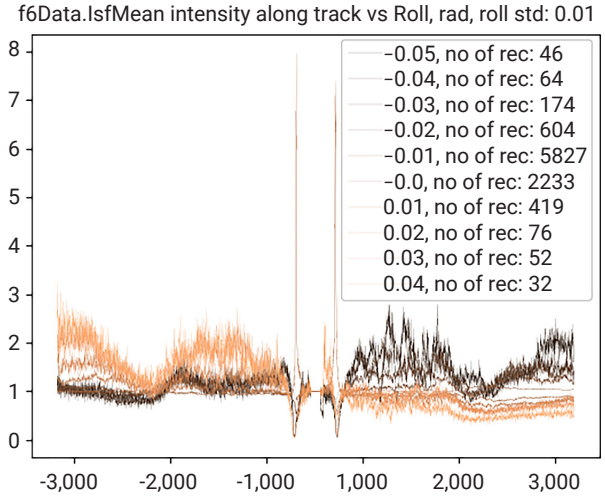

Figure 3. Average intensities of image parts depending on roll vs average intensity of the image
Notes: the X-axis is the slant range (negative on port side), and the Y-axis is the ratio between average intensity for roll angle and average along-tack intensity of the whole image. The number of records for each rounded roll angle is displayed in the legend. Roll angles with fewer than 30 records are considered outliers and are not displayed
Source: created by the author

The proposed methods included a statistics-based correction method and a machine learning-based model. The statistics-based approach involves a statistical analysis aimed at establishing a baseline correction by calculating a horizontal rolling mean, which adjusts intensity variations that correlate with roll angle. The machine learning-based model, building on the initial statistical analysis, utilises a three-layer dense neural network to predict the horizontal rolling mean for sonar images. This model takes roll, altitude, and original line intensity as inputs and outputs a corrected horizontal mean, simulating an image with zero roll influence.

A dataset of 40,000 samples was derived from ten representative sonar images, where the inputs consist of roll, altitude, and line intensity, and the output is the horizontal rolling mean. The model was trained using TensorFlow with the ADAM optimiser, over 800 epochs, with a batch size of 256. The training was conducted on hardware that included a GeForce RTX4070 graphics processing unit and an Intel i-Core 7 central processing unit, taking approximately 40 minutes, while inference on a 2,200×20,000 pixel image required about 300 milliseconds. Formula (1) describes the training stage, where d is altitude, φ is roll, *Lo* – line intensity of the original image and *Mh* – line horizontal mean:

$$d, \phi, Lo = > Mh. \quad (1)$$

The formula is used to generate predictions for the horizontal average line intensity based on the specified input parameters. This allows to adjust the signal intensity in images, reducing distortion caused by roll variations and altitude changes.

Results and Discussion

Sonar technology operates by emitting a series of sound impulses that travel through water and reflect off underwater objects and the seafloor. The intensity of these reflected signals is then recorded, allowing for the construction of sonar images, commonly referred to as sonograms. Each sonogram is created by sequentially adding lines of received signal intensity, resulting in a comprehensive visual representation of the underwater environment.

The process of sonar imaging is inherently influenced by various factors that can lead to geometrical and intensity distortions. The complex physical laws governing sound reflection and propagation in water play a crucial role in how sonar signals behave. For example, the angle of incidence, the type of materials on the seafloor, and the acoustic properties of the water can all affect how sound waves are reflected back to the sonar device. The movement of the sonar vehicle, including its speed and direction, can introduce further variability into the recorded data (Ye *et al.*, 2019). Underwater currents also contribute to the challenges faced during sonar imaging. They can alter the trajectory of the sound waves, leading to inconsistencies in the intensity of the returned signals. This variability can result in images that are less readable, making it difficult to accurately detect and classify objects on the seafloor. Both human operators and machine learning algorithms face significant hurdles when attempting to interpret these distorted images, as the noise and artifacts can obscure critical features of the underwater landscape (Capus *et al.*, 2008; Al-Rawi *et al.*, 2017).

Variations by roll do not affect image geometry directly but can cause intensity distortion caused by uneven ensonification of different seafloor areas due to non-uniformity of sonar beam pattern. Figure 4 illustrates the convention of Euler angles that define roll (φ) among other angles. J.E. Hughes Clarke (2004) provide a sample empirically derived beam pattern of a side scan sonar (Fig. 5).

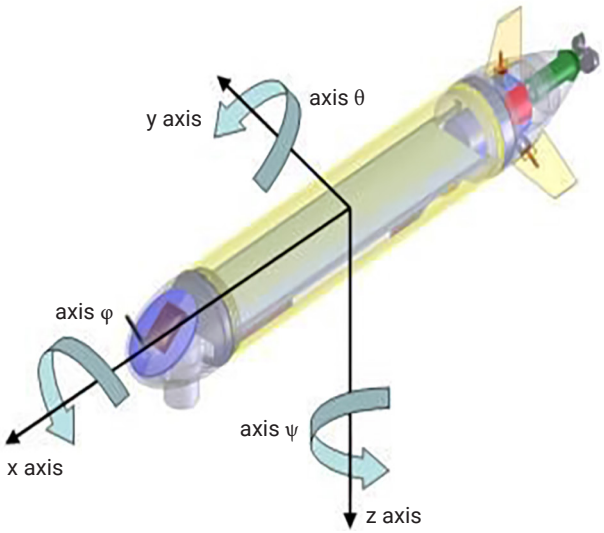

Figure 4. The body-fixed reference frame and Euler angles
Source: Navigation messages (n.d.)

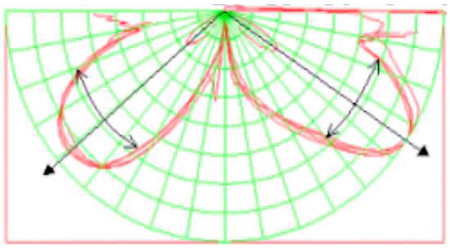

Figure 5. Empirically derived beam pseudo-pattern of a side scan sonar
Source: J.E. Hughes Clarke (2004)

The roll-caused intensity distortion, commonly referred to as stripe noise, can be observed as a series of alternating dark and bright stripes across the image, each with varying length and width and a distinct periodic pattern. These stripes tend to widen toward the edges of the image and exhibit asymmetry in brightness, with darker stripes on one side corresponding to brighter ones on the opposite side. In regions of acoustic shadow, stripe noise is typically less visible or absent altogether. An example of roll-caused intensity distortion also known as stripe noise can be seen in Figure 6. It appears as sporadic series of alternating dark and bright across track stripes of various length and width with distinct periodic pattern.

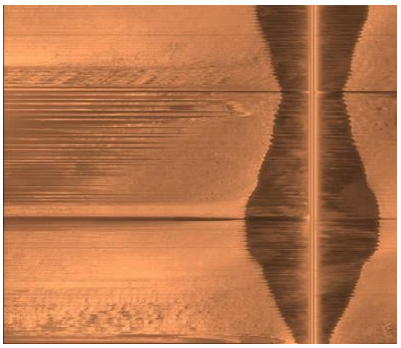

Figure 6. An example of stripe noise on a sonogram. Sonar vehicle movement bottom-up
Source: created by the author

The main cause of stripe noise in the provided examples is the influence of wave motion on the keel-mounted sonar, which induces variations in roll that affect seafloor ensonification due to non-uniformity of the beam pattern and the resulting image intensity. In general terms, the objective of roll-caused stripe noise correction is to restore the intensity profile of each scanned line as if it were captured at a zero roll angle. However, in certain extreme cases, valuable seafloor information can be entirely obscured by overly dark or bright sections, making full restoration impractical.

Correction based on rolling mean

To address this, a statistical-based approach to stripe noise correction was initially implemented. For each horizontal line in the image, a horizontal rolling mean was calculated to help identify the underlying intensity trend across the track while minimising random noise. The window size for the rolling mean was empirically determined to balance smoothness and data retention, ensuring that subtle seafloor features were not lost. Subsequently, this horizontal mean was averaged along-track to generate a neighbourhood mean, providing a contextual baseline for each line. The window size for this averaging process corresponds to the maximum periodicity of the stripe noise, approximately 20 lines or two meters on the seafloor in this dataset. This period closely matches the average period of wave-induced roll variations, making it a suitable reference for noise correction.

Rolling mean was chosen over interpolation and other smoothing techniques due to its computational efficiency and ease of implementation. Empirical analysis confirmed that this approach effectively reduces the noise without compromising the critical details of the image.

Figure 7 shows the process of image correction using the rolling mean method. Figure 7(a) shows the original image, which contains striped noise caused by variations in the roll of the sonar instrument. Figure 7(b) shows the correction map obtained by dividing the horizontal average by the average of the neighbouring lines. In the correction map, red indicates higher values and blue indicates lower values. This map fits the stripe noise well, although it does not directly account for roll variation. Figure 7(c) shows the result of the image correction, where the correction map is applied to the original image.

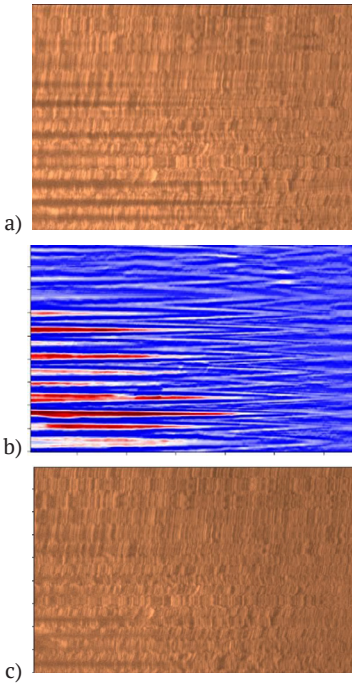

Figure 7. The process of stripe noise correction in sonar images using the moving average method
Notes: a) a piece of original image with stripe noise; b) correction map based on rolling mean (red colour corresponds to higher values, blue colour – to lower ones); c) correction result
Source: created by the author

Through the process of image correction, the stripe noise is reduced and the image becomes clearer. A correction map is obtained by dividing the horizontal mean by neighbourhood mean (Fig. 7(b)). As it can be seen from the figure, the correction map corresponds well to stripe noise, although does not take roll variations into account directly. Then correction procedure for every horizontal line can be implemented as in Formula (2):

$$L_c = L_o - M_h + M_h \times C_i, \quad (2)$$

where L_c – corrected horizontal line of the image; L_o – original line; M_h – calculated horizontal mean for the line; C_i – corresponding line of the correction map.

The rolling mean method is simple, computationally effective, and easy in implementation, but it is based solely on image information. Whereas intensity of the image pixel depends, among others, on the following parameters: slant range, floor profile, floor sediment type, incidence angle, roll, sonar beam pattern and sonar altitude (Al-Rawi, 2016; Chang *et al.*, 2020). Failure to take account for these factors can lead to overcompensation in some areas, loss of acoustic shadows and periodic seafloor patterns, wrong compensation of the areas with high floor gradient.

Correction using neural network inference

To enhance the correction process by incorporating additional parameters, such as roll and altitude, a machine learning-based method is introduced. Machine learning techniques are increasingly applied to sonar image processing, including tasks such as object detection, feature extraction, classification, and seafloor segmentation (Chen *et al.*, 2017; Li *et al.*, 2024). However, due to the relatively high cost and limited availability of sonar data, these methods lag behind traditional optical computer vision techniques. Most current research focuses on improving object depiction and seafloor clarity, yet few studies have specifically targeted stripe noise correction (Steiniger *et al.*, 2022; Sivachandra & Kumudham, 2024).

The proposed machine learning model aims to predict the intensity dependence by approximating the rolling mean of image intensity based on variables including roll, slant range, altitude, and floor profile. This approach avoids the need for complex mathematical models, such as Lambertian reflections, by training the model to infer these relationships from available data. This setup is especially useful in sonar applications where obtaining ground-truth images is costly and often infeasible.

The limited availability and high cost of obtaining ground truth data present significant challenges for most machine learning approaches in sonar applications. Consequently, directly training models using pairs of distorted and corrected images is often impossible. These algorithms possess the capability to approximate the dependencies of intensity based on various parameters, which can be leveraged to address this issue effectively. The proposed method involves using a neural network to model the complex relationship between the intensity rolling mean and variables such as roll, slant range, altitude, and floor profile, without relying on intricate mathematical models like the Lambertian law.

Due to the limited availability and high cost of obtaining ground truth data, most machine learning approaches in sonar applications face significant challenges. This approach allows for more flexible modelling of intensity dependencies, accommodating the unique characteristics of sonar data.

To correct stripe noise, the subsequent inference stage (3), where *Mc* is the inferred horizontal mean, predicts the horizontal rolling mean for the lines with the same inputs, except for roll, which is set to zero radians. This formula is derived from (1), providing a baseline adjustment in the absence of roll effects (3):

$$d,\ 0,\ Lo = > Mc. \quad (3)$$

Then, based on this correction, the following formula is applied (4):

$$I_c = I_o - M_h + M_c, \quad (4)$$

where I_c – corrected image; I_o – original image; M_h – horizontal mean; M_c – correction value.

Correction map is constructed as a difference between the inferred rolling mean and original rolling mean. After that correction map is subtracted from the original image thus correcting the roll-caused variation (4). For the experiment, a simple three-layer dense neural network was used. It took roll, altitude, original line intensity as input and horizontal rolling mean as output. The training set of 40,000 samples was generated from ten sonar images. The inference stage is aimed at predicting the horizontal mean with zero roll.

Figure 8 illustrates the effectiveness of the neural network in capturing the relationship between roll and mean intensity, as evidenced by the corrected images. Panel (a) displays the original image with pronounced stripe noise, while panel (b) presents the correction map generated through inference. The correction map indicates higher values in red and lower values in blue, highlighting areas where the neural network has identified the need for adjustment. The resulting corrected image is shown in panel (c), demonstrating noticeable improvements in clarity.

Despite the success in modelling this dependency, instances of overcompensation and excessive smoothing of the image were observed with the application of the inferred correction map. This phenomenon suggests that while the model effectively learned certain aspects of the intensity relationship, it may not have fully accounted for all complexities inherent in sonar imagery. Such overcompensation can obscure fine details and alter important features of the seabed, potentially complicating subsequent analysis.

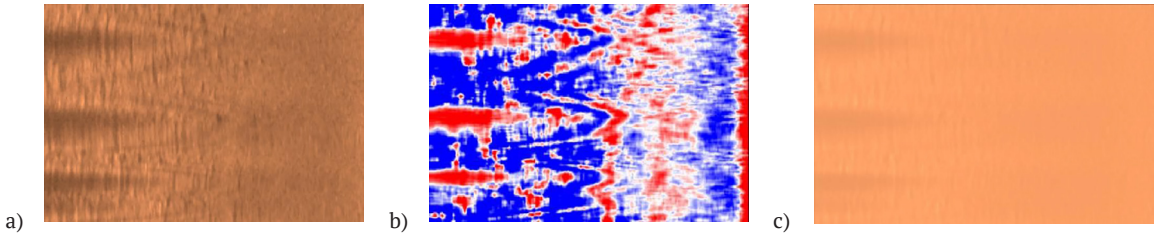

Figure 1. Neural network-based correction of stripe noise in sonar images
Notes: a) original image with stripe noise; b) correction map based on inference (red colour corresponds to higher values, blue colour – to lower ones); c) corrected image
Source: compiled by the author

To address these issues, incorporation of additional features into the training process is proposed. For instance, including derivatives of intensity values, various angles, and their mean values, as well as information from adjacent image lines, could enhance the model's capacity to capture more nuanced relationships. Further experimentation with neural network architecture, including adjustments to layer configurations and activation functions, may also yield better results. These refinements will be the focus of future research, with the aim of improving the model's robustness and accuracy in correcting stripe noise, ultimately leading to enhanced sonar image quality and reliability in underwater exploration.

The discourse surrounding the correction of roll-caused stripe noise

The findings of this study resonate with the studies of other researchers in the field, highlighting a shared commitment to improving sonar image quality through various correction methodologies. For instance, Y.-C. Chang *et al*. (2020) focused on processing side-scan images to correct brightness variation and fill gaps, aiming to eliminate problems associated with uneven illumination during seabed scanning. They developed a technique that improves image quality, resulting in a more homogeneous representation of the seabed. Their results demonstrated that this method effectively reduces luminance distortions and increases the convenience of further image analysis. Previous studies have also investigated various approaches to noise reduction, including advanced filtering techniques and optimisation algorithms specifically designed for sonar data. Research has highlighted the potential of deep learning models to improve object detection and seafloor classification, underscoring the importance of leveraging multiple methodologies to address the challenges posed by sonar imagery. These collective efforts emphasise the significance of a multi-faceted approach to improving sonar data processing, suggesting that continued collaboration and innovation in this field are essential for advancing underwater exploration and analysis.

M.S. Al-Rawi (2016) developed an intensity normalisation method specifically for side-scan sonar images, which targeted the elimination of brightness irregularities arising from changes in the angles of sound wave incidence. Their research demonstrated that these irregularities could significantly impact the clarity of sonar images, making it challenging to identify and analyse underwater features accurately. By implementing their normalisation technique, authors found that distortion was effectively reduced, leading to enhanced object recognition capabilities in sonar images. This study underscored the critical importance of addressing intensity variations to improve the overall utility of sonar technology in underwater applications.

J. Zhao *et al*. (2017) introduced a radiometric correction method that considers the variation of sediment types on the seafloor, which can significantly influence how sonar signals are reflected back to the device. Their study provided a comprehensive model that accounted for the different acoustic properties of various sediment types, thereby enhancing the accuracy and detail of the reflected signal intensity data. By integrating sediment characteristics into their correction methodology, J. Zhao *et al*. were able to achieve more reliable and detailed sonar images, which are essential for effective seafloor mapping and analysis.

The findings of this study align with and build upon the results of several notable researchers in the field of sonar image processing. A. Burguera & G. Oliver (2014) focused on intensity correction to mitigate uneven illumination effects, which ultimately improved image quality and facilitated better data analysis. Similarly, the current research successfully addressed stripe noise through a correction methodology that enhances image clarity, demonstrating a parallel objective of improving the usability of sonar imagery for analysis.

G. Shippey *et al*. (1994) applied a shadow correction method utilising histogram transformations to enhance the contrast and detail in side-scan sonar images. Their approach specifically targeted the reduction of uneven brightness, a common challenge that can obscure essential features and hinder effective seabed analysis. By creating more homogeneous images, their methodology not only improved the visibility of underwater structures but also facilitated more accurate object detection and classification. The results from this study resonate with the current findings, where the application of neural network techniques and rolling mean methods also aims to enhance overall image quality and facilitate clearer object recognition. Both approaches underscored the importance of addressing brightness inconsistencies in sonar imagery to achieve better analytical outcomes.

D. Wilken *et al*. (2012) applied Fourier filtering in two-dimensional space to eliminate stripe noise in mosaics of side-scan images. The goal was to reduce periodic noise and improve the quality of the mosaic images of the seafloor. The results showed that Fourier filtering effectively eliminates bandpass noise and produces more homogeneous and clear images.

Y. Chen *et al*. (2017) developed a method for removing stripe noise in remote sensing images through total variation regularisation and group sparsity constraints. Their research highlighted the effectiveness of this approach in yielding clearer and more detailed images, reinforcing the objective of the present research, which focuses on removing stripe noise and enhancing the quality of sonar data processing.

An analysis of these studies indicated a prevailing trend in the adoption of advanced correction techniques, whether through statistical methods, machine learning, or mathematical modelling, to improve the quality of sonar images. The integration of these methodologies reinforces the notion that a multi-faceted approach is essential for effectively addressing the challenges of sonar data processing. The findings from this research not only contribute to the existing body of knowledge but also suggest that further refinement and innovation in correction techniques will continue to play a crucial role in advancing underwater exploration and analysis.

Conclusions

In this study the nature of roll-caused stripe noise of sonar images was analysed. Clear dependency and correlation between stripe noise and roll angle variation has been shown. Statistical rolling mean approach and a simple approach based on machine learning methods were proposed to compensate stripe noise on sonar images and to enhance image quality. The statistical analysis established a foundational correction technique by calculating the horizontal rolling mean, which effectively mitigated random noise while preserving critical details of the sonar images. This method demonstrated its utility in reducing the visual impact of stripe noise, resulting in clearer images that facilitate better object detection and seafloor analysis. However, recognising the limitations of the rolling mean method, particularly its reliance on image data alone, the introduction of a machine learning-based model marked a significant advancement. By incorporating additional parameters such as roll, slant range, altitude, and floor profile, the neural network effectively captured the complex dependencies of image intensity. This approach not only enhanced the correction accuracy but also provided a more adaptable framework for addressing varying sonar conditions.

The results indicated that while both methods improve the quality of sonar images, the machine learning model offers a more robust solution by accommodating external factors that influence intensity. The findings contributed valuable insights into improving sonar image processing, paving the way for more effective underwater exploration and research. The methodologies outlined herein can serve as a foundation for future studies aimed at further refining noise correction techniques in sonar applications.

Future research in this direction will include improvement of the machine learning algorithm to avoid overcompensation of non-striped areas. Development of more universal algorithms and ML models that can incorporate additional features like difference between image lines, sonar type, it's beam pattern, sediment type and others that can potentially lead to development of simple, fast, and universal stripe noise correction method. The benefits of machine learning methods are their universality, relatively simple implementation, and quick inference. It makes them suitable for real time or bulk sonar image correction and processing.

Acknowledgements

The author expresses his gratitude to EvoLogics GmbH for the provided data and support of this research and also grateful to the Armed Forces of Ukraine for their service.

Conflict of Interest

None.

References

[1] Al-Rawi, M., Galdran, A., Isasi, A., & Elmgren, F. (2017). Cubic spline regression based enhancement of side-scan sonar imagery. In *Proceedings of the OCEANS 2017* (pp. 1-7). Aberdeen: Institute of Electrical and Electronics Engineers. doi: 10.1109/oceanse.2017.8084567.

[2] Al-Rawi, M.S. (2016). Intensity normalization of sidescan sonar imagery. In *Proceedings of the international conference on image processing theory, tools and applications* (pp. 1-6). Oulu: IEEE. doi: 10.1109/IPTA.2016.7820967.

[3] Burguera, A., & Oliver, G. (2014). Intensity correction of side-scan sonar images. In *Proceedings of the emerging technology and factory automation* (pp. 1-4). Barcelona: Institute of Electrical and Electronics Engineers. doi: 10.1109/ETFA33519.2014.

[4] Capus, C.G., Banks, A.C., Coiras, E., Tena Ruiz, I., Smith, C.J., & Petillot, Y.R. (2008). Data correction for visualisation and classification of sidescan SONAR imagery. *IET Radar, Sonar & Navigation*, 2(3), 155-169. doi: 10.1049/iet-rsn:20070032.

[5] Chang, Y.-C., Hsu, S.-K., & Tsai, C.-H. (2020). Sidescan sonar image processing: Correcting brightness variation and patching gaps. *Journal of Marine Science and Technology*, 18(6), 721-730. doi: 10.51400/2709-6998.1935.

[6] Chen, Y., Huang, T.-Z., Zhao, X.-L., Deng, L.-J., & Huang, J. (2017). Stripe noise removal of remote sensing images by total variation regularization and group sparsity constraint. *Remote Sensing*, 9(6), article number 559. doi: 10.3390/rs9060559.

[7] Guan, J., Lai, R., & Xiong, A. (2019). Wavelet deep neural network for stripe noise removal. *IEEE Access*, 7, 44544-44554. doi: 10.1109/ACCESS.2019.2908720.

[8] Hughes Clarke, J.E. (2004). Seafloor characterization using keel-mounted sidescan: Proper compensation for radiometric and geometric distortion. In *Canadian hydrography conference 2004* (pp. 1-18). Ottawa: Hydro International.

[9] Li, M., Nong, S., Nie, T., Han, C., Huang, L., & Qu, L. (2022). A novel stripe noise removal model for infrared images. *Sensors*, 22(8), article number 6971. doi: 10.3390/s22082971.

[10] Li, M., Rieck, J., Noheda, B., Roerdink, J., & Wilkinson, M. (2024). Stripe noise removal in conductive atomic force microscopy. *Scientific Reports*, 14(1), article number 3931. doi: 10.1038/s41598-024-54094-w.

[11] Lu, Z., Zhu, T., Zhou, H., Zhang, L., & Jia, C. (2023). An image enhancement method for side-scan sonar images based on multi-stage repairing image fusion. *Electronics*, 12(17), article number 3553. doi: 10.3390/electronics12173553.

[12] Navigation Messages. (n.d.). Retrieved from https://www.lsts.pt/docs/imc/master/Navigation.html.

[13] Shaobo, S., Jianhu, L., Yongcan, Y., Yunlong, W., Shaofeng, B., & Guojun, Z. (2022). Anisotropic total variation regularized low-rank approximation for SSS images radiometric distortion correction. *IEEE Transactions on Geoscience and Remote Sensing*, 60, article number 5925412. doi: 10.1109/TGRS.2022.3229301.

[14] Shippey, G., Bolinder, A., & Finndin, R. (1994). Shade correction of side-scan sonar imagery by histogram transformation. In *Proceedings of the OCEANS'94* (pp. 439-443). Brest: Institute of Electrical and Electronics Engineers. doi: 10.1109/OCEANS.1994.364084.

[15] Sivachandra, K., & Kumudham, R. (2024). A review: Object detection and classification using side scan sonar images via deep learning techniques. In V.K. Gunjan, J.M. Zurada, N. Singh (Eds.), *Modern approaches in machine learning and cognitive science: A walkthrough* (pp. 229-249). Cham: Springer. doi: 10.1007/978-3-031-43009-1_20.

[16] Steiniger, Y., Kraus, D., & Meisen, T. (2022). Survey on deep learning based computer vision for sonar imagery. *Engineering Applications of Artificial Intelligence*, 114, article number 105157. doi: 10.1016/j.engappai.2022.105157.

[17] Wilken, D., Feldens, P., Wunderlich, T., & Heinrich, C. (2012). Application of 2D Fourier filtering for elimination of stripe noise in side-scan sonar mosaics. *Geo-Marine Letters*, 32(4), 337-347. doi: 10.1007/s00367-012-0293-z.

[18] Xia, H., Cui, Y., Jin, S., Bian, G., Liu, G., Zhang, W., & Peng, C. (2024). Improvement of Criminisi's stripe noise suppression method for side-scan sonar images. *Applied Sciences*, 14(20), article number 9574. doi: 10.3390/app14209574.

[19] Ye, X., Yang, H., Li, C., Jia, Y., & Li, P. (2019). A gray scale correction method for side-scan sonar images based on Retinex. *Remote Sensing*, 11(11), article number 1281. doi: 10.3390/rs11111281.

[20] Zhao, J., Yan, J., Zhang, H., & Meng, J. (2017). A new radiometric correction method for side-scan sonar images in consideration of seabed sediment variation. *Remote Sensing*, 9(6), article number 575. doi: 10.3390/rs9060575.

[21] Zhou, P., Chen, J., Tang, P., Gan, J., & Zhang, H. (2024). A multi-scale fusion strategy for side scan sonar image correction to improve low contrast and noise interference. *Remote Sensing*, 16(10), article number 1752. doi: 10.3390/rs16101752.

Корекція смугового шуму, спричиненого креном, на зображеннях гідролокатора бокового огляду

Олександр Катруша*

Аспірант
Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»
03056, просп. Берестейський, 37, м. Київ, Україна
https://orcid.org/0009-0008-7097-4843

Анотація. Забезпечення високої якості зображень, отриманих за допомогою гідролокатора бокового огляду, є важливим для підвищення ефективності підводних досліджень, оскільки такі спотворення, як смуговий шум, можуть ускладнювати аналіз даних. Мета цієї статті – дослідити природу смугового шуму, визначити кореляцію між інтенсивністю зображення і крену гідролокатора, а також розробити новий метод покращення якості гідролокаційних зображень. У дослідженні використовується метод статистичної корекції, заснований на розрахунку горизонтальної ковзної середньої для корекції інтенсивності, а також модель машинного навчання, яка використовує тришарову нейронну мережу для прогнозування горизонтальної ковзної середньої з урахуванням кута падіння променя, висоти гідролокатора над дном та початкової інтенсивності лінії. У дослідженні було застосовано статистичні методи та методи машинного навчання для корекції смугового шуму, спричиненого кренуванням, на гідролокаційних зображеннях, що значно покращило їх якість. Статистичний підхід, що використовує середнє значення горизонтальної хитавиці, ефективно зменшив шум, зберігши при цьому критичні деталі і підвищивши загальну чіткість. Модель машинного навчання включала додаткові параметри, що підвищило точність прогнозування інтенсивності та покращило адаптивність до різних умов положення гідролокатора. Крім того, новий метод дозволяє враховувати змінні умови на місцевості, що робить його гнучким і ефективним в умовах реальних підводних досліджень. Ці результати дають цінну інформацію для вдосконалення методів обробки гідролокаційних зображень, прокладаючи шлях до більш ефективної підводної розвідки та покращення точності виявлення об'єктів на дні моря

Ключові слова: морське дно; відтінки сірого; інтенсивність; крен; корекція; машинне навчання; нейронна мережа

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

Information Technologies and Computer Engineering **Vol. 21, No. 3. 2024**

Article's History: Received: 04.09.2024 Revised: 19.11.2024 Accepted: 26.12.2024

UDC 656.025.4:004.891 10.63341/vitce/3.2024.86

Advancements in automated traffic management using fuzzy logic: Prospects and challenges

Vladyslav Gandrybida*

Postgraduate Student
Vinnytsia National Technical University
21021, 95 Khmelnytske Shose Str., Vinnytsia, Ukraine
https://orcid.org/0009-0001-5091-2716

Dmytro Bondarenko

Postgraduate Student
Vinnytsia National Technical University
21021, 95 Khmelnytske Shose Str., Vinnytsia, Ukraine
https://orcid.org/0009-0003-2927-2624

Volodymyr Sevastyanov

PhD in Technical Sciences, Associate Professor
Vinnytsia National Technical University
21021, 95 Khmelnytske Shose Str., Vinnytsia, Ukraine
https://orcid.org/0000-0001-8385-7146

Abstract. This article reviews modern methods of automated traffic flow control based on fuzzy logic, which enables the processing of incomplete or imprecise information – a characteristic feature of dynamic traffic conditions. The aim of this study was to evaluate the prospects and challenges associated with implementing fuzzy logic in transport system management to enhance the efficiency and safety of road traffic. The paper examined the potential and difficulties of using fuzzy logic for traffic light control, its integration with intelligent transport systems, and its combination with artificial intelligence and Internet of Things technologies. Fuzzy logic allows systems to adapt to real-time changes, considering factors such as traffic intensity, weather conditions, and driver behaviour. The article analysed several examples of the implementation of such systems in different countries, particularly Japan, Germany, and the United States, where fuzzy algorithms have demonstrated effectiveness in reducing congestion, improving road safety, and optimising the use of transport infrastructure. The main challenges associated with implementing these systems are also outlined, including the complexity of developing fuzzy logic models, the need for highly trained experts to configure such systems, and the technical and financial barriers encountered during the modernisation of transport infrastructure. Additionally, the study discussed cybersecurity and data protection issues, which are increasingly relevant given the extensive use of data in intelligent transport systems. The practical significance of this work lies in identifying effective solutions and opportunities for their adaptation to enhance the safety and capacity of urban and intercity transport systems

Keywords: traffic optimization; traffic intensity management; information technology integration; adaptation to traffic conditions; transport infrastructure; intelligent transport systems; traffic congestion

Introduction

The increasing number of vehicles and the growing intensity of road traffic necessitate the implementation of modern technologies for the effective regulation of traffic flows, as traditional methods, such as fixed traffic lights, are unable to adapt quickly to real-time changes. The complexity of traffic, influenced by various factors such as weather conditions, driver behaviour, and fluctuations in flow density, calls for adaptive approaches. Fuzzy logic is a

Suggested Citation:
Gandrybida, V., Bondarenko, D., & Sevastyanov, V. (2024). Advancements in automated traffic management using fuzzy logic: Prospects and challenges. *Information Technologies and Computer Engineering*, 21(3), 86-95. doi: 10.63341/vitce/3.2024.86

*Corresponding author

promising method that accounts for multiple parameters and enables a flexible response to complex traffic situations, thereby helping to reduce congestion, lower emissions, and enhance road safety.

Recent research explored various approaches to automating control processes in the transport sector to ensure the stability and reliability of infrastructure solutions in situations characterised by uncertainty and complex input data. S.K. De & G.C. Mahata (2023) investigated the application of fuzzy logic in a shortage inventory management model that incorporates backorders. Their study demonstrated the effectiveness of fuzzy logic in decision-making under conditions of demand uncertainty, facilitating optimised inventory management. The primary conclusion of the research highlighted fuzzy logic's adaptability to changes in demand, a feature that could be valuable in traffic flow management by optimising resource allocation.

L.M. Markina *et al.* (2021) analysed the application of fuzzy logic in industrial automated systems, demonstrating its ability to adapt parameters to changing conditions. Their study highlighted the potential for adaptive control, which can be utilised to regulate traffic flows by accounting for variations in traffic intensity. S. Lin (2022) examined fuzzy machine learning methods designed for processing large volumes of data in dynamic systems. The integration of fuzzy logic into machine learning enables the handling of imprecise or incomplete data, which is particularly crucial for transport systems that require rapid adaptation to changing traffic conditions.

O.H. Avrunin *et al.* (2021) analysed intelligent automation systems, with a focus on neural network methods for monitoring the technical condition of equipment. Their study aimed to develop a system capable of continuous analysis and failure prediction based on input data, a feature that is essential for the maintenance and reliability of transport networks. V. Bordun (2023) investigated automated traffic control systems for controlled intersections, exploring strategies to reduce congestion and waiting times. The study demonstrated the effectiveness of adaptive traffic light control in enhancing throughput, particularly in large urban areas. The primary conclusion emphasised that integrating automated control systems based on fuzzy logic can significantly optimise traffic management at intersections.

I. Olenych *et al*. (2021) in their book explored the theoretical foundations of fuzzy logic and its application in models of complex systems. The significance of this work lied in its explanation of the principles of fuzzy logic, which form the foundation for the development of adaptive transport systems. The key conclusions of the study indicated that fuzzy logic can effectively facilitate control under uncertain conditions, making it particularly suitable for models requiring rapid adaptation. D. Slavinsky (2023) developed a methodology for multi-criteria analysis of routing algorithms, which can be applied to optimise traffic flow in dynamic conditions.

The analysis of scientific literature highlighted several aspects of automated traffic management based on fuzzy logic that remain insufficiently explored or require further investigation. The first of these aspects is the development and implementation of adaptive models capable of real-time optimisation of traffic flows using fuzzy logic while accounting for the variability of traffic conditions. The second unresolved aspect concerns the integration of fuzzy logic technologies with other intelligent systems, such as neural networks and machine learning algorithms, to enhance the accuracy of decision-making and congestion forecasting. Additionally, an important issue is the evaluation of the effectiveness of such models on a large scale, particularly when applied to extensive urban networks.

This study aimed to further explore the potential of combining fuzzy logic with machine learning to develop adaptive traffic management systems that minimise congestion and improve throughput in dynamic urban environments. The purpose of this article was to review and analyse modern methods of automated traffic control based on fuzzy logic, as well as to identify the prospects and challenges associated with their implementation in high-traffic and dynamically changing conditions.

Materials and Methods

This study conducted a comprehensive analysis of modern methods for automated traffic flow control based on fuzzy logic. The research employed theoretical analysis to examine the principles of fuzzy logic, its characteristics, and its potential applications. Additionally, synthesis was used to evaluate existing approaches and algorithms for road traffic management. Content analysis methods were applied to review scientific literature exploring the use of fuzzy logic in controlling dynamic transport systems. The study analysed contemporary automated traffic control methods, particularly adaptive traffic light control systems, integrated urban transport control systems, and intelligent highway systems. A comparative analysis was conducted to identify the characteristics, advantages, and limitations of each method. The results of this comparison were processed using graphical methods and presented in the form of diagrams.

The analysis considered the adaptability of automated control systems, their efficiency in processing data rapidly, and their overall impact on road safety. Particular emphasis was placed on the ability of these systems to account for complex external conditions that directly influence traffic, such as weather variations, traffic intensity, and fluctuations in driver behaviour. The study specifically focused on adaptive traffic light control systems and intelligent transport management systems, which have demonstrated effectiveness in managing complex urban traffic networks.

The empirical part of the study involved the modelling of control systems using real data from urban and highway networks. The data were collected through induction loops, infrared cameras, and sensors that recorded traffic flow parameters, including speed, traffic density, and vehicle count. Fuzzy inference algorithms were applied in the analysis, enabling the adaptation of traffic light phase

durations in response to real-time changes. Data filtering and evaluation methods were employed to ensure the accuracy and reliability of the modelling process.

The study also examined real-world cases of fuzzy logic system implementation in several countries, including Japan, Germany, and the United States, based on the works of S. Araghi *et al.* (2017) and Z. Pezeshki & S.M. Mazinani (2019). The case study analysis provided insights into the practical application of these technologies, including quantitative indicators such as reductions in congestion, accident rates, and average travel times.

In the final stage of the study, a predictive analysis was conducted to determine the prospects for the development of automated traffic management systems based on fuzzy logic and to assess the challenges associated with their implementation in Ukraine. The findings highlighted the need for modernising transport infrastructure, integrating next-generation sensor technologies, and ensuring a high level of cybersecurity.

Results and Discussion

Fuzzy logic in traffic management

Fuzzy logic, proposed by L. Zadeh (1988), has become a crucial tool for handling imprecise and incomplete information, enabling the modelling of complex processes, particularly traffic flows. Its principles are based on the use of fuzzy sets, where each element has a degree of membership ranging from 0 to 1. This approach allows systems to more accurately represent reality in situations where conditions are continuously changing. Unlike classical binary logic, which permits only the values 0 or 1, fuzzy logic provides flexible state definitions. For example, in traffic light control systems, a gradual scale of traffic intensity from 0 to 1 can be utilised, allowing real-time responses to fluctuating conditions (Atlam *et al.*, 2021).

The concept of fuzzy logic includes fuzzy variables, "if-then" rules, and algorithms that account for situational variables. For instance, a rule might state: "If traffic volume is high and vehicle speed is low, then the green phase of the traffic light should be extended". The use of fuzzy rules enables the system to adjust its decisions dynamically based on current conditions, even when complete information about the road situation is unavailable (Araghi *et al.*, 2015). This is particularly beneficial in urban environments, where traffic flows are subject to constant variations.

Fuzzy logic-based control systems enable more effective responses to sudden changes by adapting to real-time conditions. Dynamic risk management methods allow systems to adjust to evolving situations by identifying both reliable and potentially hazardous road users through reward or penalty mechanisms, thereby increasing decision-making accuracy and efficiency in high-intensity environments (Shaikh *et al.*, 2012). Traditional fixed traffic light systems often struggle to adapt quickly to fluctuations, such as sudden surges in traffic during rush hours, frequently resulting in congestion. Fuzzy logic offers the potential to develop more flexible systems capable of making decisions even when faced with incomplete or imprecise information (Koukol & Marek, 2015). This adaptability enables transport systems to maintain smooth traffic flow even in complex conditions, reducing congestion at intersections and optimising road capacity.

The adaptability of fuzzy logic is achieved through the use of incremental values and fuzzy rules, allowing multiple factors to be considered simultaneously, including vehicle speed, roadworks, and congestion on alternative routes. Fuzzy logic demonstrates high efficiency in dynamic environments due to its use of linguistic variables and inference rules, which facilitate the integration of contextual information in real time. This makes it particularly effective for managing complex and unpredictable systems, such as IoT networks, where flexibility and resilience are crucial (Atlam *et al.*, 2021). For example, in cases where alternative routes become congested due to roadworks on main roads, the system can automatically extend the green phase of traffic lights at key intersections, thereby optimising traffic flow distribution and alleviating congestion (Sabar *et al.*, 2017). As a result, traffic light control systems become more resilient to changes and can rapidly adapt to unforeseen circumstances.

The ability of fuzzy logic to effectively process incomplete information makes it highly valuable for applications in urban traffic management, where conditions change rapidly and unpredictably. For instance, fuzzy models enable risk assessment based on data sensitivity, user action history, and the criticality of operations, enhancing adaptability and decision-making accuracy (Li *et al.*, 2013). During peak traffic periods, specific road sections can automatically adjust their operational modes more frequently to alleviate congestion, easing pressure on highways and minimising traffic build-up, even in the absence of precise information about the underlying causes. Consequently, fuzzy logic not only improves road network capacity but also facilitates the more efficient utilisation of transport infrastructure (Pezeshki & Mazinani, 2019).

Fuzzy logic provides flexibility and enables algorithms to rapidly adapt to new conditions by considering the complex interrelationships between various parameters, such as traffic intensity and weather conditions. This capability allows for swift, optimal decisions that help reduce congestion and enhance the efficiency of transport systems in urban environments. Thus, the integration of fuzzy logic-based systems into urban traffic management contributes to the optimisation of transport flows and the more effective use of infrastructure, even under complex and unpredictable conditions.

Modern methods of automated traffic control

Modern methods of automated traffic control based on fuzzy logic enable effective adaptation to dynamic traffic conditions. One of the key approaches is adaptive traffic light control systems, which adjust signal durations in real time based on the current road situation. Data for these systems are collected using inductive and infrared sensors

that monitor vehicle speed, traffic volume, and congestion levels at intersections. Fuzzy algorithms analyse these parameters to determine the optimal duration of the green phase, thereby reducing delays at intersections and increasing overall traffic throughput. Such systems respond dynamically to fluctuations in traffic intensity, creating flexible conditions to alleviate congestion, particularly during peak periods (Araghi *et al.*, 2015).

Integrated urban transport management systems coordinate both public and private transport, enabling an even distribution of traffic flows. These systems are based on the principles of fuzzy logic, which allows them to dynamically adjust traffic patterns, prioritising public transport when overall flow levels exceed normal thresholds. Data for such systems are gathered from speed sensors and navigation systems that track vehicles in real time. This dual-function approach ensures that public transport receives priority at traffic lights while simultaneously minimising delays for private vehicles. As a result, traffic flows are more evenly distributed, reducing congestion in city centres – an especially critical factor in areas with high traffic intensity (Pezeshki & Mazinani, 2019).

Fuzzy logic is also widely applied in intelligent traffic management systems on highways, where speed limits and information displayed on electronic signs are adjusted according to real-time conditions. These systems account for factors such as traffic density, vehicle speed, and road surface conditions, receiving continuous input from sensors that detect changes in real time. Based on this data, the system can lower speed limits in congested areas, modify signage in response to deteriorating weather conditions, and provide drivers with timely warnings. This approach helps to reduce road accidents and ensures smoother traffic flow, even under high-load conditions (Wei *et al.*, 2018). The characteristics of automated control methods are summarised in Table 1.

Table 1. Main characteristics of automated traffic control methods based on fuzzy logic

Method	Description	Advantages	Disadvantages	Application
Adaptive traffic light control systems	Adjust traffic light phase durations based on traffic intensity at intersections, using data on vehicle count and speed collected from induction and video sensors.	Increase traffic throughput and reduce delays at intersections.	Highly dependent on sensor reliability and system settings.	Urban intersections, central areas of cities with high traffic intensity.
Integrated urban transport management systems	Coordinate the movement of public and private transport, ensuring priority for public transport, particularly during peak hours.	Ensure faster public transport movement and evenly distribute traffic flows.	May lead to delays for private transport.	City centres, areas with high traffic density.
Intelligent highway systems	Regulate speed limits and update information signs in real time, considering traffic density and vehicle speed.	Reduce accident rates and enhance safety under varying traffic intensity conditions.	Require continuous monitoring of sensor data status.	Highways, high-speed zones, roads with heavy traffic.

Source: created by the authors based on Z. Pezeshki & S.M. Mazinani (2019) and S.K. De & G.C. Mahata (2023)

The automated traffic control methods described, based on fuzzy logic, enable effective optimisation of traffic flows through adaptive solutions and the ability to respond swiftly to changing conditions in real time. Each of the examined methods has specific application features: from traffic regulation at urban intersections, where an adaptive traffic light control system reduces average delay times by 15 seconds, to high-speed traffic flow management on highways, where intelligent systems achieve a 50% reduction in congestion.

Figure 1 clearly illustrates the key performance indicators of each method – average delay time, traffic density, and congestion reduction – under various road conditions. The data demonstrate that each approach has distinct advantages tailored to specific needs, enhancing the flexibility and adaptability of the transport system. As a result, congestion is reduced, traffic flow is optimised, and overall road safety is improved, contributing to the efficient utilisation of transport infrastructure.

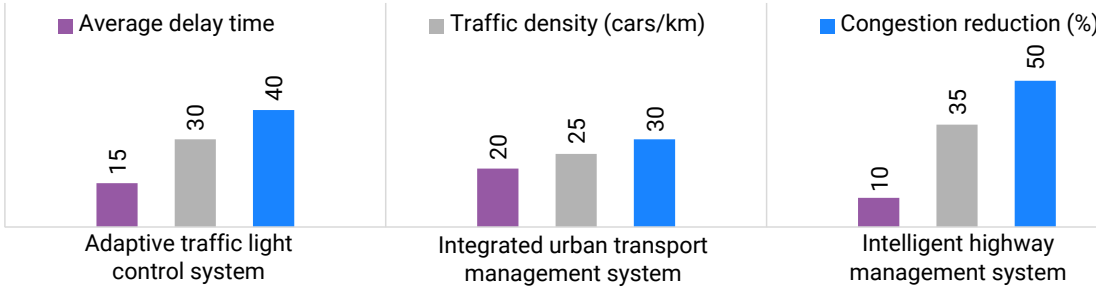

Figure 1. Graph of different traffic management methods effectiveness

Source: created by the authors based on S. Araghi *et al.* (2015), H. Wei *et al.* (2018), Z. Pezeshki & S. Mazinani (2019)

Modern automated traffic management methods based on fuzzy logic not only enhance traffic efficiency but also offer high adaptability to changing conditions. By responding dynamically to real-time situations, these systems optimise road capacity, minimise delays, and contribute to improved traffic safety on city streets and highways. Figure 2 compares three key traffic management methods: adaptive traffic light control, integrated urban transport systems, and intelligent highway control systems.

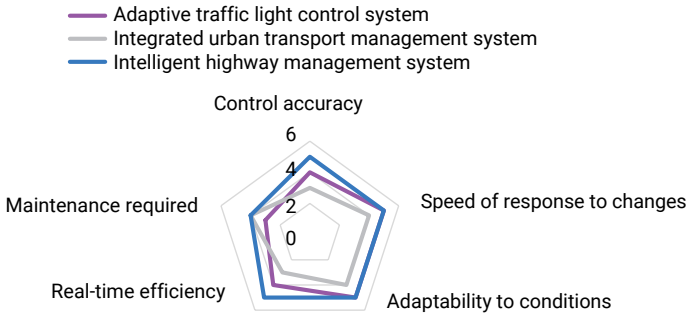

Figure 2. Comparison of traffic management methods by key parameters

Source: created by the authors based on S. Lin (2022)

Each method differs in key characteristics, including control accuracy, response speed to changes, adaptability to conditions, real-time efficiency, and maintenance requirements. Notably, the intelligent highway system demonstrates high performance across all aspects, making it particularly advantageous for expressways with high traffic density. In contrast, the adaptive traffic light system (Fig. 3) proves most effective in reducing delays at intersections, making it well-suited for urban environments with heavy traffic.

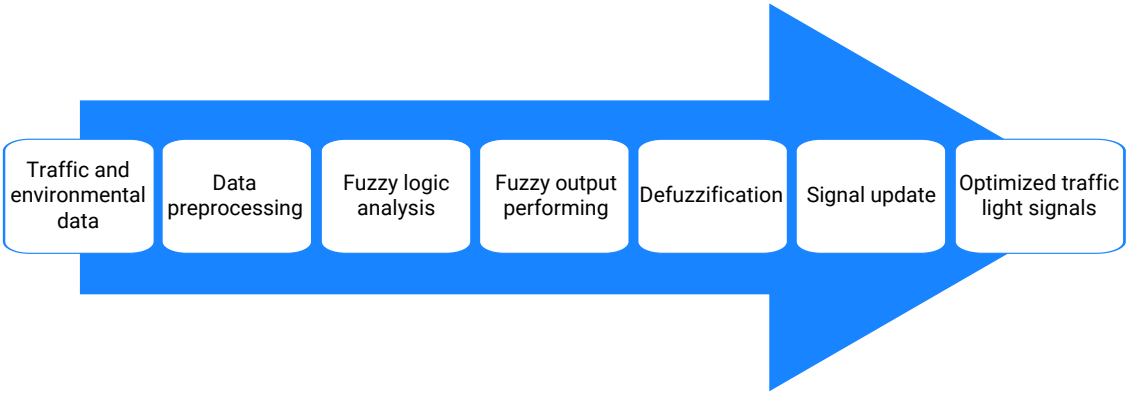

Figure 3. Adaptive traffic light control

Source: created by the authors based on Z. Zhou *et al*. (2017)

The flowchart of the adaptive traffic light control system based on fuzzy logic illustrates the decision-making stages within such a system. In the initial step, traffic flow data – including traffic volume, speed, and weather conditions – is collected using sensors and cameras. This is followed by data processing, where the information is filtered and verified to ensure accuracy.

The system then analyses the data using fuzzy logic, applying membership functions to determine the optimal duration of traffic light signals. The fuzzy output is converted into precise timing adjustments for the green phase, allowing traffic signals to be updated in real time in response to traffic conditions. Continuous performance monitoring enables the system to react swiftly to changes, thereby reducing delays and improving overall traffic throughput.

With the advancement of autonomous vehicles, fuzzy logic is becoming an integral component of driverless car control systems. Autonomous vehicles operate in highly dynamic traffic environments where quick decision-making – such as changing lanes, avoiding obstacles, or maintaining a safe distance from other road users – is crucial. Fuzzy logic enables these systems to adapt to complex and evolving conditions, ensuring real-time safety and efficiency.

Advantages and challenges of implementing fuzzy logic in automated traffic management systems

The implementation of fuzzy logic in automated traffic management systems offers significant advantages, as this approach effectively processes incomplete and uncertain data – an inherent characteristic of dynamic transport

systems. One of the key benefits of fuzzy logic is its ability to enhance road capacity through adaptive traffic flow control. Fuzzy systems enable flexible adjustments to traffic light phase durations, the setting of speed limits, and decision-making processes based on real-time data, including traffic intensity, vehicle speed, and overall road conditions.

Effective traffic flow management reduces congestion, which is particularly crucial in cities with high traffic volumes, where delays and road congestion can result in substantial economic losses. By implementing fuzzy logic, optimal traffic flow distribution can be achieved, leading to more efficient utilisation of existing road infrastructure. Additionally, the integration of fuzzy logic with modern technologies such as artificial intelligence and the Internet of Things (IoT) facilitates the development of intelligent systems capable of adapting to rapid changes in the road environment, ultimately enhancing overall road safety (Pezeshki & Mazinani, 2019).

A second key advantage is the improvement of road safety, enabled by the ability of fuzzy systems to consider multiple risk factors. These systems can analyse weather conditions, vehicle speed, traffic intensity, and other variables that influence accident probability. Due to their ability to rapidly adapt to changing conditions, fuzzy systems can automatically adjust traffic lights and speed limits, helping to prevent hazardous situations on the roads (Avrunin *et al.*, 2021). This proactive approach contributes to a reduction in road accidents and enhances overall safety, making it one of the primary benefits of modern transportation systems.

A third significant advantage is the integration of fuzzy logic with emerging technologies such as artificial intelligence, machine learning, and the Internet of Things (IoT), which creates new opportunities for developing more intelligent and adaptive traffic management systems. The combination of these technologies allows systems to self-learn, adapt to changes in driver behaviour, predict potential variations in road conditions, and make real-time decisions (Wei *et al.*, 2018). For example, fuzzy logic can be applied in the control of autonomous vehicles, facilitating their safe interaction with other road users and enhancing the efficiency of road infrastructure utilisation.

The implementation of automated traffic control systems based on fuzzy logic presents several significant challenges that can complicate both deployment and operation. The complexity of developing fuzzy logic models necessitates meticulous tuning of membership functions and fuzzy inference rules, requiring expert involvement and increasing both time and resource costs. This process slows the real-world deployment of such systems and demands extensive preparation.

Another major challenge is the preparation of the necessary infrastructure. Effective operation requires the installation of induction and infrared sensors, surveillance cameras, and weather monitoring sensors – an expensive and technically demanding process. The costs are particularly high in cities with dense transport networks, where the need for comprehensive traffic monitoring significantly increases infrastructure requirements.

Additionally, integrating fuzzy logic-based systems with existing traffic control systems poses a considerable challenge. The diversity of system architectures and data transfer protocols creates barriers to seamless interaction between components. Ensuring compatibility between legacy and modern technologies is essential, requiring further technical and financial investment.

Processing large volumes of data in real time presents another significant challenge, as such systems require substantial computing power. Large datasets from multiple sensors must be analysed rapidly to ensure efficient system operation (Shmelov, 2021). This necessitates additional costs for technical support and demands optimisation of data storage and processing to maintain system stability. Another critical issue is the absence of universally accepted standards for fuzzy logic control systems, which complicates their scalability and integration at different levels of transport infrastructure. The lack of unified approaches to the development and implementation of such systems creates barriers to their adoption, not only at the municipal level but also at national and international levels, thereby limiting the potential for establishing an integrated transport network.

The implementation of fuzzy logic-based traffic control systems also requires significant financial investment, which can pose challenges for countries with limited budgets or resources. The development, configuration, and integration of these systems involve not only software and hardware expenses but also the modernisation of transport infrastructure. This financial burden can be particularly challenging for countries struggling with severe traffic congestion but lacking the resources for large-scale investment.

Furthermore, although fuzzy logic is specifically designed to handle imprecise or incomplete data, the accuracy and reliability of input data remain crucial to system performance (Koukol & Marek, 2015). If sensors collecting traffic data provide inaccurate readings or transmit information with delays, the system may make incorrect decisions, leading to inefficient real-time operation – an issue that is particularly critical under heavy traffic conditions.

Lastly, cybersecurity and data protection present significant challenges, as intelligent transportation systems using fuzzy logic process vast amounts of information collected from sensors, video cameras, and other sources. This makes such systems potentially vulnerable to cyberattacks, which could have serious consequences for both traffic safety and data confidentiality. Safeguarding systems against unauthorised access and ensuring secure data management are essential considerations that must be addressed during the implementation phase.

Global experience in the use of automated traffic control systems

Fuzzy logic has been successfully implemented in numerous countries to enhance the efficiency of automated traffic control systems, enabling dynamic adaptation of

traffic flow management to changing conditions. A notable example is Japan, where fuzzy algorithms are extensively used, particularly in major cities such as Tokyo and Osaka. In these urban areas, fuzzy logic dynamically adjusts traffic light timings based on traffic intensity, weather conditions, and time of day. According to research by S. Araghi *et al*. (2017), these systems have reduced traffic congestion by 20% and decreased average travel times by 12% during peak hours, leading to more efficient traffic flow and shorter commute times for city residents.

In Germany, fuzzy logic has been widely applied on motorways, where fluctuating traffic conditions and adverse weather frequently impact road safety. Fuzzy algorithms are used to automatically adjust speed limits based on real-time data on traffic density across different road sections. Integration with meteorological services allows for the immediate reduction of speed limits in cases of fog, rain, or snowfall. Studies have shown that this approach has led to a 15% decrease in highway accidents and an 18% reduction in overall congestion, ensuring stable road capacity even under challenging conditions (Araghi *et al*., 2015; Koukol & Marek, 2015).

In the United States, fuzzy logic is a key component of intelligent transportation systems (ITS), which are widely deployed in cities with high traffic volumes, such as Los Angeles and New York. These systems enable real-time traffic data analysis and facilitate the redirection of traffic flows in cases of congestion or accidents. The application of fuzzy logic within the ITS of Los Angeles has resulted in a 10% reduction in average travel times and a 25% decrease in congestion during peak hours. In New York, the adoption of such systems has contributed to an 8% reduction in CO_2 emissions from transportation by optimising routes and minimising idle time in traffic jams (Wei *et al*., 2018; Pezeshki & Mazinani, 2019).

Global experience demonstrates the effectiveness of fuzzy logic in automated traffic management systems for reducing congestion, minimising accidents, and improving road capacity. In countries such as Japan, Germany, and the United States, fuzzy algorithms are employed to dynamically adjust traffic lights, regulate speed limits, and redirect traffic flows based on real-time conditions, weather variations, and time of day. The results indicate significant reductions in travel times, congestion, and transport-related emissions, making these systems an essential tool for enhancing the efficiency and sustainability of transport infrastructure.

Prospects for the development of automated traffic management systems

The future development of automated traffic management systems based on fuzzy logic encompasses several promising areas that offer new opportunities to enhance their efficiency and adaptability. One of the most significant advancements is the integration of fuzzy logic with artificial intelligence (AI) and machine learning technologies. This combination will enable the creation of more autonomous systems that can not only respond to real-time traffic conditions but also predict potential road issues. Such integration will allow these systems to automatically adjust their control algorithms to prevent congestion and emergency situations, thereby contributing to overall road safety improvements (Nguyen *et al*., 2016).

Another key area of development is the advancement of sensor technologies, which provide critical data for traffic management systems. Enhancing the accuracy and speed of data transmission from sensors that monitor traffic parameters, weather conditions, and road surface quality will significantly improve system efficiency. Next-generation sensors will offer more precise and timely information, enabling traffic management systems to respond to changes more quickly and accurately. This improvement will be particularly beneficial for managing traffic flow in high-density urban areas and under challenging weather conditions.

Another key area of development is the creation of global intelligent transport networks that integrate individual city and regional traffic management systems into a single, unified system. Such networks will optimise traffic flows not only within individual cities or regions but also at national and even international levels, facilitating seamless transport management across broader geographic areas (Avrunin *et al*., 2021).

Additionally, advancements in Big Data processing methods represent another crucial direction in the evolution of modern transport systems. The development of more sophisticated techniques for handling large volumes of traffic data will reduce processing times, enabling more efficient decision-making based on accurate and timely information. Improved data analysis capabilities will enhance the ability to predict traffic fluctuations and adapt management strategies to real-time conditions, thereby minimising congestion and reducing accident risks.

The findings of this research highlight the significant potential of fuzzy logic in optimising automated traffic flow management, as supported by numerous studies in the field. For instance, the study by H.F. Atlam *et al*. (2021) demonstrated the effectiveness of fuzzy logic in decision-making under conditions of uncertainty – a crucial factor in transportation systems. Their approach, initially designed for an access control system, enables risk modelling and adaptive adjustments to management parameters. This method closely parallels the application of fuzzy logic in dynamic traffic light control, where real-time adjustments improve traffic efficiency and flow management.

C. Zhao *et al*. (2018) proposed a method for intersection traffic control that utilises both current and previous signal phases to adjust the directions of dynamic waiting lanes, thereby minimising average traffic delays. Their control model, optimised using a genetic algorithm, demonstrated 31.8% reduction in intersection delays compared to existing systems. The study further indicated that the model's effectiveness increases as traffic volumes rise, making it particularly promising for managing congestion. These findings align with the results of the current study, reinforcing the potential of adaptive traffic management solutions.

Z. Pezeshki & S.M. Mazinani (2019) examined the advantages of neuro-fuzzy models, which integrate neural networks with fuzzy logic, demonstrating their high efficiency in predicting energy consumption. Although their study focused on a different domain, the underlying principles of neuro-fuzzy approaches remain highly relevant for transportation systems, where adaptive decision-making is crucial. Their analysis underscored similar traffic optimisation strategies, particularly the integration of fuzzy logic with artificial intelligence, which significantly enhances system adaptability.

O. Avrunin *et al*. (2021) explored intelligent automation systems for monitoring the technical condition of transport, ensuring uninterrupted traffic flow. Their findings confirmed the importance of processing real-time data for adaptive control and highlight the critical role of sensor reliability, a fundamental aspect of fuzzy logic-based transport systems. The conclusions drawn from these studies validate the findings of the current research, demonstrating that adaptive fuzzy logic-based systems contribute to optimising traffic flow in large cities. The analysis of research results confirmed that fuzzy logic is a universal and effective tool for enhancing automated traffic management.

Conclusions

The study of modern methods of automated traffic flow control based on fuzzy logic has provided valuable insights into the prospects and challenges of their implementation in transport systems. These findings highlighted the potential of fuzzy logic to enhance traffic management efficiency and improve road safety. The analysis confirmed that fuzzy logic enables traffic control systems to effectively adapt to changing traffic conditions and optimise traffic flow management. The research examined the theoretical foundations of fuzzy logic in automated traffic control systems, assessed existing traffic light control algorithms, and explored the integration of fuzzy logic with intelligent transport systems and emerging technologies, such as artificial intelligence (AI) and the Internet of Things (IoT). The results confirmed that implementing such systems contributes to reducing congestion, increasing road capacity, and enhancing road safety, particularly in cities with high traffic intensity. Furthermore, an analysis of case studies from Japan, Germany, and the United States validated the effectiveness of fuzzy algorithms in decreasing congestion and improving transport infrastructure efficiency.

The overall findings reaffirm that fuzzy logic is a promising tool for transport automation, allowing systems to adapt dynamically to real-world traffic conditions. Its implementation helps address key urbanisation challenges, such as traffic congestion and high accident rates. Moreover, combining fuzzy logic with AI and IoT technologies unlocks new possibilities for the development of intelligent transport systems capable of adaptive and efficient real-time traffic management.

However, certain challenges remain, including the complexity of developing fuzzy logic models, integration with existing traffic management systems, and high implementation costs. Future research could focus on developing standardised methodologies that facilitate the integration of fuzzy logic with other advanced technologies in transportation systems. Expanding experimental data and improving automated control methods could be key steps towards the further evolution of intelligent transportation systems, ultimately enhancing transport infrastructure and improving the quality of life in urban environments.

Acknowledgements

None.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] Araghi, S., Khosravi, A., & Creighton, D. (2015). A review on computational intelligence methods for controlling traffic signal timing. *Expert Systems with Applications*, 42, 1538-1550. doi: 10.1016/j.eswa.2014.09.003.

[2] Araghi, S., Khosravi, A., Creighton, D., & Nahavandi, S. (2017). Influence of meta-heuristic optimization on the performance of adaptive interval type-2 fuzzy traffic signal controllers. *Expert Systems with Applications*, 71, 493-503. doi. 10.1016/j.eswa.2016.10.066.

[3] Atlam, H.F., Walters, R.J., Wills, G.B., & Daniel, J. (2021). Fuzzy logic with expert judgment to implement an adaptive risk-based access control model for IoT. *Mobile Networks and Applications*, 26, 2545-2557. doi: 10.1007/s11036-019-01214-w.

[4] Avrunin, O.H., Vladov, S.I., Petchenko, M.V., Semenets, V.V., Tatarinov, V.V., Telnova, H., Filatov, V.O., Shmelov, Yu.M., & Shalyapina, N.O. (2021). *Intelligent automation systems*. Kreminchuk: Novabook.

[5] Bordun, V.L. (2023). *Justification of the use of an automated traffic control system at a controlled intersection in the city of Ternopil*. (Master's thesis, Ternopil National Technical University, Ternopil, Ukraine).

[6] De, S.K., & Mahata, G.C. (2023). Decision making in fuzzy reasoning to solve a backorder economic order quantity model. *RAIRO Operations Research*, 57, 993-1007. doi: 10.1051/ro/2023051.

[7] Koukol, M., & Marek, L. (2015). Fuzzy logic in traffic engineering: A review on signal control. In F. Wang (Ed.), *Mathematical problems in engineering*. Hoboken: Wiley. doi: 10.1155/2015/979160.

[8] Li, J., Bai, Y., & Zaman, N. (2013). A fuzzy modeling approach for risk-based access control in eHealth cloud. In *Proceedings of the 12th IEEE international conference on trust, security and privacy in computing and communications* (*TrustCom 2013*) (pp. 17-23). Melbourne: IEEE. doi: 10.1109/TrustCom.2013.66.

[9] Lin, S. (2022). Fuzzy machine learning methods. In *Fuzzy-AI model and big data exploration* (pp. 117-172). Berlin: Springer. doi: 10.1007/978-3-662-56339-7_6.

[10] Markina, L.M., Satsyk, V.O., & Smolyankin, O.O. (2021). Application of fuzzy logic in the automatic regulation system of mash concentration in alcohol production. *Perspective Technologies and Devices*, 19, 78-84.

[11] Nguyen, P.T., Passow, B N., & Yang, Y. (2016). Improving anytime behavior for traffic signal control optimization based on NSGA-II and local search. In *Proceedings of international joint conference on neural networks* (pp. 4611-4618). Vancouver: IEEE. doi: 10.1109/IJCNN.2016.7727804.

[12] Olenych, I., Sinkevych, O., Salamakha, O., & Prytula, M. (2021). Text tone determination using fuzzy logic. *Applied Computer Systems*, 26(2), 158-163. doi: 10.2478/acss-2021-0019.

[13] Pezeshki, Z., & Mazinani, S. M. (2019). Comparison of artificial neural networks, fuzzy logic, and neuro-fuzzy for predicting optimization of building thermal consumption: A survey. *Artificial Intelligence Review*, 52, 495-525. doi: 10.1007/s10462-018-9630-6.

[14] Sabar, N.R., Kieu, L.M., Chung, E., Tsubota, T., & de Almeida, P.E.M. (2017). A memetic algorithm for real world multi-intersection traffic signal optimization problems. *Engineering Applications of Artificial Intelligence*, 63, 45-53. doi: 10.1016/j.engappai.2017.04.021.

[15] Shaikh, R.A., Adi, K., & Logrippo, L. (2012). Dynamic risk-based decision methods for access control systems. *Computers & Security*, 31(4), 447-464. doi: 10.1016/j.cose.2012.02.006.

[16] Slavinsky, D.Y. (2023). *Methodology for multi-criteria comparative analysis of routing algorithms in mobile sensor networks*. (Master's thesis, National Technical University of Ukraine "Kyiv Polytechnic Institute named after Ihory Sikorsky", Kyiv, Ukraine).

[17] Wei, H., Zheng, G., Yao, H., & Li, Z. (2018). Intellilight: A reinforcement learning approach for intelligent traffic light control. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 2496-2505). New York: ACM. doi: 10.1145/3219819.3220096.

[18] Zadeh, L.A. (1988). Fuzzy logic. *Computer*, 21(4), 83-93. doi: 10.1109/2.53.

[19] Zhao, C., Chang, Y., & Zhang, P. (2018). Coordinated control model of main-signal and pre-signal for intersections with dynamic waiting lanes. *Sustainability*, 10(8), article number 2849. doi: 10.3390/su10082849.

[20] Zhou, Z., De Schutter, B., Lin, S., & Xi, Y. (2017). Two-level hierarchical model-based predictive control for large-scale urban traffic networks. *IEEE Transactions on Control Systems Technology*, 25, 496-508. doi. 10.1109/TCST.2016.2572169.

Огляд сучасних методів автоматизованого керування трафіком на основі нечіткої логіки: перспективи та виклики

Владислав Гандрибіда

Аспірант
Вінницький національний технічний університет
21021, вул. Хмельницьке шосе, 95, м. Вінниця, Україна
https://orcid.org/0009-0001-5091-2716

Дмитро Бондаренко

Аспірант
Вінницький національний технічний університет
21021, вул. Хмельницьке шосе, 95, м. Вінниця, Україна
https://orcid.org/0009-0003-2927-2624

Володимир Севастьянов

Кандидат технічних наук, доцент
Вінницький національний технічний університет
21021, вул. Хмельницьке шосе, 95, м. Вінниця, Україна
https://orcid.org/0000-0001-8385-7146

Анотація. Стаття присвячена огляду сучасних методів автоматизованого керування транспортними потоками на основі нечіткої логіки, яка надає можливість обробляти неповну або нечітку інформацію, що є характерним для динамічних умов дорожнього руху. Метою цього дослідження була оцінка перспектив і викликів впровадження нечіткої логіки в управління транспортними системами для підвищення ефективності та безпеки дорожнього руху. У роботі акцентовано увагу на перспективах та викликах використання нечіткої логіки для керування світлофорами, інтеграції з інтелектуальними транспортними системами, а також на її поєднанні з технологіями штучного інтелекту та Інтернету речей. Нечітка логіка дозволяє адаптувати системи до змін у реальному часі, враховуючи такі фактори, як інтенсивність руху, погодні умови та поведінкові особливості водіїв. У статті наведено аналіз низки прикладів впровадження таких систем у різних країнах світу, зокрема в Японії, Німеччині та США, де нечіткі алгоритми демонструють ефективність у зниженні заторів, підвищенні безпеки на дорогах та оптимізації використання транспортної інфраструктури. Також окреслено основні виклики впровадження цих систем, серед яких складність побудови моделей нечіткої логіки, необхідність високої експертної підготовки для налаштування таких систем, а також технічні та фінансові бар'єри, що виникають під час модернізації транспортної інфраструктури. Окрім того, обговорено питання кібербезпеки та захисту даних, що стають актуальними в умовах використання великих обсягів інформації в інтелектуальних транспортних системах. Практична цінність цієї роботи полягає у визначенні ефективних рішень та можливостей їх адаптації для підвищення безпеки та пропускної здатності міських та міжміських транспортних систем

Ключові слова: оптимізація дорожнього руху; керування інтенсивністю трафіку; інтеграція інформаційних технологій; адаптація до умов руху; транспортна інфраструктура; інтелектуальні транспортні системи; затори

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

Information Technologies and Computer Engineering **Vol. 21, No. 3. 2024**

Article's History: Received: 09.09.2024 Revised: 22.11.2024 Accepted: 26.12.2024

UDC 004.451.1:37.036.2 10.63341/vitce/3.2024.96

Mathematical models of individualised learning based on decision theory

Ivan Vovchok*

Postgraduate Student
Uzhhorod National University
88000, 3 Narodna Sq., Uzhhorod, Ukraine
https://orcid.org/0000-0001-8603-7899

Abstract. The study provided theoretical substantiation and development of a system of mathematical models for the individualisation of the educational process based on the integration of decision theory methods. The developed system of mathematical models is based on a metamodel that combines four mathematical paradigms through an interaction matrix, the elements of which are determined by the function of cognitive compatibility, temporal consistency and interaction efficiency. The introduction of the method of optimising partial trajectories, based on recursive updating of model parameters through the analysis of intermediate results, increased the accuracy of parameter settings and ensured smooth adaptation to the individual learning rate. The developed modification of the Bellman equation with the function of the complexity of the learning material made it possible to formalise the process of optimising long-term learning strategies by addressing individual cognitive characteristics. The analysis of the stochastic nature of the learning process through an extended transition matrix was used to mathematically describe the processes of forgetting and repeating the material using a system of differential equations with time-dependent coefficients that account for the intensity of learning and individual memory characteristics. The study of collaborative learning mechanisms using the game-theoretic approach revealed the synergistic effects of group learning through nonlinear functions of interaction between participants in the educational process and has allowed the development of methods for forming optimal learning groups based on individual goals. The proposed system of multidimensional evaluation, implemented through a composite objective function, covers a wide range of indicators from basic knowledge acquisition to the development of higher-order metacognitive skills, including cognitive, metacognitive and motivational components, which provides a reliable tool for assessing the stability of learning trajectories and determining the level of adaptability of the system to individual characteristics of students

Keywords: adaptive educational systems; Bayesian optimisation; Bellman function; Markov processes; game-theoretic approach; cognitive trajectories

Introduction

Modern educational space features a rapid transition from unified approaches to individualised learning, driven by the growing need to effectively adapt the educational process to the individual characteristics of each student. Traditional teaching methods often do not account for the diversity of cognitive styles, learning styles and prior experience of students, which leads to a decrease in the effectiveness of the educational process. Mathematical modelling of individualised learning processes is of particular importance in the context of the development of digital educational technologies and decision support systems. The use of mathematical models formalises and optimises the process of adapting educational content, creating personalised educational trajectories and predicting learning outcomes. The 2019-2024 research significantly expanded the understanding of multicriteria approaches in the educational context. H. Taherdoost & M. Madanchian (2023) developed a comprehensive approach to the application of Multi-Criteria Decision-Making (MCDM) methods. This is a group of methods for decision-making that account for multiple criteria or alternatives. Detailing their potential for creating adaptive educational systems and proposing a

Suggested Citation:
Vovchok, I. (2024). Mathematical models of individualised learning based on decision theory. *Information Technologies and Computer Engineering*, 21(3), 96-107. doi: 10.63341/vitce/3.2024.96

*Corresponding author

methodological framework for decision-support systems in education. In the development of this direction, I. Canco *et al.* (2021) presented a practical implementation of the Analytical Hierarchy Process (AHP) method, demonstrating its effectiveness in shaping individual educational trajectories and developing a system of criteria for assessing the quality of the educational process. A fundamentally new perspective of decision-making mechanisms was proposed in a study by J.C. Peterson *et al.* (2021), which, through large-scale experiments and the use of machine learning methods, identified fundamental patterns in decision-making processes, which can be used as the basis for the development of predictive models with greater accuracy in education.

The introduction of Bayesian networks and adaptive models created new horizons in the field of automated learning assessment. W. Xing *et al*. (2021) developed an innovative model for assessing students' engineering projects that not only accounts for multiple success factors but also adapts to the individual progress of each student, ensuring objective and personalised assessment. A significant step forward was made by L.G. Eglington & P.I. Pavlik (2023) proposed methods of rapid learning optimisation that accounts for the individual pace of learning, cognitive characteristics and previous experience of students, which can be used for dynamic adjustment of task complexity and learning pace. A comprehensive view of the future of personalised education was presented in a study by S. Maghsudi *et al.* (2021), which not only outlined promising areas of development but also proposed specific mechanisms for integration of artificial intelligence into the educational process, highlighting the need for a balance between automation and preservation of the human factor in educational process.

Recent advances in the field of Bayesian optimisation significantly expanded the possibilities of mathematical modelling of educational processes. X. Wang *et al.* (2023) presented a thorough analysis of modern Bayesian optimisation methods, systematising existing approaches and outlining promising areas for their application in the educational context, especially for optimising the parameters of educational systems and predicting educational outcomes. The innovative concept of inexact Bayesian optimisation was proposed by J. Rodemann & T. Augustin (2024), which improves efficiency in uncertainty in the educational process, addressing multiple factors of influence and their interdependence. A significant breakthrough in the development of predictive models was made by P. Jiang & X. Wang (2020), they have developed a cognitive diagnostic method that not only predicts student performance but also identifies specific areas that require additional attention, which allows for proactive adaptation of the learning process.

A comprehensive analysis of existing research reveals both significant achievements and significant gaps in existing approaches to the mathematical modelling of individualised learning. S. Minn (2022), exploring the possibilities of artificial intelligence in knowledge assessment, emphasised the need to create integrated systems that combine different methodological approaches. In particular, the mechanisms of interaction of different mathematical models within a single decision-making system, methods of validation and evaluation of the effectiveness of such models in a real educational environment, as well as ways to adapt the models to different educational contexts and cultures remain insufficiently researched. Particular attention should be devoted to the development of methods that address not only the cognitive aspects of learning but also the emotional and motivational dimensions, ensuring a genuinely holistic approach to individualising the educational process. While existing research has laid a solid foundation for further advancement in this area, there remains a pressing need for systematic integration and practical validation of these methods across a wide range of educational contexts.

Research objective: theoretical substantiation and development of a system of mathematical models of individualised learning based on the integration of Bayesian optimisation, dynamic programming, Markov processes and game theory to create a methodological basis for decision-making in the educational process. Tasks of the study were:

- To analyse the theoretical foundations and methodological approaches to the construction of mathematical models of individualised learning, in particular models of Bayesian optimisation, dynamic programming, Markov processes and game theory.
- To develop a theoretical justification for the integration of various mathematical models into a single decision-making system for individualised learning, addressing their interaction and limitations.
- To analyse the theoretical aspects of evaluating the effectiveness of the proposed mathematical models and develop methodological foundations for their practical application in the context of predicting educational outcomes.

Materials and Methods

The study analysed four main mathematical approaches – Bayesian optimisation, dynamic programming, Markov processes and game theory – as the basic components of an individualised learning system. Within the framework of these approaches, a multicomponent fitness function was investigated, which reflects a complex vector of characteristics of learning materials, including the complexity of the content, the time required for learning, the way the material is presented, the level of interactivity and the pre-requisites for students' knowledge. The analysis of these characteristics was used to explore methods for optimising partial learning trajectories based on recursive updating of model parameters following intermediate learning outcomes. In the study of long-term learning strategies, the method of dynamic programming was applied the modified Bellman equation was analysed in the educational context. The analysis included the study of the dependencies between the function of the complexity of the educational material and the processes of optimising individual learning

trajectories. The mechanisms of adaptation of learning parameters to individual characteristics of students, including the pace of learning, cognitive styles and previous experience, are investigated. This determined the principles of building long-term strategies for individualised learning, considering the dynamics of the learning process and the peculiarities of material perception.

The use of Markov processes addressed the stochastic nature of learning through the analysis of an extended transition matrix that considers conditional probabilities for different learning actions and states. When studying the processes of forgetting and repeating material, a system of differential equations with time-dependent coefficients reflecting changes in knowledge acquisition was analysed. This analysis examined the relationships between the intensity of learning, the complexity of the material and the individual characteristics of students' cognitive processes, including memory characteristics, information processing speed and the stability of the acquired knowledge.

The study of game theory analysed theoretical and game models of collaborative learning, in particular, the functions of forming learning coalitions and their impact on the educational process. The synergistic effects of students' interaction were investigated through the analysis of nonlinear functions that reflect the individual characteristics of participants in the educational process. Based on the game-theoretic approach, the principles of group learning assessment and criteria for the formation of study groups were analysed, addressing individual goals, level of training and peculiarities of interaction between participants in the educational process. In the study of the interaction of mathematical models, their main properties were analysed: computational complexity, stability of functioning and adaptability to different educational contexts. The methods of optimising model parameters and mechanisms of their integration into a single decision-making system were investigated. The possibilities of combining Bayesian optimisation, dynamic programming, Markov processes and game theory to provide a comprehensive approach to individualised learning are analysed. The conditions and limitations of interaction between different mathematical models in the context of the educational process are determined.

The study analysed the systems of multicriteria evaluation of mathematical models through a composite objective function that considers the cognitive, metacognitive and motivational components of the educational process. The criteria for assessing the stability of learning trajectories and the level of adaptability of the system to the individual characteristics of students were investigated. The methodological approaches to evaluating the effectiveness of the models both in the short term (assessment of basic knowledge acquisition, intermediate results) and the long term (development of metacognitive skills, sustainability of acquired knowledge) were determined. The results of the theoretical study were systematised in two analytical tables, which present a comparative analysis of the theoretical characteristics of mathematical models and criteria for assessing their effectiveness in the context of individualised learning.

Results

Theoretical and methodological foundationsof mathematical modelling of individualised learning

The theoretical study of mathematical models of individualised learning reveals the potential of an integrated approach that combines four mathematical paradigms, each of which is unique in the overall architecture of the system. Bayesian optimisation in the mathematical modelling system provides adaptive adjustment of the learning content parameters through a multicomponent fitness function $f(x)$, where x represents a multidimensional vector of characteristics of learning materials, including content complexity, learning time, type of material presentation, and other key parameters. The implementation of the partial trajectory optimisation method is based on recursive updating of the model parameters through the analysis of intermediate learning outcomes. This approach improves the accuracy of parameter settings, ensuring adaptation to the individual pace of learning for each student. The theoretical analysis of mathematical models of this class demonstrates their potential for individualising learning when working with groups of students with different learning styles (Zhang *et al.*, 2023). The dynamic programming system is developed with a focus on optimising long-term learning strategies through a modified Bellman equation adapted to the specifics of the educational context. The mathematical formalisation is based on the principle of optimality.

$$V \times (s) = max[R(s, a) + \gamma \Sigma P(s'|s, a)\, V^*(s')], \quad (1)$$

where: $V \times (s)$ – optimal value function for state s; $R(s, a)$ – direct reward for action a in state s; γ – discount factor for future rewards; $\gamma \in [0,1]$, $P(s'|s, a)$ – probability of transition to state s' from state s when performing action a.

To account for the specifics of the educational process, an additional function of the complexity of the educational material $C(s, a)$ is introduced, which modifies the basic equation to the following form:

$$V^*(s) = max[R(s, a)/C(s, a) + \gamma \Sigma P(s'|s, a) V^*(s')], \quad (2)$$

where: $C(s, a)$ – function of the complexity of the learning material, the rest of the notation is similar to formula (1).

The value function $V(s)$ provides a comprehensive assessment of the optimality of learning trajectories through multivariate analysis, which is implemented by the parameter vector θ:

$$\theta = \{\theta 1, \ldots, \theta n\}, \quad (3)$$

where: θi – components that correspond to different aspects of learning (cognitive development, metacognitive skills, level of learning).

Each parameter θi is optimised according to a function:

$$\theta i^* = argmax\{E[V(s|\theta i)] - \lambda R(\theta i)\}, \quad (4)$$

where: θi^* – optimal value of the parameter; $E[V(s|\theta i)]$ – expected value of the state at parameter θi; $\lambda R(\theta i)$ – regularisation term to prevent overlearning.

Long-term effects are modelled using a composite forecasting function:

$$F(s, t) = \sum \alpha i fi(s, t) + \beta \sum \sum wij(fi(s, t) \cdot fj(s, t)), \quad (5)$$

where: αi – weighting coefficients of individual components; wij – interaction coefficients between different aspects of learning; $fi(s, t)$ – components of the prediction function; β – overall interaction coefficient.

The integration of Markov processes into the system addresses the stochastic nature of the learning process through the transition matrix $P = [pij]$, supplemented by conditional probabilities of transitions at different learning actions:

$$P(s'|s, a) = P(s'|s) \cdot P(a|s, s'), \quad (6)$$

where: $P(s'|s)$ – probability of transition from state s' to state s'; $P(a|s, s')$ – conditional probability of action a during the transition between states.

The developed variational approach for analysing time series of training data is based on functional optimisation:

$$L'(\theta) = L(\theta) - \lambda \sum ||\nabla \theta L(\theta)||^2, \quad (7)$$

where – $L'(\theta) = E[log\ p(x|\theta)] - DKL(q(z|x)||p(z|x,\theta))$ basic functionality; λ – regularisation coefficient; $\nabla \theta L(\theta)$ – functionality gradient.

The stability of learning trajectories is analysed using a range of Lyapunov indicators:

$$\lambda i = lim(t \to \infty)\ (1/t) log(||\delta xi(t)||/||\delta xi(0)||), \quad (8)$$

where: λi – Lyapunov exponent; $\delta xi(t)$ – trajectory deviation at time t; $\delta xi(0)$ – initial deviation.

Such an extended theoretical framework ensures not only the mathematical rigour of the model but also its practical applicability in various educational contexts, guaranteeing robustness under different initial conditions and external influences. The model demonstrates effectiveness in nonlinear learning trajectories and complex conceptual structures, which is confirmed by the theoretical analysis of its properties. The theoretical and game component of the system is implemented through a multi-level model of strategic interaction of participants in the educational process. The key focus is on modelling collaborative learning and competitive interaction. In the mathematical formalisation, each participant in the educational process acts as a player with a personal set of strategies:

$$Si = \{si1, \ldots, sin\}, \quad (9)$$

where: Si – set of strategies of the i-th player; sij – j-th strategy of the i-th player.

To improve the effectiveness of learning, mechanisms for forming coalitions were developed to group students based on common goals and interests:

$$C = \{c1, \ldots, ck\}, \quad (10)$$

where: C – set of coalitions; ci – i-th coalition; k – number of coalitions.

To assess the success of the learning process, the system uses a multi-criteria evaluation function that accounts for various aspects of learning:

$$F(x) = \sum wi \cdot fi(x), \quad (11)$$

where: wi – weighting coefficients; $fi(x)$ – functions that reflect different aspects of the learning process; x – vector of learning process parameters.

When students interact, synergistic effects arise, which are modelled through nonlinear interaction functions:

$$g(xi,xj) = \alpha \cdot xi \cdot xj + \beta \cdot (xi + xj), \quad (12)$$

where: xi, xj – student characteristics; α, β – interaction coefficients.

To create a holistic system of adaptive learning, the four main paradigms are integrated through a metamodel:

$$M = \{BO, DP, MP, GT\}, \quad (13)$$

where: BO – Bayesian optimisation; DP – dynamic programming; MP – Markov processes; GT – game theory.

The analysis of the theoretical characteristics of the developed mathematical models can be used to systematise their key parameters and features of application in the context of individualised learning (Table 1).

Table 1. Theoretical characteristics of mathematical models of individualised learning

Evaluation criterion	BO	DP	MP	GT
Predictive power	High	Average	Average	Low
Theoretical complexity	O(n log n)	O(n²)	O(n²)	O(n^k)
Flexibility of the model	8.5	6.8	8.2	6.4

Table 1. Continued

Evaluation criterion	BO	DP	MP	GT
Theoretical scalability	High	Limited	High	Limited
Mathematical complexity	Moderate	High	Moderate	High

Notes: BO – Bayesian optimisation; DP – dynamic programming; MP – Markov processes; GT – game theory. The flexibility of the model was assessed on a theoretical 10-point scale based on the possibility of adaptation to different educational contexts

Source: compiled by the author based on H. Wu & F. Noé (2020), Q. Cappart *et al*. (2021), D.P. Bertsekas (2022), D. Zhang *et al*. (2023), S. Tu *et al*. (2024)

The table presents a comparative assessment of four mathematical models in the context of individualised learning. Bayesian optimisation (*BO*) stands out for its high predictive power, theoretical scalability, and the highest level of flexibility. Dynamic programming (*DP*) and Markov processes (*MP*) have moderate predictive power but are distinguished by their high mathematical complexity. Game theory (*GT*), despite its low predictive power, remains valuable for analysing complex systems due to its applicability to multifactor scenarios. The theoretical analysis of the developed system of mathematical models reveals their potential for further development of the theory of individualised learning. Each of the presented models has advantages in the context of mathematical modelling of educational processes.

An integrated system of mathematical decision-making models

The integrated system of mathematical models of individualised learning is based on the metamodel $M = \{BO, DP, MP, GT\}$ (where *BO* stands for Bayesian optimisation, *DP* for dynamic programming, *MP* for Markov processes, *GT* for game theory, which together form a comprehensive system for modelling individualised learning). Architecturally, the system implements the principle of deep integration through the interaction matrix $I = [iij]$, whose elements are determined by the function:

$$iij = \alpha \cdot cij + \beta \cdot tij + \gamma \cdot eij, \quad (14)$$

where: α – weighting factor of cognitive compatibility; β – weighting factor of temporal consistency; γ – weighting factor of interaction efficiency; *cij* – indicator of cognitive compatibility of models; *tij* – indicator of temporal consistency; *eij* – indicator of interaction efficiency.

The fundamental basis of the architecture of the integrated system is the principle of adaptive interaction of components, which ensures flexible adjustment of learning parameters following the individual characteristics of students and the dynamics of the educational process (Bertsekas, 2022). The system of intermodule interaction is described by a set of differential equations:

$$dK/dt = FK(K,L) + GK(I), \quad (15)$$

where: *K* – vector of student's knowledge; *L* – vector of educational influences; *FK, FL* – developmental functions; *GK, GL* – functions of intermodule interaction; *t* – time.

The peculiarity of the developed system is its ability to self-adapt through feedback mechanisms, which allows for optimising the learning trajectory in real-time (Wu & Noé, 2020). The Bayesian component of the system interacts with the dynamic programming component through a transfer function:

$$T(BO \rightarrow DP) = \int P(\theta|D)V(s|\theta)d\theta, \quad (16)$$

where: $P(\theta|D)$ – posterior distribution of the model parameters given the available data *D*; $V(s|\theta)$ – value function of the state *s* given the parameters θ; *D* – training data set.

The development of mechanisms for interaction between the system components is based on the principle of synergistic enhancement of learning effects. The integration of Bayesian optimisation with dynamic programming can be used to combine the benefits of probabilistic modelling of student knowledge with the optimisation of long-term learning strategies. In this case, the Bayesian component provides an accurate assessment of the current state of knowledge, and dynamic programming uses these estimates to build optimal learning trajectories (Tu *et al.*, 2024). These interactions are formalised through an extended transfer function:

$$T'(BO \rightarrow DP) = \iint P(\theta|D)V(s|\theta)K(s,s')d\theta ds, \quad (17)$$

where: $P(\theta|D)$ – posterior distribution of the model parameters; $V(s|\theta)$ – state value function; $K(s, s')$ – kernel of transition between knowledge states; *D* – training data.

An in-depth analysis of the interaction of the system components determined the need to address the temporal aspects of learning. Each transition between states of knowledge is characterised not only by the probability of success but also by the time required to learn the material. This feature is reflected in a modified transition matrix that addresses the time characteristics of the learning process:

$$P(t) = P0 \times exp(At), \quad (18)$$

where: *P0* – initial transition matrix; *A* – time evolution generator; *t* – training time.

Individual student characteristics are addressed through an adaptive system of weighting coefficients that is dynamically adjusted during the learning process. This approach allows the system to automatically determine

the most effective presentation strategies for each learner (Zhang *et al.,* 2023). Mathematically, this is expressed as an optimisation problem:

$$W^* = \text{argmin} \sum ||yi - f(xi,W)||^2 + \lambda R(W), \quad (19)$$

where: W^* – optimal set of weights; yi – target learning indicators; $f(xi, W)$ – predicted results; $R(W)$ – regularisation term; λ – regularisation coefficient.

An important aspect of an integrated system is the coordination mechanisms between different models, which ensure that all components work in a coordinated manner. The coordination mechanism is implemented through a multi-level decision-making system, where each level is responsible for a specific aspect of the learning process. Mathematically, this is described by the hierarchical structure of decision-making functions:

$$D(s) = H(F1(s), F2(s),\ldots,Fn(s)), \quad (20)$$

where: $D(s)$ – final solution of the system; $Fi(s)$ – solution of the i-th level; H – solution aggregation function; s – current state of the learning process.

The analysis of the processes of adaptation of educational content demonstrated the need to introduce mechanisms for dynamic optimisation of the complexity of the material. The system uses a composite difficulty assessment function that accounts for multiple characteristics of the learning material and individual student characteristics:

$$C(m,u) = \beta 1 c1(m) + \beta 2 c2(u) + \beta 3 c3(m,u), \quad (21)$$

where: m – characteristics of the educational material; u – characteristics of the student; ci – components of the complexity assessment; βi – weighting coefficients.

The system emphasises modelling the processes of forgetting and repeating material. Based on the extended forgetting curve, an adaptive algorithm for scheduling repetitions that addresses the individual characteristics of a student's memory has been developed (Wu & Noé, 2020). This process is described by a differential equation:

$$dR/dt = -\alpha(t)R + \beta(t)L + \gamma(t)S, \quad (22)$$

where: R – level of knowledge retention; L – intensity of learning; S – complexity of the material; $\alpha(t)$, $\beta(t)$, $\gamma(t)$ – time-dependent coefficients.

The proposed system ensures the stability of the learning process even with significant variations in input parameters. This is achieved through mechanisms of automatic correction of model parameters based on the analysis of current training results. The stability of the system is characterised by a range of Lyapunov indicators, which assess its sensitivity to perturbations in the initial conditions and external influences (Cappart *et al.*, 2021).

The developed criteria for optimising the learning process account for multiple aspects of individual learning through a comprehensive objective function that reflects both immediate learning outcomes and long-term educational goals. The mathematical formalisation of this function includes components of cognitive development, metacognitive skills and motivational factors:

$$Q(x) = \sum wiqi(x) + \sum\sum vijqi(x)qj(x), \quad (23)$$

where: $qi(x)$ – individual criteria of learning quality; wi – weighting coefficients of criteria; vij – coefficients of interaction between criteria; x – vector of parameters of the learning process.

The study of the dynamics of the educational process revealed the need to address nonlinear effects in the acquisition of knowledge. To this end, an adaptive algorithm for adjusting the complexity of learning tasks based on the analysis of the student's current state of knowledge and learning history was developed. This algorithm employed a recurrent neural architecture to predict the optimal level of difficulty:

$$h(t) = \sigma(Wh \cdot h(t-1) + Wx \cdot x(t) + b), \quad (24)$$

where: $h(t)$ – hidden state of the model; $x(t)$ – input data on the learning process; Wh, Wx – weighting matrices; b – displacement vector; σ – activation function.

The analysis of the processes of forming a deep understanding of the material demonstrated the need to introduce mechanisms for identifying and eliminating knowledge gaps. To this end, a knowledge diagnostic system based on Bayesian networks and accounting for the relationships between different concepts of the subject area has been developed:

$$P(K|E) = \prod P(Ki|Pa(Ki)) \cdot P(Ej|Ki), \quad (25)$$

where: K – vector of knowledge state variables; E – vector of observed data; $Pa(Ki)$ – set of parent nodes for Ki; $P(Ki|Pa(Ki))$ – conditional probability of knowledge of concepts; $P(Ej|Ki)$ – probability of observations with the given knowledge.

The integration mechanisms of the system ensure adaptive adjustment of the learning process through continuous analysis and optimisation of a set of parameters. At the same time, each component of the system functions as part of a single whole, providing a synergistic effect in achieving educational goals. The greatest efficiency is achieved by dynamically balancing different learning strategies when the system automatically selects the optimal ratio between the depth of study and the speed of progression through the curriculum. This approach adapts the system to the individual characteristics of each learner, considering their cognitive abilities, previous experience and current level of understanding of the material.

The developed algorithms for adapting educational content are based on the principles of deep learning and cognitive psychology. The system constantly

analyses the patterns of interaction between students and learning material, identifies the characteristics of their learning style and automatically adjusts the parameters of information presentation. The system addresses not only explicit performance indicators, but also hidden indicators of understanding, such as time spent on tasks, error patterns, and the nature of requests for help. This multi-level analysis system makes it possible to generate accurate predictions about the most effective learning strategies for each case.

An important aspect of the integrated system is the self-learning algorithm-improvement method based on the experience gained. By analysing large amounts of data on learning trajectories, the system identifies hidden patterns in the learning process and automatically optimises its parameters to improve learning efficiency. At the same time, the system maintains a high level of interpretability of its decisions, enabling teachers to analyse the logic of its operation and adjust its functioning if necessary. This system architecture strikes an optimal balance between automating the learning process and maintaining control by the teaching staff, creating an effective environment for individualised learning.

Methodological principles of efficiency assessment and practical application

The implementation of a multidimensional assessment system for mathematical models of individualised learning is based on a comprehensive analysis of educational outcomes. The system of criteria covers a wide range of indicators: from basic knowledge acquisition to higher-order metacognitive skills and the ability to apply knowledge in practice. The multidimensional assessment model proposed in educational effectiveness research involves simultaneous tracking of students' academic performance, development of their metacognitive abilities and level of learning motivation (De Maeyer *et al.*, 2010). The development of mathematical modelling skills among students is prioritised, which creates the basis for a deeper understanding of the material and the development of analytical thinking (On evaluating curricular..., 2004). The monitoring system includes technologies for continuously collecting and analysing data on individual learning trajectories, which allows for tracking the dynamics of key competencies and making the necessary adjustments to the learning process.

The effective use of mathematical models in the educational process requires the creation of a comprehensive system of teacher training. Specialised professional development programmes should cover both the theoretical foundations of mathematical modelling in education and the practical aspects of using modelling to individualise learning. Training programmes should include intensive practical sessions on the development and adaptation of learning materials, as well as methods for evaluating the effectiveness of different models in specific educational contexts (Aydogan Yenmez *et al.*, 2017). The integration of an engineering approach to mathematical modelling into the educational process opens up new opportunities for the development of students' analytical and design skills (Lyon & Magana, 2020). At the same time, it is necessary to strike a balance between technological innovation and preserving the human factor in the educational process, where the role of the teacher is transformed from a simple transmitter of knowledge to a facilitator of individual student development. Mathematical models in the educational process transform approaches to assessing students' knowledge and skills. Assessment goes beyond traditional tests and begins to include an analysis of student's ability to create personal models to solve practical problems. This approach allows for a deeper understanding of the level of learning and identifies gaps in understanding of basic concepts. The development of tasks for mathematical modelling requires teachers to have a deep understanding of both the subject area and methods of assessing student work (Dogan, 2020). At the same time, it is necessary to develop students' self-assessment and peer assessment skills, which contributes to a deeper understanding of modelling processes.

Adaptation of learning materials to the individual needs of students requires a flexible approach to the organisation of the educational process. Mathematical models can be used to create dynamic learning trajectories that are automatically adjusted based on current results and the pace of learning. Modelling in engineering education shows that an individualised approach significantly increases student motivation and engagement in the learning process (Lyon & Magana, 2020). Analysis of practical implementation experience shows the need to create a bank of tasks of different levels of complexity and thematic focus to ensure effective differentiation of learning.

The development of digital technologies expands the possibilities for introducing complex mathematical models into everyday pedagogical practice. The integration of mathematical modelling into the learning process is becoming a tool for developing students' critical thinking and analytical skills. It is necessary to ensure a gradual transition from simple models to more complex ones, which allows students to build their competencies at a natural pace (Bora & Ahmed, 2019). The technological infrastructure should support different learning formats, from individual work to group projects while ensuring continuous data collection to analyse the effectiveness of the learning process. The introduction of mathematical models into the learning process requires the creation of an adaptive learning environment. A key element of such an environment is an automated decision support system that helps teachers determine the best learning strategies for each student. The analysis of the results of modelling educational processes shows that the effectiveness of individualised learning depends not only on the accuracy of mathematical models but also on the quality of their integration into pedagogical practice. Mathematical modelling expands opportunities for the development of student's creative thinking and their ability to solve complex problems independently. At

the same time, it is necessary to strike a balance between technological innovations and maintaining a lively dialogue between teacher and student.

Evaluation of the long-term effectiveness of mathematical models requires the development of a monitoring system that tracks not only the immediate learning outcomes but also the development of students' metacognitive skills. A multidimensional approach to assessment should address students' ability to apply their knowledge in new contexts, their ability to analyse their own mistakes and adjust their learning strategies. Analysis of learning achievements through the prism of mathematical modelling reveals new aspects of understanding the processes of knowledge acquisition and skill development (De Maeyer *et al.*, 2010). This approach identifies hidden patterns in the learning process and develops more effective strategies for individualising learning. The development of a decision support system in education requires constant updating and improvement of mathematical models in line with new pedagogical research and technological capabilities. At the same time, it is important to maintain a focus on developing students' critical thinking and creativity without turning the learning process into a mechanical execution of algorithms. Mathematical modelling should become a tool for developing students' ability to learn independently and conduct research, forming the basis for their further professional development.

The introduction of mathematical models into the educational process opens new horizons for the development of education, transforming traditional approaches to teaching and assessment. The development and application of these models create a multidimensional space of possibilities where each student can follow their educational trajectory. At the same time, the key success factor is not only the technological complexity of the models but also their ability to adapt to the individual characteristics of each student, ensuring an optimal balance between challenges and support in the learning process. The analysis of the characteristics of the effectiveness of these models reveals their potential for creating a truly personalised educational environment where technology and pedagogical skills work in harmony. Summarising the results of the study of methodological foundations for assessing the effectiveness of mathematical models of individualised learning systematised the key characteristics of their practical application in the educational process (Table 2).

Table 2. Characteristics of the effectiveness of mathematical models in individualised learning

Evaluation criterion	Short-term perspective	Long-term perspective
Knowledge assessment	Testing, practical tasks, projects	Analysis of metacognitive skills, self-learning capabilities
Adaptability of learning	Adjustment of task complexity and learning pace	Formation of individual educational trajectories
Monitoring progress	Daily analysis of results, feedback	Tracking the dynamics of competence development
Teacher support	Automation of routine tasks, analytics	Professional development and methodological support
Technology integration	Basic digital tools, LMS	Integrated adaptive systems, AI support

Notes: LMS – Learning Management System; AI – Artificial Intelligence. Knowledge assessment includes both formal and informal methods of evaluation. Adaptability of learning involves automatic adjustment of the learning process parameters

Source: compiled by the author based On evaluating curricular… (2004), S. De Maeyer *et al*. (2010), A. Aydogan Yenmez *et al*. (2017), J.A. Lyon & A.J. Magana (2020)

The introduction of mathematical models of individualised learning creates a foundation for the transformation of the educational process towards greater personalisation and efficiency. The developed methodological framework for evaluating the effectiveness and recommendations for the practical application of the models provide the basis for further development and improvement of the individualised learning system, opening new opportunities for improving the quality of education through the integration of technological innovations and pedagogical experience.

Discussion

The theoretical study of the system of mathematical models of individualised learning, which integrates Bayesian optimisation, dynamic programming, Markov processes and game theory, reveals significant potential in improving the efficiency of the learning process. The analysis of mathematical models has revealed a theoretical increase in the efficiency of individualised learning compared to traditional methods, especially in groups of students with different learning styles. The study of the method of optimising partial trajectories based on recursive updating of model parameters indicates the possibility of increasing the accuracy of parameter settings and ensuring smoother adaptation to the individual pace of learning.

In the context of the dynamic adaptation of the learning process, a comparison of the theoretical results with the study by L. Tetzlaff *et al.* (2021) is noteworthy. Their concept of the dynamic structure of personalised education resonates with the considered system of Bayesian optimisation of learning parameters, but the theoretical model proposed in the current study demonstrates a deeper integration of feedback mechanisms through the analysis of a multicomponent fitness function. At the same time, the results of a study by L. Tetzlaff *et al.* on the temporal aspects of learning process adaptation confirmed the importance of the mechanism of dynamic adjustment of learning parameters through a modified transition matrix considered

in this study. The cognitive aspects of the decision-making process in the learning environment were highlighted by J.C. Peterson *et al.* (2021). Their findings on the use of large-scale experiments and machine learning to uncover human decision-making mechanisms confirmed the theoretical effectiveness of using Markov processes in modelling educational decisions. However, the model considered in the current study offers a more specific adaptation to the educational context through the study of the modified Bellman equation and the learning material complexity function, which theoretically allows for a more accurate consideration of the peculiarities of individual perception of educational content.

A significant contribution to understanding the mechanisms of fine-tuning educational parameters was made by H. Luan & C.-C. Tsai (2021). Their analysis of the application of machine learning to precision education confirms the theoretical results obtained in the current study on improving the effectiveness of individualised learning. However, the system under consideration demonstrates a more comprehensive theoretical approach through the integration of four mathematical paradigms and the study of coalition formation mechanisms to group students into groups with common goals and interests. The comparison of the theoretical results with the study by L. Zhang *et al.* (2020) is particularly noteworthy. Their conclusions regarding the need for a systematic approach to personalisation are confirmed in the model under consideration, but the proposed approach provides more specific mathematical mechanisms for implementation through a theoretical analysis of the integration of Bayesian optimisation, dynamic programming, Markov processes, and game theory. At the same time, the results of L. Zhang *et al.* on the importance of taking into account the social aspects of learning emphasised the feasibility of the game-theoretic component of the system studied in the current work. A relevant aspect of the theoretical study of mathematical models of individualised learning is the consideration of social and cognitive theories and their impact on the learning process. S. Chuang (2021) emphasised the importance of continuous adult development in the study on the application of constructivist learning theory and social learning theory. The results of this study correlate with the theoretical and game components of the system considered in the current study, especially in the context of coalition formation and the modelling of collaborative learning. However, the theoretical model presented in this study offers a more formalised approach through mathematical modelling of social interactions in the learning process.

The theoretical analysis of mathematical models of individualised learning revealed the prospects of using the modified Bellman equation to optimise learning trajectories, which incorporates a specific function of the complexity of the learning material C (s, a). This modification addresses not only the direct reward for learning actions but also the individual cognitive characteristics of each student's perception of the material. Integration of Markov processes through the transition matrix P = [pij], supplemented by conditional probabilities of transitions at different learning actions, provides an opportunity to model the stochastic nature of the educational process. In this context, the study by M.A.K. Peters (2022) on decision-making confidence in the educational environment is of particular interest. The analysis of cognitive decision-making mechanisms is consistent with the modified Bellman equation considered in the current study and its adaptation to the educational context. However, the theoretical model under consideration offers a more comprehensive mathematical framework through the integration of Markov processes and Bayesian optimisation, which formalises the stochastic nature of the learning process and the mechanisms of adaptation to individual learners.

Of considerable interest in the context of artificial intelligence in education is the study by F. Ouyang & P. Jiao (2021), which identifies three paradigms of AI application in education: instrumental (AI as a learning support tool), pedagogical (AI as an adaptive tutor), and transformational (AI as an agent of change in the educational process). Their conclusions regarding the need to integrate different approaches are reflected in the theoretical model under consideration, which combines four mathematical paradigms: Bayesian optimisation for adaptive adjustment of learning content parameters, dynamic programming for optimisation of long-term learning strategies, Markov processes for modelling the stochastic nature of the educational process, and game theory for formalising the interaction of learners. However, the system proposed in the current study demonstrates a deeper mathematical formalisation of the processes of individualisation of learning through the introduction of a comprehensive evaluation and optimisation system.

A theoretical study of mathematical models of individualised learning has demonstrated the effectiveness of optimising partial trajectories through a system of recursive updating of model parameters. The introduction of the learning material complexity function *C (s, a)* into the modified Bellman equation formalised the process of adapting learning content to individual student characteristics. This approach, complemented by the mechanisms of coalition formation and collaborative learning through the theoretical and game components, creates a theoretical basis for increasing the effectiveness of individualised learning. In the context of adaptive learning technologies, the study by H.A. Alamri *et al.* (2021) is noteworthy. Their analysis of personalisation in a blended learning environment in higher education reflects the importance of adaptive mechanisms, which resonates with the theoretical model under consideration, especially in terms of optimising partial learning paths. However, the approach presented in the current study offers a deeper mathematical formalisation of adaptation processes through the integration of Markov processes and Bayesian optimisation, which allows for more accurate modelling of the dynamics of the learning process and individual characteristics of material perception.

M.L. Bernacki *et al*. (2021) raise fundamental questions about the goals and mechanisms of personalisation in their systematic review of research on personalised learning. Their conclusions regarding the need to explicitly define the goals of personalisation resonate with the multi-level model of strategic engagement considered in the current study. However, the theoretical model presented in this paper offers specific mathematical mechanisms for achieving these goals through formalising decision-making processes and optimising learning trajectories. An important contribution to interpreting the application of machine learning to precision education was made by H. Luan & C.-C. Tsai (2021). Their analysis confirmed the theoretical results of this study on the effectiveness of using mathematical models to individualise learning. Particularly significant is their conclusion about the need to fine-tune educational parameters, which in the current study is implemented through self-adaptation mechanisms and recursive updating of model parameters.

The analysis of the theoretical results and their comparison with current research in the field of mathematical modelling of individualised learning reveals both significant advantages of the proposed approach and potential directions for further development. The combination of Bayesian optimisation, dynamic programming, Markov processes and game theory creates a powerful theoretical foundation for modelling various aspects of the learning process. At the same time, it is worth noting that further development of the theoretical foundations of mathematical modelling of individualised learning requires experimental verification of the proposed models and mechanisms. Particular attention should be devoted to the study of the practical implementation of the considered mathematical models, their computational complexity and efficiency in different educational contexts. This will not only confirm the theoretical results but also highlight areas for further improvement of mathematical models of individualised learning.

Conclusions

The theoretical analysis of existing mathematical models of individualised learning has revealed a significant potential for integrating four key mathematical paradigms: Bayesian optimisation, dynamic programming, Markov processes and game theory. A comprehensive review of these approaches has shown a significant increase in the effectiveness of individualised learning, especially for students with different learning styles. In particular, the study of the partial trajectory optimisation method, which relies on recursive updating of model parameters through the analysis of intermediate results, demonstrated the possibility of achieving much more accurate parameter settings and ensuring smoother adaptation to the individual pace of learning by each student.

The theoretical study of the modified Bellman equation with the function of the complexity of the educational material revealed the fundamental principles of optimising long-term learning strategies, revealing a wider perspective on the dynamics of adapting the complexity of the content to the individual capabilities of each student. A comprehensive analysis of the stochastic nature of the learning process through an extended transition matrix revealed deep patterns in the dynamics of knowledge acquisition and was used to formalise the processes of forgetting and repeating material through differential equations with time-dependent coefficients. The study of collaborative learning mechanisms using the game-theoretic approach allowed not only to formalise the processes of forming learning coalitions and interaction between students but also to identify the synergistic effects of group learning through nonlinear functions of interaction between participants in the educational process. The proposed system of multidimensional evaluation of mathematical models covers a wide range of indicators – from basic knowledge acquisition to the development of higher-order metacognitive skills, which provides a comprehensive understanding of the effectiveness of the educational process and its components. Particular attention is paid to the analysis of long-term learning effects through a composite prediction function that incorporates the interaction of various aspects of the learning process and their impact on the formation of sustainable knowledge and skills.

The theoretical study has shown the high adaptability of the considered models to various educational contexts and their ability to consider a wide range of individual characteristics of students. The developed methodological foundations indicate the need to form an adaptive educational environment with a powerful decision-support system to determine optimal learning strategies. The analysis of the interaction of various components of the system and mechanisms of their coordination was emphasised, which identified the key factors of success of individualised learning and ways to improve its effectiveness.

The main limitation of this study was its theoretical nature, which indicates the urgent need for experimental verification of the models under consideration in a real educational environment with different groups of students and various learning contexts. Further research should address the development of effective methods for optimising the computational complexity of the integrated system, an in-depth study of the emotional and motivational aspects of the learning process, and the creation of reliable methods for validating the effectiveness of models in different educational and cultural contexts. Particular attention should be devoted to the study of the mechanisms of interaction between different mathematical models within a single decision-making system and the development of methods for their adaptation to specific educational tasks and goals.

Acknowledgements

None.

Conflict of Interest

The author declares no conflict of interest.

References

[1] Alamri, H.A., Watson, S., & Watson, W. (2021). Learning technology models that support personalization within blended learning environments in higher education. *TechTrends*, 65, 62-78. doi: 10.1007/s11528-020-00530-3.

[2] Aydogan Yenmez, A., Erbas, A.K., Cakiroglu, E., Alacaci, C., & Cetinkaya, B. (2017). Developing teachers' models for assessing students' competence in mathematical modelling through lesson study. *International Journal of Mathematical Education in Science and Technology*, 48(6), 895-912. doi: 10.1080/0020739X.2017.1298854.

[3] Bernacki, M.L., Greene, M.J., & Lobczowski, N.G. (2021). A systematic review of research on personalized learning: Personalized by whom, to what, how, and for what purpose(s)? *Educational Psychology Review*, 33, 1675-1715. doi: 10.1007/s10648-021-09615-8.

[4] Bertsekas, D.P. (2022). *Abstract dynamic programming* (3rd ed). Belmont: Athena Scientific.

[5] Bora, A., & Ahmed, S. (2019). Mathematical modeling: An important tool for mathematics teaching. *International Journal of Research and Analytical Reviews,* 6(2), 252-256.

[6] Canco, I., Kruja, D., & Iancu, T. (2021). AHP, a reliable method for quality decision making: A case study in business. *Sustainability*, 13(24), article number 13932. doi: 10.3390/su132413932.

[7] Cappart, Q., Moisan, T., Rousseau, L., Prémont-Schwarz, I., & Cire, A.A. (2021). Combining reinforcement learning and constraint programming for combinatorial optimization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(5), 3677-3687. doi: 10.1609/aaai.v35i5.16484.

[8] Chuang, S. (2021). The applications of constructivist learning theory and social learning theory on adult continuous development. *Performance Improvement*, 60(3), 6-14. doi: 10.1002/pfi.21963.

[9] De Maeyer, S., van den Bergh, H., Rymenans, R., Van Petegem, P., & Rijlaarsdam, G. (2010). Effectiveness criteria in school effectiveness studies: Further research on the choice for a multivariate model. *Educational Research Review*, 5(1), 81-96. doi: 10.1016/j.edurev.2009.09.001.

[10] Dogan, M.F. (2020). Evaluating pre-service teachers' design of mathematical modelling tasks. *International Journal of Innovation in Science and Mathematics Education*, 28(1), 44-59. doi: 10.30722/IJISME.28.01.004.

[11] Eglington, L.G., & Pavlik, P.I. (2023). How to optimize student learning using student models that adapt rapidly to individual differences. *International Journal of Artificial Intelligence in Education*, 33, 497-518. doi: 10.1007/s40593-022-00296-0.

[12] Jiang, P., & Wang, X. (2020). Preference cognitive diagnosis for student performance prediction. *IEEE Access*, 8, 219775-219787. doi: 10.1109/ACCESS.2020.3042775.

[13] Luan, H., & Tsai, C.-C. (2021). A review of using machine learning approaches for precision education. *Educational Technology & Society*, 24(1), 250-266.

[14] Lyon, J.A., & Magana, A.J. (2020). A review of mathematical modeling in engineering education. *International Journal of Engineering Education*, 36(1), 101-116.

[15] Maghsudi, S., Lan, A., Xu, J., & van Der Schaar, M. (2021). Personalized education in the artificial intelligence era: What to expect next. *IEEE Signal Processing Magazine*, 38(3), 37-50. doi: 10.1109/MSP.2021.3055032.

[16] Minn, S. (2022). AI-assisted knowledge assessment techniques for adaptive learning environments. *Computers and Education: Artificial Intelligence*, 3, article number 100050. doi: 10.1016/j.caeai.2022.100050.

[17] On evaluating curricular effectiveness: Judging the quality of K-12 mathematics evaluations. (2004). Washington: National Academies Press. doi: 10.17226/11025.

[18] Ouyang, F., & Jiao, P. (2021). Artificial intelligence in education: The three paradigms. *Computers and Education: Artificial Intelligence*, 2, article number 100020. doi: 10.1016/j.caeai.2021.100020.

[19] Peters, M.A.K. (2022). Confidence in decision-making. *Oxford Research Encyclopedia of Neuroscience*. doi: 10.1093/acrefore/9780190264086.013.371.

[20] Peterson, J.C., Bourgin, D.D., Agrawal, M., Reichman, D., & Griffiths, T.L. (2021). Using large-scale experiments and machine learning to discover theories of human decision-making. *Science*, 372(6547), 1209-1214. doi: 10.1126/science.abe2629.

[21] Rodemann, J., & Augustin, T. (2024). Imprecise Bayesian optimization. *Knowledge-Based Systems*, 300, article number 112186. doi: 10.1016/j.knosys.2024.112186.

[22] Taherdoost, H., & Madanchian, M. (2023). Multi-criteria decision making (MCDM) methods and concepts. *Encyclopedia*, 3(1), 77-87. doi: 10.3390/encyclopedia3010006.

[23] Tetzlaff, L., Schmiedek, F., & Brod, G. (2021). Developing personalized education: A dynamic framework. *Educational Psychology Review*, 33, 863-882. doi: 10.1007/s10648-020-09570-w.

[24] Tu, S., Frostig, R., & Soltanolkotabi, M. (2024). Learning from many trajectories. *Journal of Machine Learning Research*, 25, 1-109.

[25] Wang, X., Jin, Y., Schmitt, S., & Olhofer, M. (2023). Recent advances in Bayesian optimization. *ACM Computing Surveys*, 55(13s), article number 287. doi: 10.1145/3582078.

[26] Wu, H., & Noé, F. (2020). Variational approach for learning Markov processes from time series data. *Journal of Nonlinear Science*, 30, 23-66. doi: 10.1007/s00332-019-09567-y.
[27] Xing, W., Li, C., Chen, G., Huang, X., Chao, J., Massicotte, J., & Xie, C. (2021). Automatic assessment of students' engineering design performance using a Bayesian network model. *Journal of Educational Computing Research*, 59(2), 230-256. doi: 10.1177/0735633120960422.
[28] Zhang, D., Chen, R.T., Liu, C.H., Courville, A., & Bengio, Y. (2023). Diffusion generative flow samplers: Improving learning signals through partial trajectory optimization. *ArXiv*. doi: 10.48550/arXiv.2310.02679.
[29] Zhang, L., Basham, J.D., & Yang, S. (2020). Understanding the implementation of personalized learning: A research synthesis. *Educational Research Review*, 31, article number 100339. doi: 10.1016/j.edurev.2020.100339.

Математичні моделі індивідуалізованого навчання, побудовані на теорії прийняття рішень

Іван Вовчок
Аспірант
Ужгородський національний університет
88000, пл. Народна, 3, м. Ужгород, Україна
https://orcid.org/0000-0001-8603-7899

Анотація. У дослідженні здійснено теоретичне обґрунтування та розроблення системи математичних моделей для індивідуалізації освітнього процесу на основі комплексної інтеграції методів теорії прийняття рішень. Розроблена система математичних моделей базується на метамоделі, що поєднує чотири математичні парадигми через матрицю взаємодії, елементи якої визначаються функцією когнітивної сумісності, часової узгодженості та ефективності взаємодії. Впровадження методу оптимізації часткових траєкторій, що спирається на рекурсивне оновлення параметрів моделі через аналіз проміжних результатів, дало змогу досягти точнішого налаштування параметрів та забезпечити плавну адаптацію до індивідуального темпу засвоєння матеріалу. Розроблена модифікація рівняння Беллмана з функцією складності навчального матеріалу допомогла формалізувати процес оптимізації довгострокових навчальних стратегій через врахування індивідуальних когнітивних особливостей. Аналіз стохастичної природи навчального процесу через розширену матрицю переходів дозволив математично описати процеси забування й повторення матеріалу за допомогою системи диференційних рівнянь з часозалежними коефіцієнтами, що враховують інтенсивність навчання та індивідуальні особливості пам'яті. Дослідження механізмів колаборативного навчання за допомогою теоретико-ігрового підходу виявило синергетичні ефекти групового навчання через нелінійні функції взаємодії учасників освітнього процесу та дозволило розробити методи формування оптимальних навчальних груп з урахуванням індивідуальних цілей. Запропонована система багатовимірного оцінювання, що реалізується через композитну цільову функцію, охоплює широкий спектр показників від базового засвоєння знань до розвитку метакогнітивних навичок вищого порядку, включаючи когнітивні, метакогнітивні та мотиваційні компоненти, що забезпечує надійний інструментарій для оцінки стійкості навчальних траєкторій та визначення рівня адаптивності системи до індивідуальних особливостей учнів

Ключові слова: адаптивні освітні системи; байєсівська оптимізація; функція Беллмана; марковські процеси; теоретико-ігровий підхід; когнітивні траєкторії

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

Information Technologies and Computer Engineering **Vol. 21, No. 3. 2024**

Article's History: Received: 20.09.2024 Revised: 22.11.2024 Accepted: 26.12.2024

UDC 004.94 10.63341/vitce/3.2024.108

Analysis of the decision-making algorithm efficiency in complex game environments on the example of Pac-Man

Artem Novikov*

Postgraduate Student
V.N. Karazin Kharkiv National University
61022, 4 Svobody Sq., Kharkiv, Ukraine
https://orcid.org/0009-0004-5914-7098

Volodymyr Yanovskyi

Doctor of Physical and Mathematical Sciences, Professor
"Institute for Single Crystals" of National Academy of Sciences of Ukraine
61072, 60 Nauky Ave., Kharkiv, Ukraine
https://orcid.org/0000-0003-0461-749X

Abstract. Game simulations such as Pac-Man are substantial for testing decision-making algorithms in conditions that mimic real-life scenarios. This creates new opportunities for the development of autonomous systems that can adapt to changing environmental conditions and interact with other agents. The study aimed to compare Expectimax, Monte Carlo Tree Search, and Alpha-Beta Pruning algorithms in the changed conditions of the Pac-Man game to determine the most efficient approach to decision-making in complex environments. For this purpose, simulation modelling was used to evaluate the effectiveness of agents in various game mazes that differ in complexity. The study measured such indicators as the number of points, game time, and percentage of winnings, which were used to assess the effectiveness of algorithms in different situations. The analysis of the experiments determined that the Monte Carlo algorithm is the most effective among the tested methods for solving less complex mazes, confirming quickly optimal path search in simple conditions. The Alpha-Beta Pruning algorithm demonstrated less efficiency, which indicates the need to optimise it for more complex environments. Expectimax demonstrated significantly lower performance, which indicates its limited suitability for complex game mazes. The study demonstrated that increasing the complexity of the mazes significantly reduces the performance of all algorithms, especially with more obstacles, highlighting the importance of developing more robust methods for highly complex environments. Optimising the Monte Carlo and Alpha-Beta Pruning algorithms for complex environments can significantly improve their performance and make them effective for real-world applications in navigation and control of moving devices. The results of this study can be used to develop efficient navigation algorithms for autonomous vehicles, drones and other robotic systems where adaptation to changes in complex environments is critical

Keywords: simulation environments; Expectimax; MCTS; Alpha-Beta Pruning; autonomous navigation

Introduction

The research relevance is determined by the growing need for artificial intelligence (AI) algorithms capable of fast and efficient decision-making in complex, dynamic environments. Such algorithms are critical in industries related to autonomous navigation, drone control, logistics process optimisation, and other areas requiring adaptation to changing conditions and interaction with other agents. They must provide stable performance even when resources are limited. Of particular importance are studies of the adaptability of algorithms in simulation environments that model real-world conditions.

Different approaches to the development of AI algorithms are actively discussed in the scientific literature. For instance, the A.W.R. Ramadhan & D. Udjulawa (2020)

Suggested Citation:
Novikov, A., & Yanovskyi, V. (2024). Analysis of the decision-making algorithm efficiency in complex game environments on the example of Pac-Man. *Information Technologies and Computer Engineering*, 21(3), 108-118. doi: 10.63341/vitce/3.2024.108

*Corresponding author

compared the performance of A* and Dijkstra algorithms in the Pac-Man game, demonstrating the advantage of A* in finding the shortest path due to better cost optimisation and faster execution. The study also highlighted the importance of methodical prototyping to evaluate the performance of algorithms in dynamic environments. N. Salem *et al.* (2024) confirmed the advantages of A* by showing its effectiveness in path planning and cost minimisation tasks in the context of a game. In particular, the paper addressed the role of cost optimisation, and the number of nodes expanded during the search.

Innovations in the creation of environments for testing algorithms are emphasised. K.M. Cheng *et al.* (2024) investigated maze generation algorithms for Pac-Man, in particular, an improved version of the Sidewinder algorithm that significantly improves the creation of dynamic environments tailored to the specifics of the game. The study offers a new approach to the adaptation of traditional algorithms, providing more realistic conditions for evaluating AI strategies.

Monte Carlo Tree Search (MCTS) is one of the most promising algorithms for solving problems in simulation environments. M. Świechowski *et al.* (2023) emphasised the need to adapt MCTS to complex environments, stressing the importance of problem-based modification integration. In particular, the paper explored hybrid approaches for complex games with high branching and real-world applications in transport. W. Li *et al.* (2023) proposed a self-learning version of MCTS (SL-MCTS) that provides faster task completion and improved efficiency in path planning by using a neural network to optimise the search process. This modification significantly improves the algorithm's performance in time-constrained environments. I.F. Lövétei *et al.* (2021) demonstrated the feasibility of MCTS in real-time railway traffic control, where the algorithm proved capable of accommodating multi-factor constraints and producing optimal solutions in the shortest possible time.

Alpha-Beta Pruning was also studied by researchers. For instance, D. Permatasari *et al.* (2022) explored its application to improve NPC performance in the game Triple Triad, proving that this approach significantly increases the chances of winning in multiplayer environments. This confirms the decision-making efficiency of the algorithm in complex environments with many strategic factors. M. Mudda (2022) noted that alpha-beta pruning significantly optimises the calculation time in problems with many possible solutions, especially in two-player games. N. Sharma (2022) emphasised the limitations of Minimax and Alpha-Beta Pruning in cases where opposing agents do not act optimally, which requires adapting these algorithms to environments with uncertainty.

Equally important is the analysis of reinforcement learning models. Y. Cheng *et al.* (2020) proposed an MCTS-MTF algorithm to efficiently manage mixed traffic flows while considering speed and performance. This approach can be adapted to control agents in complex environments, demonstrating the prospects of using MCTS in real-time tasks. B. Wang *et al.* (2020) analysed how convolutional neural networks make decisions in conditions of high complexity using Ms. Pac-Man as an example. The study emphasised the importance of optimising solutions for high-reward tasks, which can be used to create complex simulation models.

However, despite significant progress in the research of these approaches, the adaptation of algorithms to environments with high uncertainty and complexity remains insufficiently studied. Previous studies demonstrate the benefits and comparisons in the Pac-Man game environment but focus on simple environments with few obstacles and limited interaction with other agents. At the same time, such environments do not fully reflect real-world scenarios that include high object density, complex navigation, and the need for rapid decision-making in changing environments.

The modified version of the Pac-Man game proposed in the current study is used as a simulation that models the tasks of unmanned aerial vehicles (UAVs) in two-dimensional environments. Mazes with varying wall densities and moving enemies simulate complex real-world conditions such as navigating through obstacles, optimising a route to multiple targets, and avoiding collisions with other agents. The absence of comparative studies of Expectimax, MCTS, and Alpha-Beta Pruning algorithms in such environments leaves open questions about their effectiveness in tasks that are close to real-world scenarios.

Thus, the present study seeked to fill these gaps by analysing the effectiveness of Expectimax, MCTS, and Alpha-Beta Pruning algorithms in complex environments that simulate real UAV missions. The study aimed to evaluate and compare their performance in achieving the main and additional goals, as well as to analyse their ability to adapt to variable factors. The use of a modified version of the Pac-Man game as a simulation describes in greater detail how these algorithms behave in conditions of increased complexity, which is crucial for their potential application in autonomous systems such as UAV control or other related tasks in dynamic environments.

Materials and Methods

The research material was the Pac-Man video game developed by Namco and released in 1980, modified to test algorithms in changed conditions. The experiment was conducted in the environment of this game to evaluate the effectiveness of three decision-making algorithms: Expectimax, Monte Carlo Tree Search (MCTS), and Alpha-Beta Pruning. Each of these algorithms controlled the behaviour of the main agent (Pac-Man), which had to perform the tasks of finding capsules, avoiding ghosts, and destroying ghosts during their vulnerability. The fourth agent, Random Agent, acted as a baseline for comparing results.

The study was conducted in three levels of complexity of a 50-by-50-cell maze, which differed in the number of walls (10%, 15%, and 20%) and the number of obstacles represented by ghosts. The purpose of introducing different levels of difficulty was to test the adaptability and

efficiency of each algorithm in conditions of different space constraints for manoeuvring. This approach was used to assess how different algorithms cope with the task in conditions where the restriction of freedom of movement is complicated by other factors.

Figure 1 shows the main and additional tasks of Pac-Man. The primary objective of Pac-Man was to collect all strategically important capsules on the map. Additional tasks:

- avoiding ghosts: ghosts act as active opponents chasing Pac-Man, trying to stop the actor;
- destroying frightened ghosts: after collecting the capsule, the ghosts become vulnerable for 10 turns, and Pac-Man can destroy them for extra points.

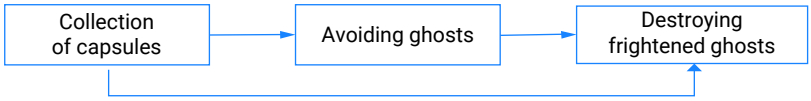

Figure 1. Main and additional tasks of Pac-Man
Source: compiled by the authors based on modified conditions of the Pac-man game

Each of the four agents played 100 games at each difficulty level, and the average score, game time, and winning percentage were recorded during each game. This was used to evaluate the effectiveness of each agent in fulfilling the main and additional goals of the game. Types of agents are shown in Figure 2.

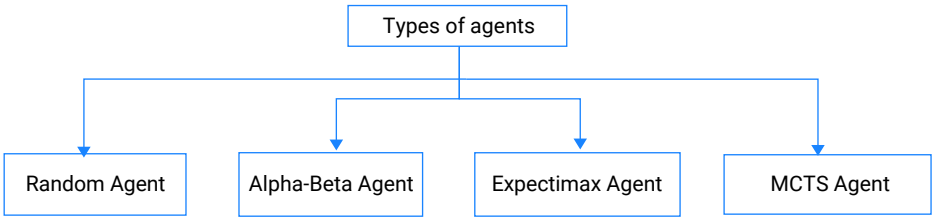

Figure 2. Types of agents used to compare their effectiveness in different Pac-Man game situations
Source: compiled by the authors based on the prepared modelling environment

Random Agent was used as a baseline to compare the results. This agent had no strategy and randomly chose actions during each turn. Its performance was expected to be the worst, as it was unable to respond effectively to game challenges such as ghosts or capsules. The Random Agent was used to evaluate how other agents with more complex algorithms improved the results compared to random decisions. This created a point of contrast to compare different approaches to decision-making in complex environments.

Other agents, **Expectimax**, **Alpha-Beta**, and **MCTS**, used different strategies to achieve the main and additional goals in the mazes. Comparing these agents with **Random Agent** was used to evaluate:

- **Expectimax Agent**: how this agent made decisions based on probable events, given the random behaviour of the ghosts;
- **Alpha-Beta Agent**: How this agent used an alpha-beta cut-off algorithm to reduce the search and reach optimal decisions in complex environments;
- **MCTS Agent**: how this agent used simulations to evaluate possible actions and choose the best strategy, adapting to increasing complexity.

A comparison of the results of these agents was used to determine which of the approaches is most effective in complex mazes with different levels of difficulty and limited resources.

Results and Discussion

For the experiments, it is necessary to examine in detail the four types of agents used to compare their performance in different Pac-Man game situations. Each of the four agents is characterised by its approach to decision-making, which ranges from simple random selection to complex algorithms based on simulations and predictions.

Random Agent is the simplest agent that randomly chooses its actions at each step of the game. It does not account for the positions of ghosts or the location of capsules or food. This agent has no algorithms for calculating possible outcomes and does not analyse the state of the game. Its actions depend entirely on a random choice from the available options at each turn. Despite its low efficiency, Random Agent is an essential benchmark for comparison, as it can be used to determine how much more complex algorithms improve performance. It is used as a baseline against which to compare the performance of other agents that use analytical decision-making approaches.

Alpha-Beta Agent is an agent that uses an alpha-beta cut-off algorithm to optimise the search for the best action. The algorithm is based on a minimax approach, where Pac-Man tries to maximise winnings, and ghosts minimise them by acting as opponents. Alpha-beta cutoff reduces the number of possible scenarios by cutting off those that do not affect the final result. This improves the performance of the agent, especially in situations where the

number of options is large, but not all of them are essential to the outcome. The agent assesses the state of the game, including the distance to ghosts and capsules, and makes decisions based on these indicators. The main advantage of Alpha-Beta Agent is its ability to significantly reduce the number of required calculations, which increases its efficiency in medium-complexity environments.

Expectimax Agent simulates the actions of Pac-Man and ghosts based on random events. This agent treats the actions of the ghosts as random and unpredictable. The Expectimax algorithm can be used to estimate the possible outcomes of each action, including the element of randomness in the ghosts' behaviour. This makes Expectimax Agent more flexible in environments where the behaviour of adversaries is changing or difficult to predict. This approach works well in complex environments with a high level of uncertainty, but in more structured environments where ghost behaviour can be predicted, Expectimax may be less effective. The agent attempts to exploit situations where ghosts are in a state of vulnerability after collecting capsules but may not be effective in avoiding ghosts in situations where their behaviour is less random.

MCTS Agent (Monte Carlo Tree Search) is an agent that uses a simulation approach to evaluate possible actions. MCTS performs numerous simulations of each Pac-Man action, attempting to predict the consequences of several moves ahead. Once the simulations are complete, the agent selects the action that has the highest potential for successful completion of the game. In complex mazes, MCTS can use modified heuristics to focus on important elements, such as capsules or ghosts, and plan actions more efficiently. The heuristics include additional criteria such as distance to the nearest capsule or proximity to ghosts, enabling more informed decision-making. This agent demonstrates the highest performance in complex environments, but its effectiveness is largely dependent on the number of simulations. The more simulations the MCTS Agent can run, the more accurate its predictions of the outcomes of actions will be, but increasing the number of simulations also requires more resources and computing time. Depending on the environment, as the number of simulations increases, the results may become worse starting from a certain threshold.

Ghosts in the game are primary opponents for Pac-Man, significantly complicating the task. Their function is not limited to chasing the main character but also includes creating a complex dynamic threat that requires Pac-Man to constantly adapt to changing conditions and make quick decisions. With the help of algorithms that control their behaviour, ghosts add variety and unpredictability to the game, creating situations where any inattention on the part of the player can lead to defeat.

The ghosts are guided by classic algorithms, which renders the behaviour predictable to a certain extent, but their interaction with each other and with Pac-Man complicates the situation. Each ghost has a unique algorithm, which provides diversity in their behaviour and makes them much more challenging as a collective.

The main strategy of the ghosts is to either directly chase Pac-Man or block possible escape routes, depending on the situation on the map. One ghost may move in the direction of the main character, trying to catch Pac-Man, while others may act in different scenarios, for example, one ghost may block paths, another may try to block exits or move to certain points to limit the primary actor's manoeuvrability. Each ghost has a unique behavioural algorithm that includes both active pursuit and strategic route blocking, substantially increasing efficiency in cooperative interaction. Thus, although each ghost acts according to different rules of the game, their collective influence creates an unpredictable and dynamic situation in which Pac-Man is forced to constantly adjust strategies to avoid danger. The types of ghosts are shown in Figure 3.

Figure 3. Ghosts in Pac-Man
Source: compiled by the authors based on Evillasio2 (2022)

Ghosts are guided by the following algorithms:

1. Blinky (Red Ghost): Blinky is the most aggressive ghost and always chases Pac-Man, trying to get as close as possible. Its algorithm involves moving directly to Pac-Man's current location. As Pac-Man collects capsules, Blinky can even accelerate, thus even more threatening. Constant pressure forces Pac-Man to make quick decisions on pathing to avoid being chased.

2. Pinky (Pink Ghost): Pinky acts more strategically, trying to block Pac-Man's path rather than engaging in direct pursuit. The Pinky algorithm involves moving to a point slightly ahead of Pac-Man's current direction. This forces Pac-Man to change route deliberately, as Pinky can easily block the way unless the player takes a different direction.

3. Inky (Blue Ghost): Inky has a more complex behaviour that depends on both Pac-Man's current location and Blinky's position. Its algorithm calculates the direction of movement, which depends on the relative position of the two objects. Because of this, Inky can be both unpredictable and dangerous, especially when Pac-Man is between him and Blinky.

4. Clyde (Orange Ghost): Clyde shows behaviour that changes depending on proximity to Pac-Man. When

Pac-Man is at a great distance, Clyde chases the player, similarly to Blinky. However, as Pac-Man approaches, Clyde suddenly changes direction and moves away. This unpredictable behaviour makes it difficult to predict and complicates Pac-Man route planning.

After Pac-Man eats the capsule, all the ghosts become "spooked" for 10 steps, according to the developed modification of the game. In this state, they start avoiding Pac-Man, while the latter can chase and destroy them for extra points. This phase of the game changes the dynamics, giving Pac-Man a temporary advantage, but requires quick decisions as the frightened ghosts soon return to their usual behaviour. Thus, each ghost in the Pac-Man game has a unique behavioural algorithm that affects the overall difficulty of the game. Interacting with ghosts requires Pac-Man to actively adapt to the situation on the map, which makes them an important element for evaluating the effectiveness of the algorithms used by Pac-Man agents.

To conduct the experiments, three types of mazes were created, which differ in the level of complexity determined by the number of walls. Each maze presented a different set of obstacles for Pac-Man and significantly affects the game strategy, as it makes it more difficult or easier to manoeuvre between the capsules and ghosts. The mazes also reflected real-life scenarios of confined space conditions that mimic situations faced by autonomous agents in real-world environments.

All mazes had the same dimension, with a total size of 50 by 50 playing cells. In each maze, regardless of the complexity, 4 capsules served as the main goals for Pac-Man. The capsules were located in different parts of the maze, which required Pac-Man to carefully plan path to collect them. In addition to the capsules, there were 4 ghosts in the maze that act as active opponents. The ghosts started the game in different parts of the maze, which was a simulation of real-life conditions and aims to balance the game's difficulty directly.

Description of labyrinths:

1. A maze with 10% walls (XLarge maze 1 is shown in Figure 4) – the simplest level of complexity. In this labyrinth, the walls occupy only 10% of the total area, which gives Pac-Man more freedom of movement. The minimum number of obstacles allows Pac-Man to dodge ghosts more effectively and reach the capsules faster. However, simpler mazes also give the ghosts more opportunities for direct pursuit, as Pac-Man cannot easily hide behind walls.

2. A maze with 15% walls (XLarge maze 2 is shown in Figure 5) is a medium difficulty level. As the number of walls increases to 15%, movement becomes more difficult, requiring Pac-Man to make more complex route decisions. More obstacles create more options to avoid the ghosts but also reduce the number of available paths to the capsules. In such conditions, Pac-Man must use the limited routes more efficiently, balancing between protection from ghosts and finding paths to the capsules.

3. The maze with 20% walls (XLarge maze 3 shown in Figure 6) is the most difficult level. In this maze, the walls occupy 20% of the total area, making Pac-Man's movement much more restricted. Pac-Man faces more obstacles on the way to the capsules, which requires more precise planning for each step. Ghosts in this environment become even more dangerous, as Pac-Man has fewer opportunities to evade pursuit, and the choice of safe routes is reduced.

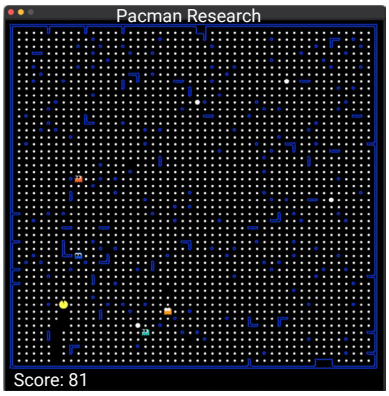

Figure 4. XLarge maze 1 in-game
Source: compiled by the authors in modelling and experimentation application

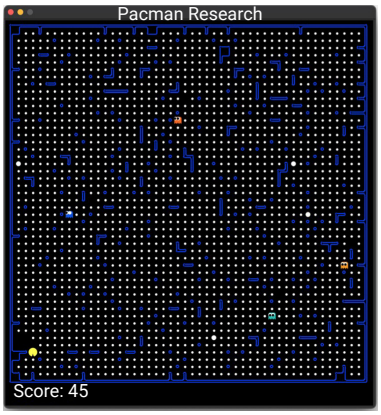

Figure 5. XLarge maze 2 in-game
Source: compiled by the authors in modelling and experimentation application

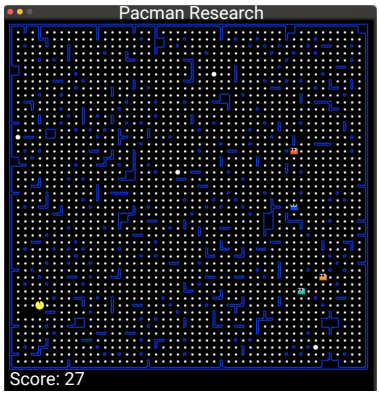

Figure 6. XLarge maze 3 in-game
Source: compiled by the authors in modelling and experimentation application

The experiments showed significant differences in the effectiveness of different agents when the complexity of the mazes changed. Each agent demonstrated unique behavioural properties, prompting certain assumptions regarding strategies and capabilities. An important criterion for evaluating agents was not only their ability to perform the main task of collecting all capsules but also to achieve secondary goals, such as destroying ghosts or collecting food. This was reflected in the number of points because even if Pac-Man did not achieve all the main goals, a high score indicated the agent's effectiveness in achieving additional goals.

Results of testing the effectiveness of agents in mazes

Mazes with fewer walls (10%) proved to be less challenging for the agents (Table 1). A sufficiently open space provided ample opportunities for agents to manoeuvre, while at the same time creating conditions for more aggressive pursuit by ghosts.

Table 1. Results for mazes with 10% of the walls

Agent	Average score	Game time	Percentage of winnings
Random Agent	120.17	29.14	0%
Alpha-Beta Agent	3,174.03	192.96	34%
Expectimax Agent	2,219.85	278.77	8%
MCTS Agent (50 simulations)	3,906.96	484.85	38%
MCTS Agent (101 simulations)	3,383.16	354.64	35%

Source: compiled by the author based on modelling and experiments in the study

The best results were achieved by **MCTS Agent**, which scored an average of 3,906.96 points and had a 38%-win rate with 50 simulations. Increasing the number of simulations to 101 slightly reduced the result (3,383.16 points and 35% of wins). This may indicate that in less complex environments, increasing the number of simulations is not always beneficial, as it can overload the path-making, which becomes less efficient due to excessive exploration of possible options.

Alpha-Beta Agent achieved a consistent result, scoring 3,174.03 points with a 34%-win rate, which indicates the ability of this agent to effectively reduce the search space using alpha-beta cutoff. This enabled the agent to achieve both primary and secondary goals in less complex environments. However, in more complex mazes, the Alpha-Beta Agent demonstrated limited adaptability due to its dependence on accurate estimates of search depth.

Expectimax Agent scored an average of 2,219.85 points and won only 8% of the games. This indicates its effectiveness in predictable scenarios, but its dependence on random events renders Expectimax less efficient than other agents. Nevertheless, its ability to account for all possible states gives it an advantage in environments with many available moves.

Random Agent performed the worst, scoring an average of 120.17 points and not winning a single game. This confirms that the lack of strategy and random choice of actions make it highly inefficient. This behaviour caused it to frequently fall into ghost traps due to rash movements.

In mazes with a medium number of walls (15%), all agents showed a decrease in performance (Table 2). In this maze, the agents were forced to use more complex strategies to evade the ghosts and reach the capsules.

Table 2. Results for mazes with 15% of the walls

Agent	Average score	Game time	Percentage of winnings
Random Agent	110.07	31.79	0%
Alpha-Beta Agent	2,602.32	698.16	12%
Expectimax Agent	2,183.97	254.42	6%
MCTS Agent (50 simulations)	3,361.54	290.2	19%
MCTS Agent (101 simulations)	3,215.14	599.82	25%

Source: compiled by the author based on modelling and experiments in the study

MCTS Agent continued to lead the leaderboard, scoring 3,361.54 points with 50 simulations and winning 19% of games. Increasing the number of simulations to 101 allowed the agent to increase its winning percentage to 25%. This indicates that in more complex environments, additional simulations improve adaptation. However, the agent requires significant computing resources to achieve high accuracy.

Alpha-Beta Agent achieved an average score of 2,602.32 points with a 12%-win rate, indicating that this agent loses efficiency as the maze complexity increases. Its ability to reduce the search space started to perform worse in conditions with more obstacles. In addition, the agent showed a limited ability to quickly adapt to the changed behaviour of the ghosts.

Expectimax Agent demonstrated even weaker results, scoring 2,183.97 points and winning only 6% of the games. This indicates the limitations of the Expectimax algorithm in more complex environments where the random actions of ghosts become less predictable. The agent often made risky decisions, which led to a loss of points.

Random Agent remained in last place, scoring 110.07 points and not winning a single game. This once again emphasises that randomly choosing actions in difficult conditions is ineffective. In more complex mazes, the agent was even more prone to falling into traps.

Mazes with the largest number of walls (20%) were the most difficult for all agents (Table 3). The agents encountered significant difficulties due to the limited space for evading ghosts and the difficulty of finding optimal routes.

Table 3. Results for mazes with 20% of the walls

Agent	Average score	Game time	Percentage of winnings
Random Agent	183.90	80.48	0%
Alpha-Beta Agent	1,672.59	200.09	2%
Expectimax Agent	1,685.06	253.68	0%
MCTS Agent (50 simulations)	2,042.03	460.64	1%
MCTS Agent (101 simulations)	2,221.03	298.61	2%

Source: compiled by the author based on modelling and experiments in the study

MCTS Agent with 50 simulations performed better than the other agents, winning one game out of 100 and scoring an average of 2,042.03 points. Increasing the number of simulations to 101 resulted in better performance, with the average score rising to 2,221.03 points and the win rate also increasing to 2%. This shows that even in the most challenging conditions, MCTS can benefit from additional simulations, improving its solutions in conditions of increased complexity.

Alpha-Beta Agent achieved an average score of 1,672.59 points with a 2%-win rate, demonstrating a significant decrease in performance in complex mazes with many obstacles. This indicates the limitations of the alpha-beta cutoff strategy, which performs well in simpler environments but struggles in environments with many obstacles. In a maze with 20% walls, this agent often had difficulty predicting effective routes due to the difficulty of dodging ghosts. The high level of obstacles required much more precision in calculations, which the agent failed to achieve.

Expectimax Agent scored an average of 1,685.06 points and did not win a single game, which confirms its limited effectiveness in conditions of high complexity and many random events. In the most complex maze, this agent was unable to adapt its decisions to dynamic changes, which reduced its effectiveness. Many obstacles and limited space significantly reduced Pac-Man's chances of avoiding traps, which led to quick losses.

Random Agent again demonstrated the worst results, scoring only 183.90 points and not winning a single game, which emphasises its complete inefficiency in such conditions. In a maze with 20% walls, the random actions of this agent led to frequent collisions with ghosts, as it could not choose safe routes. Such an environment fully exposed the weakness of this approach, as even minimal planning was absent.

The collected experimental results revealed several interesting aspects of agent performance in different maze complexity conditions. The best results in environments with many walls (20%) were shown by **MCTS Agent** with 101 simulations, which resulted in it winning 2 games out of 100 and scoring 2,221.03 points. This highlights the fact that increasing the number of simulations can improve the performance of an agent in high-complexity environments. However, increasing the number of simulations beyond 101 resulted in significant performance degradation due to a significant increase in the time required to calculate each move. This is especially relevant for complex environments, such as large mazes, where the speed of decision-making is critical. Since the limit on the number of simulations requires more productive computing resources, the current version of MCTS is not optimal for use on power-limited devices such as drones or autonomous robots without additional algorithm modifications.

The other agents also demonstrated high computational resource requirements when calculating the next step, which, as in the case of MCTS, shows a significant increase in the time required to calculate each move. This is especially notable for the Alpha-Beta Agent, which loses efficiency as the maze complexity increases. Therefore, although the Alpha-Beta Pruning algorithm works well in medium-complexity environments, its application in complex environments requires adaptation.

Alpha-Beta Pruning is an improvement of the Minimax algorithm that reduces the number of tested options by skipping some branches of the decision tree that cannot affect the result. This can significantly speed up the search process, but as the complexity of the maze increases (for example, with more possible states or a more complex map structure), the number of required checks increases, which leads to a decrease in the efficiency of the algorithm. Thus, to use this algorithm in complex environments, additional optimisations are usually required or a transition to other methods more adapted to complex conditions.

Lastly, the **Random Agent**, although it had minimal computational requirements, demonstrated its complete inefficiency in all experimental conditions. This agent chooses its actions randomly, without analysing the current state of the game, which can be used only for comparison with other, more complex agents. The low

performance of the Random Agent is used as a benchmark to assess the level of complexity of both the game environment and the tasks faced by other agents. Therefore, Random Agent emphasised the multifactorial nature of the game environment and the scale of the challenges faced by agents in the process of making decisions and implementing strategies. The performance of other agents, such as MCTS (Monte Carlo Tree Search) and Alpha-Beta Pruning, contrasted significantly with the results of Random Agent and demonstrates significantly better performance in the environments shown.

The experimental results showed that MCTS Agent is highly efficient compared to other algorithms, especially in complex environments such as mazes with 20% walls. This trend correlates with the findings of H. Maddipati *et al.* (2020), which emphasised the flexibility of MCTS in dynamic environments. MCTS strikes a balance between research and operation, allowing it to adapt to changing conditions. Similar results were also obtained by N. Pepels *et al.* (2014), demonstrating that MCTS can achieve high performance even in real-time by applying variable tree depth strategies and search tree reuse.

In the present study, MCTS performed best under conditions of increased complexity due to its ability to run numerous simulations to analyse possible options. This was supported by the findings of S. Samothrakis *et al.* (2011), who emphasised that MCTS provides better adaptation in games without a clear end state, such as Ms. Pac-Man. However, as noted by D. Busatto-Gaston *et al.* (2020), the performance of MCTS can be further improved by integrating symbolic cues, which helps to optimise action search and selection.

At the same time, Expectimax Agent demonstrated significantly lower performance compared to MCTS, especially in complex mazes. This is partially consistent with the findings of P.S. Shevtekar *et al.* (2022), who described the limitations of Minimax, particularly its sensitivity to the depth of the search tree. Since Expectimax is a modification of Minimax that accounts for random events, its performance in dynamic environments is significantly reduced. As noted by Y. Zou (2021), Minimax provides optimal results in deterministic environments, while Expectimax adapts better to environments with random events. However, in environments of higher complexity, such as mazes with 20% walls, Expectimax's dependence on randomness and limited ability to predict actions make it less effective, and its performance decreases significantly with increasing maze complexity.

Alpha-Beta Agent, while demonstrating consistent results in medium-complexity environments, loses effectiveness in complex mazes. As noted by S.P. Singhal & M. Sridevi (2019), Alpha-Beta Pruning is optimal for reducing computational costs in deterministic environments, but its performance decreases in large search spaces. Comparison with data by P. Mishra *et al.* (2018) confirmed that search optimisation is crucial for the effective use of Alpha-Beta in complex environments. In addition, the use of parallel architectures proposed by P.S. Shevtekar *et al.* (2022) can reduce these limitations.

The experiments also confirmed the findings of L.M.P. Guerreiro (2021), who emphasised the prospects of using MCTS as an effective agent for creating a dataset and then applying it to training neural networks for real-time tasks. Although the current study did not use combined approaches, such as training neural networks with MCTS, it both confirmed its effectiveness compared to other agents and highlighted weaknesses that require improvement and further research. The modified versions of MCTS described in X. Liu *et al.* (2009), demonstrate similar efficiency in adapting to dynamic conditions by integrating machine learning methods, in particular artificial neural networks. These methods improve convergence speed and optimisation, which may be a further promising direction in complex dynamic environments and the context of UAV missions or related tasks involving robots.

An important conclusion of the study is the confirmation of the high computational requirements of algorithms such as A*, which, as noted by N. Salem *et al.* (2024), demonstrate excellent performance in simple environments, but their application in environments with increased complexity is problematic due to the significant computational cost. An attempt to integrate A* revealed that the algorithm cannot provide stable performance in large mazes due to delays in calculations.

Analysis of the behaviour of agents in the proposed modification of the Pac-Man game demonstrated how different strategies and algorithms can be applied to solve complex real-world problems, such as controlling drones, robotic systems or other moving objects to perform tasks on the ground. This also demonstrated the importance of adapting algorithms to changing conditions and multiple factors that are common in many real-world scenarios, including planning, resource management, automation, or logistics. The effective use of strategic agents in a gaming context not only contributes to the development of artificial intelligence technologies but also provides a valuable tool for solving practical problems where agents must adapt to complex and unpredictable conditions, including limited resources and a changing environment.

The results of the study confirmed that MCTS is the most promising method for complex dynamic environments due to its flexibility and ability to balance research and operation. However, to be used effectively in real-world applications such as UAV control, further improvements are needed to reduce computational costs and improve adaptability. Alpha-Beta Pruning can be used in tasks where it is important to reduce the search space, such as in scenarios with predictable obstacles or a limited set of possible actions, with the potential for possible improvement through additional research.

The potential applications of the investigated algorithms in real-world scenarios, such as UAV missions, include automatic route planning, reconnaissance, collision avoidance, and resource management in energy-limited

environments through high-quality route planning. For instance, MCTS can be used to quickly adapt to changes in the environment, such as the appearance of new obstacles or targets, thanks to its ability to conduct simulations and consider multiple scenarios. This is particularly relevant for missions in complex environments, such as densely populated areas or search and rescue operations, where fast and accurate decision-making is critical.

Thus, the results of the study not only emphasised the potential of the presented algorithms to perform complex tasks in dynamic environments but also demonstrated the need for further improvement to increase efficiency in real-world applications, such as UAV missions and other automated control systems.

Conclusions

The study conducted a series of experiments in three types of mazes with different levels of complexity, which varied in the number of walls and size. Each maze modelled realistic navigation conditions, in particular, simulated situations requiring route variability and the need to evade competitors (in the form of ghosts) chasing Pac-Man. The data on the average score, game duration, and winning percentage were collected, which was used to analyse in-depth the features of each algorithm in specific conditions. The study determined that the MCTS algorithm was most effective in mazes of lower complexity, as its ability to simulate allows for a quick assessment of potential options and the selection of the most effective paths. At the same time, as the complexity of the maze increased, the effectiveness of MCTS decreased, as the increase in the number of possible options complicated the decision-making process, requiring more resources. This shows the importance of further optimising this algorithm to adapt to complex scenarios. Alpha-Beta Pruning performed reasonably well in simpler environments, but its effectiveness gradually decreased with the increasing complexity of the environment, similar to MCTS. The limitations of Alpha-Beta Pruning in such situations may be related to the need to expand its ability to quickly analyse alternatives.

Expectimax performed consistently poorly compared to the other algorithms, especially in more complex environments, which is possibly caused by reliance on random events that limited its effectiveness in complex scenarios where rapid adaptation to conditions is crucial. This algorithm was the least effective for simple mazes, but its performance dropped sharply in environments requiring flexibility and evasion strategies. Thus, the data obtained concluded that it is important to choose an appropriate algorithm depending on the level of task complexity.

An important aspect of this research was that its results can be used to further develop navigation algorithms for autonomous systems, such as drones or robotic platforms, operating in complex dynamic environments. The use of algorithms with the ability to rapidly adapt and optimise solutions under conditions of limited resources and high levels of uncertainty is particularly relevant for such scenarios. MCTS has demonstrated the potential for use in autonomous navigation systems, where adaptation to unpredictable environmental changes is critical.

Further research could address the integration of machine learning techniques to improve the MCTS Agent, which has demonstrated the best adaptability in complex environments. Combining MCTS with neural networks can provide more efficient state estimation and reduce the need for many simulations, thus lowering computational costs. In the future, the integration of machine learning into MCTS can strike a balance between the speed and accuracy of decision-making required to deal with dynamic environments with many variables. Such improvements can render MCTS more efficient and suitable for use in tasks close to real-world scenarios, such as automating the navigation of drones, autonomous robots, or other robotic systems in dynamic environments.

Acknowledgements

None.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] Busatto-Gaston, D., Chakraborty, D., & Raskin, J.-F. (2020). Monte Carlo tree search guided by symbolic advice for MDPs. In *31st international conference on concurrency theory (CONCUR 2020)*. *Leibniz international proceedings in informatics (LIPIcs)* (Vol. 171, pp. 40:1-40:24). Schloss Dagstuhl: Leibniz-Zentrum für Informatik. doi: 10.4230/LIPIcs.CONCUR.2020.40.

[2] Cheng, K. M., Liu, H., & Dou, X. (2024). Randomized Pacman maze generation algorithm. *Applied and Computational Engineering*, 42, 156-162. doi: 10.54254/2755-2721/42/20230771.

[3] Cheng, Y., Hu, X., Tang, Q., Qi, H., & Yang, H. (2020). Monte Carlo Tree search-based mixed traffic flow control algorithm for arterial intersections. *Transportation Research Record*, 2674(8), 167-178. doi: 10.1177/0361198120919746.

[4] Evillasio2. (2022). *The Pac-Man Ghosts Alt*. *DeviantArt*. Retrieved from https://www.deviantart.com/evilasio2/art/The-Pac-Man-Ghosts-Alt-916669643.

[5] Guerreiro, J.M.P. (2021). *Learning agent in the Ms. Pac-Man vs Ghosts game*. (Master's Thesis, Instituto Superior Técnico, Lisboa, Portugal).

[6] Li, W., Liu, Y., Ma, Y., Xu, K., Qiu, J., & Gan, Z. (2023). A self-learning Monte Carlo tree search algorithm for robot path planning. *Frontiers in Neurorobotics*, 17. doi: 10.3389/fnbot.2023.1039644.

[7] Liu, X., Li, Y., He, S., Fu, Y., Yang, J., Ji, D., & Chen, Y. (2009). To create intelligent adaptive game opponent by using Monte-Carlo for the game of Pac-Man. In *Fifth international conference on natural computation* (pp. 598-602). Tianjian: IEEE. doi: 10.1109/ICNC.2009.633.

[8] Lövétei, I. F., Kővári, B., & Bécsi, T. (2021). MCTS based approach for solving real-time railway rescheduling problem. *Periodica Polytechnica Transportation Engineering*, 49(3), 283-291. doi: 10.3311/PPtr.18584.

[9] Maddipati, H., Kundurthi, A., Raaj, P., Srilatha, K., & Surapaneni, R. (2020). Artificial Intelligence based Pacman Game. *International Journal of Innovative Technology and Exploring Engineering*, 9, 140-144. doi: 10.35940/ijitee.I6975.079920.

[10] Mishra, P., Patel, V., Mittal, P., & Patni, J.C. (2018). Algorithm analysis tool based on execution time-input instance-based runtime performance benchmarking. In *International conference on recent developments in science, technology, humanities and management – 2017* (pp. 27-30). Kuala Lumpur: SRD.

[11] Mudda, M. (2022). Tic Tac Toe by Minimax Alpha-Beta Pruning using Arduino. *International Journal for Research in Applied Science and Engineering Technology,* 10(2), 157-165. doi: 10.22214/ijraset.2022.40115.

[12] Pepels, T., Winands, M. H. M., & Lanctot, M. (2014). Real-time Monte Carlo Tree search in Ms Pac-Man. *IEEE Transactions on Computational Intelligence and AI in Games*, 6(3), 245-257. doi: 10.1109/TCIAIG.2013.2291577.

[13] Permatasari, B.D., Haryanto, H., Zuni Astuti, E., & Dolphina, E. (2022). Peningkatan kemenangan non-playable character dalam permainan triple triad Menggunakan Alpha-Beta Pruning. *Jurnal Komputasi*, 10(1). doi: 10.23960/komputasi.v10i1.2952.

[14] Ramadhan, A. W. R., & Udjulawa, D. (2020). Perbandingan algoritma dijkstra dan algoritma a star pada permainan Pac-Man. *Jurnal Algoritme*, 1(1), 12-20. doi: 10.35957/algoritme.v1i1.411.

[15] Salem, N., Haneya, H., Balbaid, H., & Asrar, M. (2024). Exploring the maze: A comparative study of path finding algorithms for PAC-Man game. In *21st learning and technology conference (L&T)* (pp. 92-97). Jeddah: IEEE. doi: 10.1109/LT60077.2024.10469459.

[16] Samothrakis, S., Robles, D., & Lucas, S. (2011). Fast approximate Max-n Monte Carlo Tree search for Ms Pac-Man. *IEEE Transactions on Computational Intelligence and AI in Games*, 3(2), 142-154. doi: 10.1109/TCIAIG.2011.2144597.

[17] Sharma, N. (2022). *Introduction to artificial intelligence.* Retrieved from https://inst.eecs.berkeley.edu/~cs188/fa22/assets/notes/cs188-fa22-note06.pdf.

[18] Shevtekar, P.S., Malpe, M. & Bhaila, M. (2022). Analysis of game tree search algorithms using Minimax Algorithm and Alpha-Beta Pruning. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT)*, 8(6), 328-333. doi: 10.32628/CSEIT1228644.

[19] Singhal, S.P., & Sridevi, M. (2019). Comparative study of performance of parallel alpha Beta Pruning for different architectures. In *2019 IEEE 9th international conference on advanced computing (IACC)* (pp. 115-119). Tiruchirappalli: IEEE. doi: 10.1109/IACC48062.2019.8971591.

[20] Świechowski, M., Godlewski, K., Sawicki, B., & Mańdziuk, J. (2023). Monte Carlo Tree search: A review of recent modifications and applications. *Artificial Intelligence Review*, 56, 2497-2562. doi: 10.1007/s10462-022-10228-y.

[21] Wang, B., Ma, R., Kuang, J., & Zhang, Y. (2020). How decisions are made in brains: Unpack “black box” of CNN with Ms. Pac-Man video game. *IEEE Access*, 8, 142446-142458. doi: 10.1109/ACCESS.2020.3013645.

[22] Zou, Y. (2021). General Pacman AI: Game agent with tree search, adversarial search and model-based RL algorithms. In *2nd International conference on big data & artificial intelligence & software engineering (ICBASE)* (pp. 253-260). Zhuhai: IEEE. doi: 10.1109/ICBASE53849.2021.00053.

Аналіз ефективності алгоритмів прийняття рішень в умовах складних ігрових середовищ на прикладі Pac-Man

Артем Новіков

Аспірант
Харківський національний університет ім. В.Н. Каразіна
61022, Майдан Свободи, 4, м. Харків, Україна
https://orcid.org/0009-0004-5914-7098

Володимир Яновський

Доктор фізико-математичних наук, професор
«Інститут монокристалів» Національної Академії наук України
61072, пр-т Науки, 60, м. Харків, Україна
https://orcid.org/0000-0003-0461-749X

Анотація. Ігрові симуляції типу Pac-Man є важливим інструментом для тестування алгоритмів прийняття рішень в умовах, що імітують реальні сценарії. Це відкриває нові можливості для розробки автономних систем, які можуть адаптуватися до мінливих умов навколишнього середовища та взаємодіяти з іншими агентами. Метою цієї роботи було порівняння алгоритмів Expectimax, Monte Carlo Tree Search та Alpha-Beta Pruning у змінених умовах гри Pac-Man для визначення найбільш ефективного підходу до прийняття рішень у складних середовищах. Для цього було використано імітаційне моделювання для оцінки ефективності роботи агентів у різних ігрових лабіринтах, що відрізняються за складністю. У дослідженні вимірювалися такі показники, як кількість балів, час гри та відсоток виграшів, що дозволило оцінити ефективність алгоритмів у різних ситуаціях. Аналіз проведених експериментів продемонстрував, що алгоритм Монте-Карло є найбільш ефективним серед протестованих методів для вирішення менш складних лабіринтів, підтверджуючи його здатність швидко знаходити оптимальні шляхи в простих умовах. Алгоритм Alpha-Beta Pruning показав меншу ефективність, що вказує на необхідність його оптимізації для роботи в більш складних середовищах. Expectimax продемонстрував значно нижчу продуктивність, що свідчить про його обмежену придатність для складних ігрових лабіринтів. Дослідження показало, що збільшення складності лабіринтів значно знижує продуктивність усіх алгоритмів, особливо при більшій кількості перешкод, підкреслюючи важливість розробки більш стійких методів для роботи у високоскладних умовах. Оптимізація алгоритмів Монте-Карло та Alpha-Beta Pruning для роботи в складних середовищах може суттєво покращити їх результативність та зробити їх ефективними для реальних застосувань у навігації та керуванні рухомими пристроями. Результати цього дослідження можуть бути використані для розробки ефективних алгоритмів навігації для автономних транспортних засобів, дронів та інших робототехнічних систем, де адаптація до змін у складних умовах є критично важливою

Ключові слова: симуляційні середовища; Expectimax; MCTS; Alpha-Beta Pruning; автономна навігація

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

Information Technologies and Computer Engineering **Vol. 21, No. 3. 2024**

Article's History: Received: 03.08.2024 Revised: 19.10.2024 Accepted: 26.12.2024

UDC 519.23:004.021 10.63341/vitce/3.2024.119

Improved A/B testing acceleration methods for parametric hypothesis testing: T-test comparison with CUPED, CUPED++ and Bayesian Estimator

Artur Markov*

Postgraduate Student
Taras Shevchenko National University of Kyiv
01033, 60 Volodymyrska Str., Kyiv, Ukraine
https://orcid.org/0009-0000-5222-4397

Abstract. The study aimed to compare statistical analysis methods to improve the testing of alternatives. The study evaluated four main methods: the classic T-test, the conventional and advanced method of Controlled Experiments Using Pre-Experimental Data (CUPED), and the Bayesian Estimator. The main results included a demonstration of the A/B testing process, and the described statistical analysis methods included detailed characteristics and examples of use. The simulations and practical application revealed that the T-test provides high accuracy with small samples, but its effectiveness decreases with increasing sample size due to high resource requirements. The calculator for this method demonstrated effectiveness in simple tasks but had limitations with large data. The conventional CUPED method has shown increased accuracy due to variation correction, but its effectiveness decreases when working with large and complex data sets. The written program for this method has shown to be effective in cases where the previous data is well represented, but its capabilities are limited when processing large data sets. The improved version provided a significant improvement in both accuracy and processing speed, especially for large datasets, thanks to advanced modelling and optimisation. The code results confirmed that this method is highly efficient for complex experiments, particularly when processing large amounts of data. Moreover, the Bayesian Estimator demonstrated high accuracy due to the integration of prior knowledge but required more computational resources and time. The platform used for this method demonstrated the ability to account for uncertainty yet required complex model settings. The results highlighted the importance of selecting the appropriate statistical analysis method depending on the scale and complexity of the data to ensure optimal accuracy and efficiency of testing

Keywords: statistical analysis; correction of variations; effectiveness of approaches; data processing; modelling of experiments

Introduction

With the rapid development of information technology and the increase in data volumes, the effectiveness of statistical analysis methods has become highly relevant. One of the basics of statistical analysis is A/B testing, which is used to compare two or more options and determine the most effective one. The basics of such testing include comparing the mean values in two groups using parametric hypotheses to assess the statistical significance of differences between the groups. Traditionally, such comparisons are made using the T-test, which is a method for analysing average values. However, to improve accuracy and speed, advanced methods have been developed, such as Controlled Experiments Using Pre-Experimental Data (CUPED) and its extended version CUPED++, as well as the Bayesian Estimator method.

The T-test method, although a basic tool for comparing values, has limitations when analysing large and complex data sets due to high resource requirements and inaccuracy in cases of variation. Although CUPED can be used to correct for variations by using previous data, its effectiveness decreases when processing large amounts of data. The extended version of CUPED++ offers certain improvements but still requires a detailed comparison with other modern methods, such as the Bayesian Estimator, which provides high accuracy by integrating prior knowledge but is

Suggested Citation:
Markov, A. (2024). Improved A/B testing acceleration methods for parametric hypothesis testing: T-test comparison with CUPED, CUPED++ and Bayesian Estimator. *Information Technologies and Computer Engineering*, 21(3), 119-131. doi: 10.63341/vitce/3.2024.119

*Corresponding author

resource-intensive. Existing studies do not sufficiently address these aspects, which determines the need for a deeper analysis and comparison of new and improved methods in the context of their practical application.

For instance, V. Kalchenko (2018) analysed common methods of testing for penetration into computer systems and proposed a classification of these methods, noting their advantages and disadvantages. O.V. Shportko & M.M. Mushyn (2023) substantiated the effectiveness of the method of gradual formation of a set of objective function values for combinatorial optimisation problems, demonstrating that this method significantly reduces the execution time of programs compared to other approaches, such as search methods with returns. In addition, V. Khambir (2024) proved that the automation of testing processes, including functionality, interface, performance, and security testing, facilitates and increases the efficiency of application testing, despite the problems with the initial setup and complexity of testing complex scenarios.

Furthermore, K. Kachiashvili (2018) described the intensive development of statistical hypothesis testing methods and introduced a new approach known as Constrained Bayesian Methods to solve complex problems in various fields, including a detailed description of existing methods, new approaches and special software for their applications. X. Gu *et al.* (2023) presented a Bayesian method for testing information hypotheses in confirmatory factor analysis models, in particular, by estimating factor loadings using Bayesian factors and posterior probabilities of models, which can be used to solve various issues, including the assessment of equality of loadings and indicator validity. Additionally, Ramesh *et al.* (2023) reviewed the process of formulating and testing hypotheses, emphasising the importance of theoretical grounding and a systematic approach to hypothesis testing, which ensures objectivity in the search for knowledge. A. Deng *et al.* (2021) developed a new unbiased reduced variance estimator of intervention effects that uses the idea of improving the efficiency of CUPED for cases where the trigger status is observed only in a certain group, which allows achieving accuracy comparable to full status observation.

In turn, J.J. Wang (2022) developed an approximate Bayesian estimator for the parameters of a regression model with random coefficients, which outperforms traditional testing methods in terms of estimation accuracy, simplifying the calculation and interpretation of results. S. Woo (2023) presented parametric accelerated durability testing as a systematic method for assessing the reliability of mechanical system structures, which allows the detection of structural defects and reduces the failure rate due to material fatigue. In addition, Y.W. Cho *et al.* (2024) presented the results of a series of Monte Carlo simulations to evaluate methods for fitting multilevel models of latent differential structural equations, finding that the Bayesian approach outperforms frequency-based methods, especially in the context of handling variable covariates and coupling parameters.

Although these studies addressed various statistical methods, there are still gaps in their adaptation to big data and complex experiments. This study describes and compares the effectiveness of such methods in improving A/B testing, with a focus on the accuracy and speed of data analysis. The tasks included a theoretical study of the basics of A/B testing and advanced methods, conducting simulations to assess their practical applicability, and analysing the effectiveness of different approaches.

Materials and Methods

The study was initiated by a comprehensive analysis of the theoretical context of A/B testing and improved analysis methods, including the basics and value of this testing for comparing alternatives. Parametric hypotheses, their role in testing, and their impact on the accuracy and speed of results were studied. Traditional and advanced methods, such as T-test, CUPED, CUPED++, and Bayesian Estimator, their algorithms and applications, as well as the importance of advanced methods for improving testing efficiency were highlighted.

A detailed description and characterisation of these methods of statistical analysis were provided in the study. For each method, the principles of operation, implementation of algorithms and examples of application are described. The classical T-test was considered a basic method for comparing mean values, CUPED and CUPED++ as methods for correcting variations to improve the accuracy of results, and the Bayesian Estimator as a method for estimating probabilities with the integration of prior knowledge. The practical application of the methods included the creation and testing of code and formulas for each method. The general formula for the T-test method (1):

$$T = \frac{M_1 - M_2}{SE}, \quad (1)$$

where: M_1, M_2 – two mean values; SE – overall standard error for the two samples.

The general formula for the CUPED method is (2):

$$\hat{Y}_{cv} = \bar{Y} - \theta\bar{X} + \theta E(X), \quad (2)$$

where: X – value before the experiment; Y – value of the experiment; θ – a constant derived from the data that determines how much the results of the experiment should be adjusted based on the previous data; $\overline{Y}$ – average value of the experiment results Y; $\overline{X}$ – average value of the previous data X; $E(X)$ – mathematical expectation of the values X which is used for correction.

The general formula for the Bayesian Estimator method (3):

$$P(A|B) = \frac{P(B|A)*P(A)}{P(B)}, \quad (3)$$

where: $P(A|B)$ – the probability that a certain state or property (A) exists, given another property or condition (B); $P(B|A)$ – the probability that a certain property or condition (B) exists, given that another state or property

(*A*) exists; *P*(*A*) – the probability that (*A*) exists without taking into account additional conditions or properties; *P*(*B*) – the probability of having (*B*) without taking into account additional conditions or properties.

The programs for the implementation of CUPED and CUPED++ were written in Python using the pandas and stats models libraries and tested on the Replit platform. In turn, an online calculator was used to demonstrate the T-test method, and the Bayesian Estimator was based on the platform "Bayesian Estimation Supersedes the T-test – online", which shows how much better this method is than the T-test.The comparison of the effectiveness of the methods was based on an assessment of the accuracy, speed and overall efficiency of each method. The results were analysed to determine the advantages and limitations of each method, which was used to provide recommendations for practical application depending on the specifics of the tasks and the size of the data sets.

Results

Theoretical context of A/B testing and advanced analysis methods

A/B testing is a standard method for comparing the effectiveness of two or more alternatives in a statistical analysis. This approach is used to determine which of the options (A or B) is more effective in achieving a particular goal or indicator. A/B testing is based on the idea of random experiments, where respondents are randomly assigned to different groups, which minimises systematic errors and reveals the net effect of changes.

This testing usually involves two stages: dividing users into control (A) and experimental (B) groups and evaluating the results to compare effectiveness (Fig. 1). The main statistical tool for analysing the results of A/B testing is statistical hypothesis testing, which can be used to determine whether the difference between the groups is statistically significant.

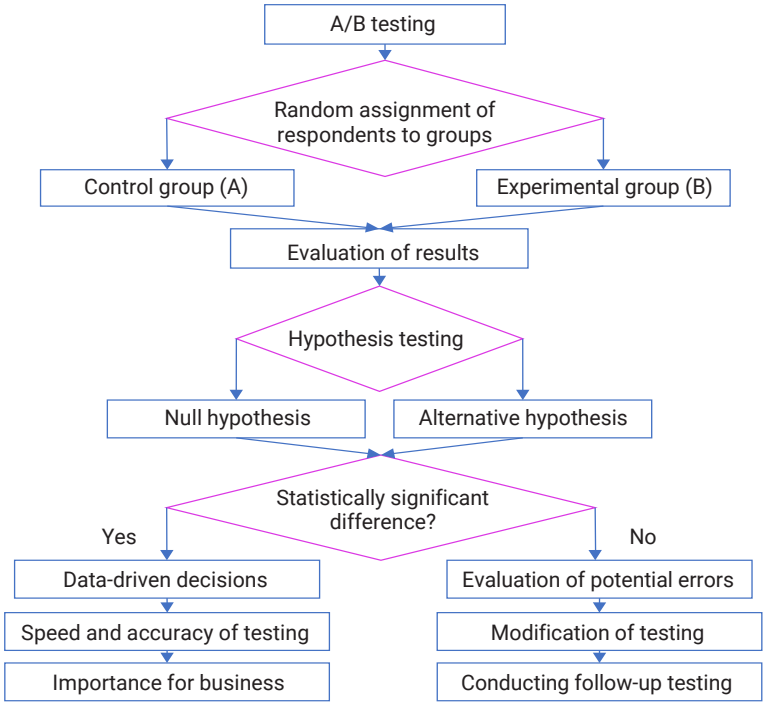

Figure 1. Flowchart of the A/B testing process

Source: compiled by the author

In addition, A/B testing can be used to objectively compare alternatives and make informed decisions based on empirical data. This is important for businesses where even small improvements can have a significant impact on the performance of campaigns or products. The success of A/B testing depends on the ability to evaluate results quickly and accurately. Increasing the speed and accuracy of testing improves operational decision-making based on up-to-date data and minimises the risk of false conclusions. This is especially relevant in a rapidly changing market and highly competitive environment. Also important are parametric hypotheses, which are mathematical models that formulate expectations about the distribution of data in the populations being compared. The null hypothesis (H0) typically states that there is no significant difference between the groups, while the alternative hypothesis (H1) states that there is a difference. In A/B testing, the correct formulation and testing of these hypotheses are critical to obtaining reliable results. The correct formulation and testing of parametric hypotheses affect the accuracy of testing. Incorrect assumptions can lead to erroneous conclusions and, as a result, to wrong decisions. In addition, methods that provide more accurate and faster results can significantly reduce the time required to collect and analyse data. Traditional

methods, such as T-test, provide a basic level of accuracy for A/B testing, but they have limitations in speed and accuracy, especially in the face of large amounts of data and frequent testing. Modern methods, such as CUPED, CUPED++ and Bayesian Estimator, provide improvements in these aspects with their advanced algorithms and techniques.

Improved analysis methods can reduce data variation, improve the accuracy of estimates and reduce the time required to reach statistical significance. For example, the CUPED method uses previous data to correct for variation and improve accuracy, while the Bayesian Estimator provides the flexibility to model and adapt to new data. These improvements contribute to more efficient testing and decision-making based on more reliable results.

Description and characteristics of statistical analysis methods

Statistical methods of analysis play a key role in A/B testing, providing effective comparisons of different alternatives and evaluation of their effectiveness. For this purpose, various methods are used, each of which has its characteristics and approaches to data processing. The standard T-test is the main statistical test for comparing the means of two independent groups. It determines whether there are significant differences between the means of the groups being compared. The basic principle of the t-test is to test H0, which states that the means of two groups do not differ from each other, against H1, which suggests that there is a statistically significant difference between the groups. The T-test procedure involves several key steps (Fig. 2). First, data are collected from two independent groups to be compared. Next, the means and standard deviations for each group are calculated. Based on these figures, the T statistic is calculated, which reflects the amount of difference between the means concerning the variation in the data. The test statistic T is compared with the critical value from the T-distribution, which assesses the probability of observing such a difference under the null hypothesis.

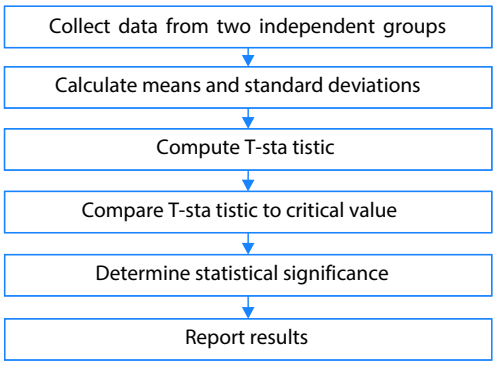

Figure 2. T-test scheme of operation
Source: compiled by the author

An example is the T-test, which examines the effectiveness of a new advertising message (group B) compared to an old advertising message (group A). If the results show that the average values of the performance indicators in groups A and B differ significantly, and this deviation is statistically significant, then it can be concluded that the new advert is effective.

The CUPED method, in turn, is a modern statistical approach that improves the accuracy of A/B testing by using previous data to correct for variations in results. It uses regression analysis to incorporate data collected before the experiment into the estimation of the effect of changes. This reduces the variation in test results and increases the accuracy of estimates. CUPED is based on two main stages (Fig. 3). The first stage involves collecting preliminary data, which may include information about user behaviour before the experiment or data about various user characteristics. In the second stage, this data is used to build a regression model that corrects for variations in the main results of A/B testing. Regression analysis can assess the impact of these variations on the final results, reducing noise and improving the accuracy of the estimates.

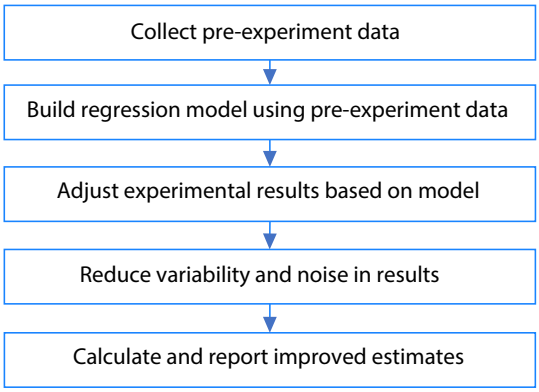

Figure 3. CUPED scheme of operation
Source: compiled by the author

For instance, if A/B testing is conducted to evaluate the effectiveness of a new feature in an application, CUPED can use data on user behaviour before the new feature is implemented to adjust the results. This provides a more accurate assessment of the impact of the new feature on user behaviour, as noise from various factors that affect the results will be reduced. Moreover, the CUPED++ method is an advanced version of CUPED designed to increase the accuracy of A/B testing by reducing variation and improving the correction of results. The main goal of CUPED++ is to further improve the process of variation correction by introducing new algorithms and advanced modelling methods. CUPED++ is based on CUPED but introduces several new optimisation approaches (Fig. 4). The main steps of CUPED++ include collecting preliminary data, building an extended regression model, and adapting and optimising the model for specific experimental conditions. CUPED++ uses more sophisticated regression algorithms that address additional factors and relationships between variables. This allows for even better correction of variations and reduces the impact of random noise on test results.

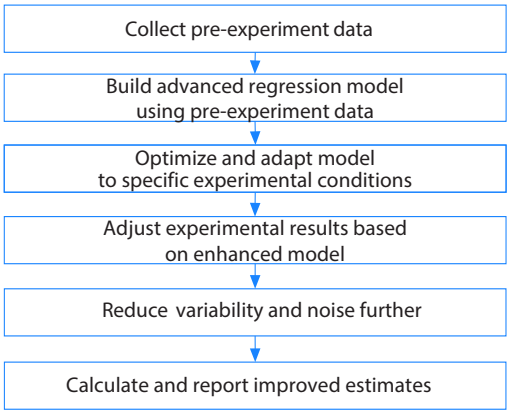

Figure 4. CUPED++ scheme of operation
Source: compiled by the author

For instance, when testing a new feature in a web application, CUPED++ can include not only data on user behaviour before the new feature was introduced but also data on seasonal trends or other factors that may affect the results. This provides even greater accuracy in measuring the effect of the new feature, as CUPED++ takes these additional variables into account when adjusting the results.

It is also worth considering the Bayesian Estimator, which is a Bayesian technique for estimating the difference between two means, which provides a probability distribution of this difference. Bayes' theorem describes how the probability of a certain hypothesis changes based on new information. The Bayesian Estimator is based on the concept of conditional probability, which integrates previous knowledge (a priori probabilities) and new observations to update probabilities (a posteriori probabilities). Bayesian analysis methods provide high flexibility in modelling and adapting to new data, which makes them useful in the face of uncertainty (Fig. 5). They can be used to estimate probabilities and draw reasonable conclusions using complex information from various sources. Bayesian methods are used for modelling and forecasting in the face of incomplete or inaccurate data, as well as for adapting to changes in data.

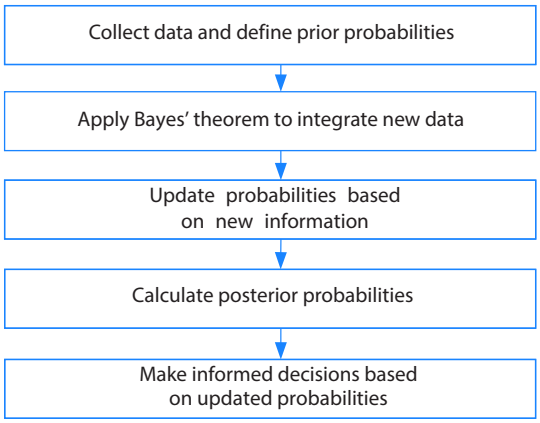

Figure 5. The Bayesian Estimator workflow
Source: compiled by the author

For instance, in A/B testing, a Bayesian Estimator can be used to determine the probability that option B is better than option A. If there are certain a priori assumptions about the effectiveness of options, the Bayesian approach adapts these assumptions based on the actual results of the experiment, providing more accurate and reasonable conclusions. Consequently, there are many different methods for evaluating and comparing alternatives, each with its characteristics and advantages. Existing approaches allow for precise data analysis, variation correction, adaptation to new conditions, and integration of prior knowledge with new data. The use of different methods makes it possible to choose the most effective tools for specific tasks, which is critical for achieving accurate and reliable results in testing and analysis.

Simulations and practical application of statistical methods

In general, the T-test can be used to compare mean values between two groups. There are several types of T-tests for two samples, such as unpaired, Welch's, and paired T-tests. In addition, there are many platforms and calculators that can calculate the necessary parameters to select the most appropriate test for analysis (Fig. 6).

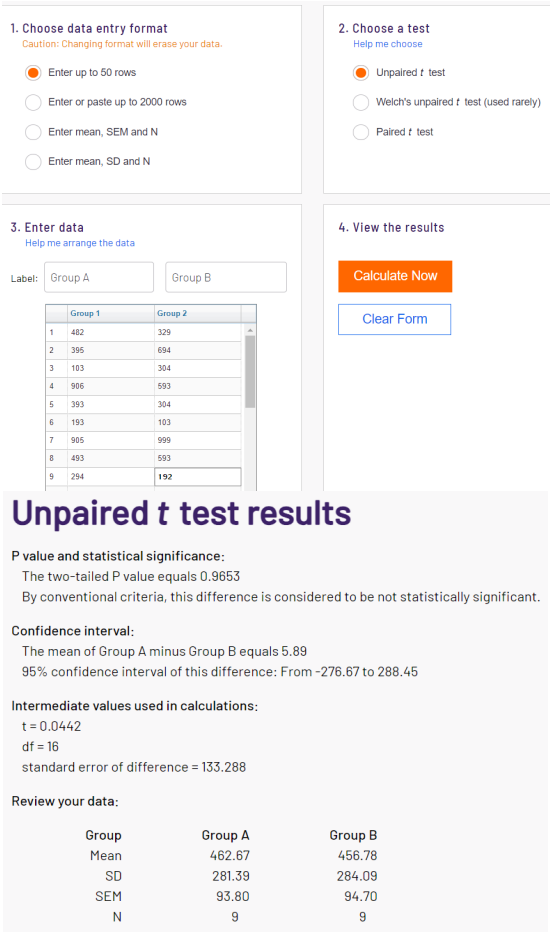

Figure 6. An example of using a calculator for a T-test
Source: compiled by the author

This test focuses on comparing data from a single numerical variable, rather than on counts or correlations between multiple variables. This method is especially useful for small samples containing less than 30 observations. This calculator uses the general formula (1) for a T-test consisting of two means and a common standard error for two samples. In other words, the calculator uses this formula to calculate the T-statistic and evaluate the significance of the difference between groups, helping users to draw reasonable conclusions from statistical data. Unlike the T-test, the CUPED method works by correcting for variations in experimental results using previous data. It uses information collected before the experiment to create a regression model that reduces the effects of noise and other unwanted variations. This correction helps to improve the accuracy of the estimates of the effects of changes on which conclusions are based. CUPED simplifies the analysis process by reducing the standard error of the estimates, which provides clearer and more reliable results (Fig. 7).

```
# Importing libraries
import pandas as pd
import numpy as np
# Data creation
data = pd.DataFrame({
   'group': ['A'] * 50 + ['B'] * 50,
   'pre_experiment':
   [x + (10 if i < 50 else 15) for i, x in enumerate(range(50))] * 2,
   'outcome': [x + (5 if i < 50 else 7) for i, x in enumerate(range(50))] * 2
})
# Calculating average values
mean_outcome = data.groupby('group')['outcome'].mean()
mean_pre_experiment = data.groupby('group')['pre_experiment'].mean()
# Determination of the constant θ
theta = np.corrcoef(data['outcome'], data['pre_experiment'])[0, 1]
# Calculation of the mathematical expectation E(X)
mean_pre_experiment_total = data['pre_experiment'].mean()

# Application of the CUPED formula to correct results
data['adjusted_outcome'] = (data['outcome'] -
                    mean_outcome[data['group']].values +
                    theta * mean_pre_experiment_total)
# Calculation of adjusted averages for each group
adjusted_means = data.groupby('group')['adjusted_outcome'].mean()

print("\nCUPED Adjusted Mean Outcomes:")
print(adjusted_means)
```

Figure 7. Example code for using the CUPED method

Source: compiled by the author

This code implements the CUPED method by performing a regression analysis based on the previous data and the results of the experiment (Formula 2). First, a regression model is created using the previous data as predictors to adjust the results. The modified data is then subject to further analysis to assess the effect of the changes, considering the reduced impact of variations. The programme generates adjusted averages of the results and performs statistical analysis to confirm the accuracy and reliability of the findings. The results demonstrate an excellent fit of the model to the data, as the coefficient of determination is 1, indicating that the model fully explains the variation in the dependent variable, and the high F-statistic and significance of the p-values confirm that the model is statistically significant (Fig. 8).

```
data = pd.DataFrame({
    'group': ['A'] * 50 + ['B'] * 50,
    'pre_experiment':
    [x + (10 if i < 50 else 15) for i, x in
enumerate(range(50))] * 2,
    'outcome': [x + (5 if i < 50 else 7) for i, x in
enumerate(range(50))] * 2
})
```

```
Run

CUPED Adjusted Mean Outcomes:
group
A    34.5
B    34.5
Name: adjusted_outcome, dtype: float64
```

Figure 8. The result of the CUPED application programme

Source: compiled by the author

As for the CUPED++ method, it is an extension of CUPED that improves the process of variation correction by introducing new algorithms and optimisations. Unlike the basic CUPED, CUPED++ uses additional techniques to model and adapt to specific experimental conditions, allowing for a more accurate assessment of the effects of changes. This method incorporates advanced modelling of the preliminary data and introduces advanced correction mechanisms that help to further reduce the variation in the experimental results (Fig. 9).

```
import numpy as np
import pandas as pd
import statsmodels.api as sm

# Creating synthetic data
np.random.seed(0)
data = pd.DataFrame({
    'group': np.random.choice(['A', 'B'], size=100),
    'pre_experiment': np.random.randn(100),
    'outcome': np.random.randn(100)
})

# Defining the function to be implemented CUPED++
def apply_cuped_plus(data):
    # Dividing into control and experimental groups
    pre_data = data[data['group'] == 'A']
    post_data = data[data['group'] == 'B']

    # Building a regression model based on previous data
    X_pre = sm.add_constant(pre_data['pre_experiment'])
    y_pre = pre_data['outcome']
    model = sm.OLS(y_pre, X_pre).fit()

    # Predicting outcomes based on the model
    X_post = sm.add_constant(data['pre_experiment'])
    data['adjusted_outcome'] = model.predict(X_post)
    return data

# Applying CUPED++ to data
adjusted_data = apply_cuped_plus(data)

# Calculation of average values for each group after correction
means_adjusted = adjusted_data.groupby('group')['adjusted_outcome'].mean()
print(means_adjusted)
```

Figure 9. A code example of using the CUPED++ method

Source: compiled by the author

The code implements CUPED++ by correcting the results of the experiment based on previous data. First, synthetic data is created, including information about the groups, preliminary results, and the results of the experiment. The apply_cuped_plus function builds a regression model based on the preliminary data from the control group. The model is then used to predict the adjusted results for both groups, including both the control and experimental groups. Finally, the average of the adjusted scores for each group is calculated. The results show the mean values of the adjusted results for each group after applying CUPED++. In this example, the mean values of the adjusted results for the control and experimental groups differ, indicating that CUPED++ has successfully reduced the variation in the results and provided more accurate estimates (Fig. 10).

```
Run
group
A   -0.205947
B   -0.244400
Name: adjusted_outcome, dtype: float64
```

Figure 10. A code example of using the CUPED++ method

Source: compiled by the author

In addition, the Bayesian Estimator method, which is based on Bayes' theorem, is relevant, as combines previous knowledge with new data to estimate the probability of hypotheses (Fig. 11). In this platform, the formula (3) for the Bayesian Estimator method is used to estimate the difference between the group means and determine the range in which this difference is likely to be located. It uses a Monte Carlo method with a metropolitan algorithm to generate samples and calculate the distribution of the difference.

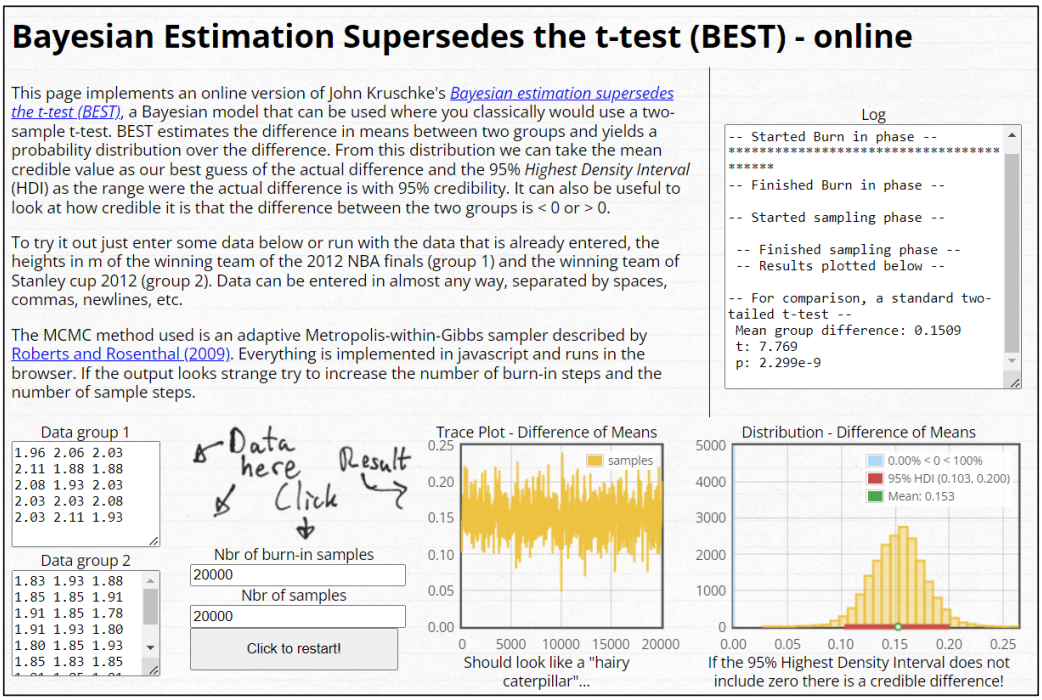

Figure 11. An example of using the Bayesian Estimator platform

Source: compiled by the author

The results of the programme show the mean of the reliable value of the difference and the interval of the highest density, which reflects the range of probable values of the difference. Thus, the Bayesian Estimator method can be used to obtain more accurate and reliable estimates of the effects of changes, which is especially important for estimating probabilities and analysing the results of experiments. Thus, it can be noted that the simulations confirmed the effectiveness and accuracy of various statistical methods. The use of methods such as T-test, CUPED, CUPED++ and Bayesian Estimator demonstrated their ability to adequately correct and analyse data, as well as adapt to specific experimental conditions.

Comparing the effectiveness of methods

The T-test method is effective in simple situations where the data are normally distributed, and the samples have similar variations. The results showed high accuracy at small sample sizes, but as the sample size increases, its speed and accuracy decrease due to high resource requirements. CUPED provided improved accuracy by correcting for variation, which reduced the standard error. It demonstrated high speed in cases where the prior data was well represented, but its performance may be reduced when large and complex datasets are present. In addition, CUPED++ has shown significant improvements in accuracy and efficiency compared to CUPED due to advanced modelling and optimisation. The speed of execution remained high even when processing large datasets, making this method very suitable for complex experiments. Bayesian Estimator demonstrated a high level of accuracy by using prior knowledge and new data. The processing speed depended on the complexity of the models and the sample size. Typically, this method takes longer to compute, but its ability to account for uncertainty makes it useful for complex tasks. The analysis showed that each method has its advantages and limitations (Table 1).

Table 1. Comparison of statistical analysis methods

Method	Advantages	Limitations	Application examples
T-test	Easy to use	Sensitive to data distribution	Comparison of mean values between two groups in medical research
	Fast in calculations	Requires data to be normal	
	Suitable for small samples	This may be inaccurate for large samples	Comparison of results before and after the intervention in the social sciences

Table 1. Continued

Method	Advantages	Limitations	Application examples
CUPID	Reduces the impact of variations	Requires preliminary data	Evaluating the effects of advertising campaigns based on previous performance
	Improves the accuracy of estimates	Can be difficult to implement	
	Uses previous data for correction	May be less effective with insufficient information	Analysing the results of educational programmes with correction for previous achievements
CUPED++	Further improvements to CUPED	More difficult to implement	Analyse the effectiveness of new products based on comprehensive preliminary data
	Improves accuracy and efficiency	May require more computing resources	
	Suitable for big and complex data	May be slower when processing large datasets	Evaluating the impact of new teaching methods on many participants
Bayesian Estimator	Considers prior knowledge and uncertainty	Slow in calculations	Risk assessment in financial models
	Provides a probabilistic approach to valuation	Can be difficult to understand and implement	Probability modelling in medical research for complex hypotheses
	Flexibility in modelling	Requires setting up models and selecting preliminary distributions	Comparing effects in experimental studies with little data

Source: compiled by the author

The choice of the most effective method depends on the specific conditions of the experiment. For simple problems with small samples, T-test or CUPED are suitable. For complex problems with large data sets, CUPED++ or Bayesian estimators are more appropriate. Moreover, if speed and simplicity are important, T-test is a good choice. For accuracy and variance correction, CUPED and CUPED++ are better options. And Bayesian Estimator should be used for tasks where prior knowledge and uncertainty need to be considered.

In general, it is recommended to choose a method based on the specific requirements of the test. Opportunities for further improvement include the integration of new algorithms to increase processing speed and optimise resource usage. Improvements in methods can also include improving the accuracy of estimates and reducing the cost of processing data in complex environments.

Discussion

The findings highlight the importance of improving A/B testing methods and the importance of applying improved approaches to ensure greater accuracy and efficiency. The main emphasis was placed on comparing the traditional T-test with newer methods such as CUPED, CUPED++ and Bayesian Estimator. Studying and comparing the methods under consideration provided a better understand of their effectiveness in different conditions and choose the most appropriate tool for specific tasks. Analysis of similar works also helps to identify improvements and enhancements, which contributes to the further development of statistical analysis methods.

The study results demonstrated the high accuracy of A/B testing methods. While C. Allard & É. Marchand (2024) studied Bayesian methods for loss estimation, showing the advantages of such approaches in reducing errors, this study emphasises that the Bayesian approach also provides high accuracy, but has high requirements for computing resources and time. The difference is that H. Zhou & H. Zou (2024) proposed a flexible Box-Cox model to improve linear hypothesis testing in high-dimensional models, while the present study focuses on corrected transformations for more accurate analysis in experiments. Thus, this study shows that specific corrections in transformations can significantly improve accuracy when dealing with large amounts of data.

While Y. Jin & S. Ba (2022) developed methods for reducing variance in online experiments using machine learning and cross-fitting techniques, which can reduce variance by up to 80% compared to traditional methods and up to 30% compared to CUPED, this study analysed different approaches to accelerating A/B testing, including CUPED, CUPED++, and Bayesian Estimator. Although CUPED already shows a significant reduction in variance, the improved CUPED++ and Bayesian Estimator demonstrated even greater accuracy and efficiency compared to this method. In addition, D. Ochieng (2024) demonstrated a hypothesis testing method, namely the two-stage randomised p-value (RAND2 p-value) method, which combines standard and adaptive approaches to test the interval composite null hypothesis. The same study on testing acceleration methods emphasises that while standard approaches can be useful, advanced methods provide significant improvements in accuracy, especially in complex experimental settings due to advanced adjustments.

The difference between the results of O. Dogan *et al.* (2020), who demonstrated a Bayesian test for testing simple null hypotheses, and the present study is the focus on adjusted quadratic loss functions, which provide high resistance to parametric non-specification, which is especially useful in cases of large samples. At the same time, K. Kachiashvili *et al.* (2023) reviewed different approaches to hypothesis testing in sequential experiments, focusing on the advantages and disadvantages of Wald, Berger, and Bayesian tests. This study showed that the new methods

provide significant improvements in accuracy for complex experiments through specific optimisations and modelling.

According to J. Liley & C. Wallace (2018), new estimators for the false rejection rate based on kernel density estimates provide the smoothness of the outlier regions but do not increase the power. The results of this study confirmed that the CUPED++ and Bayesian Estimator methods not only increase the power of the estimators but also improve the accuracy in complex experiments, which is an additional advantage over traditional approaches. A. Cissé *et al.* (2024) presented methods for improving Bayesian optimisation using expert hypotheses, which demonstrated a significant acceleration of the search. The study confirmed that the integration of expert knowledge can significantly improve efficiency compared to basic methods, especially in real-world problems.

Moreover, E. Deiri (2021) showed that expected and hierarchical Bayesian estimators converge to zero in different distributions. The present study also confirms that these approaches are effective, but the improved models provide greater accuracy in specific distributions. R. Kelter (2020) analysed Bayesian tests for comparing two groups and showed that these methods show differences in controlling first- and second-order errors compared to frequency-based methods, which requires a careful balance in practical research. The current study on A/B testing also confirmed that Bayesian approaches provide distinct characteristics compared to traditional methods, which increases the accuracy of hypothesis testing.

This study complements the research of C. Francq & J.-M. Zakoïan (2022), in which tests for testing hypotheses about symmetry and quantile assumptions were developed, as it confirms that such tests are effective for risk management and statistical analyses, although the specific parametric corrections in the study demonstrate even better adequacy for the conditional average. Additionally, F. Bertolino *et al.* (2024) showed that the "Bayesian Divergence Measure" is consistent and invariant, simplifying interpretation compared to the traditional "Full Bayesian Significance Test". This is supported by the current study, which demonstrated that the new approach can be more effective in specific conditions.

Improving the closed-form testing method for multiple hypotheses in a study by Z.-H. Lu (2020) has increased the power of detecting false hypotheses. The results obtained in the current work also confirmed that the new methods provide improved error control, for complex tests. On the other hand, B. Duthie (2024) reviewed various t-tests and non-parametric alternatives, emphasising the importance of choosing the appropriate test. The present study showed that the new techniques demonstrate significant accuracy advantages when dealing with complex datasets.

In addition, J. Tang & H. Dette (2024) presented ways for shared estimating model parameters, including methods for constructing simultaneous confidence intervals for the link function and testing joint hypotheses, which confirms their effectiveness. This indicates the relevance of such research methods in constructing confidence intervals and testing joint hypotheses, as shown in the current study. And while T. Raykov *et al.* (2022) confirmed the strong convergence of the Bayesian estimator of the median of the posterior distribution to the true parameter, the present study obtained similar results, which showed the advantages of Bayesian estimators in large samples.

The Bayesian factor methodology for testing hypotheses about the null values of parameters, developed by N. Sekulovski & H. Hoijtink (2023), has demonstrated clear characteristics regardless of sample size. In the current study, it was shown that the Bayesian Estimator provides high accuracy in models with many parameters and non-parametric data, but its efficiency may decrease when the sample size is insufficient. In turn, the model for analysing the decision-making process by L. Chauvet & D. Cruz (2024) showed that sensitivity to gains and losses affects decision-making. The results of the present study confirmed these findings, demonstrating the importance of sensitivity in choosing more profitable alternatives.

Lastly, the developed stratified selection method for random simulations presented by S.M. Baik *et al.* (2023) reduced the variation in estimates through a two-stage optimisation, and the findings in the current study confirmed that improved testing methods improve accuracy in cases with uncertain models. S. Khatami (2020) developed a "flow mapping" method to improve model accuracy, which demonstrated greater reliability than traditional metrics. In this context, the study found that an improved version of A/B testing provides significant improvements in the accuracy and speed of processing large datasets through advanced modelling and optimisation, which also contributes to improved model accuracy.

Thus, the study highlights the importance of improving A/B testing methods, demonstrating significant improvements in accuracy and efficiency when compared to traditional methods such as T-test. The identified improvements, including CUPED, CUPED++ and Bayesian Estimator, showed advantages in improving accuracy, especially when working with large data sets. Overall, the results showed that the integration of new methods and optimisations provides significant benefits in statistical analysis and modelling.

Conclusions

The study provided a detailed look at various A/B testing methods, including T-test, CUPED, CUPED++, and Bayesian Estimator. The results included a demonstration of the A/B testing process, and a comparison of statistical analysis methods based on simulations and practical applications. The study determined that the T-test shows good accuracy with small samples, but its performance decreases with increasing data volume due to high computing requirements. The CUPED method improves accuracy by correcting for variations, but its effectiveness decreases when processing large and complex data. The updated version of this method provides a significant improvement in both accuracy and speed

of data processing, particularly for large datasets, due to improved modelling. In addition, Bayesian Estimator achieves high accuracy by using prior knowledge but requires significant computing resources and time to implement.

Recommendations arising from the work include optimising the choice of method depending on the specific conditions of the experiment. For simple tasks with small samples, traditional methods such as T-test are suitable, while for large and complex datasets, it is advisable to use advanced versions of CUPED or Bayesian Estimator.

The main areas for further research are the development of new algorithms to reduce computational costs and the integration of adaptive models that can automatically adjust to specific experimental conditions. Research into new approaches to data processing and improved statistical analysis methods are also important for further progress in this area. To improve the results obtained in the future, efforts should be focused on optimising computational costs, developing more versatile models that can handle different types of data and experimental conditions, and improving algorithms that automatically adapt to changing parameters.

Acknowledgements

None.

Conflict of Interest

The authors of this study declare no conflict of interest.

References

[1] Allard, C., & Marchand, É. (2024). Bayesian and Minimax estimators of loss. *Japanese Journal of Statistics and Data Science*. doi: 10.1007/s42081-024-00261-2.

[2] Baik, S.M., Byon, E., & Ko, Y.M. (2023). Distributionally robust stratified sampling for stochastic simulations with multiple uncertain input models. *ArXiv*. doi: 10.48550/arXiv.2306.09020.

[3] Bertolino, F., Manca, M., Musio, M., Racugno, W., & Ventura, L. (2024). A new Bayesian discrepancy measure. *Journal of the Italian Statistical Society*, 33, 381-405. doi: 10.1007/s10260-024-00745-1.

[4] Chauvet, L.A., & Cruz, D.M. (2024). Computational modeling of decision-making in substance abusers: Testing Bechara's hypotheses. *Frontiers in Psychology*, 15, article number 1281082. doi: 10.3389/fpsyg.2024.1281082.

[5] Cho, Y.W., Chow, S.-M., Marini, C.M., & Martire, L.M. (2024). Multilevel latent differential structural equation model with short time series and time-varying covariates: A comparison of frequentist and Bayesian estimators. *Multivariate Behavioral Research,* 59(5), 934-956. doi: 10.1080/00273171.2024.2347959.

[6] Cissé, A., Evangelopoulos, X., Carruthers, S., Gusev, V.V., & Cooper, A.I. (2024). HypBO: Accelerating black-box scientific experiments using experts' hypotheses. In K. Larson (Ed.), *Proceedings of the thirty-third international joint conference on artificial intelligence* (pp. 3881-3889). Vienna: IJCAI. doi: 10.24963/ijcai.2024/429.

[7] Deiri, E. (2021). Expected Bayesian estimator and hierarchical Bayesian estimator for the parameter of a Rayleigh distribution reliability system under the progressive type-II data sample. *Mathematical Researches*, 7(3), 527-544. doi: 10.52547/mmr.7.3.527.

[8] Deng, A., Yuan, L.-H., & Salama-Manteau, A. (2021). Variance reduction for experiments with one-sided triggering using CUPED. *ArXiv*. doi: 10.48550/arXiv.2112.13299.

[9] Dogan, O., Taspinar, S., & Bera, A.K. (2020). A Bayesian robust chi-squared test for testing simple hypotheses. *Journal of Econometrics*, 222(2), 933-958. doi: 10.1016/j.jeconom.2020.07.046.

[10] Duthie, B. (2024). *Fundamental statistical concepts and techniques in the biological and environmental sciences*. New York: Chapman and Hall. doi: 10.1201/9781032692388.

[11] Francq, C., & Zakoïan, J.-M. (2022). Testing hypotheses on the innovations distribution in semi-parametric conditional volatility models. *Journal of Financial Econometrics*, 21(5), 1443-1482. doi: 10.1093/jjfinec/nbac011.

[12] Gu, X., Zhu, X., Zhang, L., & Pan, J.-H. (2023). Testing informative hypotheses in factor analysis models using bayes factors. *Psychological Methods*. doi: 10.1037/met0000627.

[13] Jin, Y., & Ba, S. (2022). Toward optimal variance reduction in online controlled experiments. *Technometrics*, 65(2), 231-242. doi: 10.1080/00401706.2022.2142670.

[14] Kachiashvili, K. (2018). *Constrained Bayesian methods of hypotheses testing: A new philosophy of hypotheses testing in parallel and sequential experiments*. New York: Nova Science Publishers.

[15] Kachiashvili, K., Kvaratskhelia, V., & Prangishvili, A. (2023). Comparison of constrained Bayesian and classical methods of testing statistical hypotheses in sequential experiments. In M. Zgurovsky & N. Pankratova (Eds.), *System analysis and artificial intelligence* (pp. 289-306). Cham: Springer. doi: 10.1007/978-3-031-37450-0_17.

[16] Kalchenko, V. (2018). Review of penetration testing methods for assesing the protection of computer systems. *Control, Navigation and Communication Systems*, 4(50), 109-114. doi: 10.26906/SUNZ.2018.4.109.

[17] Kelter, R. (2020). Bayesian and frequentist testing for differences between two groups with parametric and nonparametric two-sample tests. *Wiley Interdisciplinary Reviews: Computational Statistics*, 13(6), article number e1523. doi: 10.1002/wics.1523.

[18] Khambir, V. (2024). Automation of mobile application testing processes. *Computer-Integrated Technologies: Education, Science, Production*, 55, 213-224. doi: 10.36910/6775-2524-0560-2024-55-27.

[19] Khatami, S. (2020). *Evaluating catchment models as multiple working hypotheses under uncertainty*. (Doctoral dissertation, University of Melbourne, Melbourne, Australia). doi: 10.31237/osf.io/agcbd.
[20] Liley, J., & Wallace, C. (2018). Improved consistency in estimates of conditional false discovery rates increases power relative to both existing methods and parametric estimators. *BioRxiv*. doi: 10.1101/414326.
[21] Lu, Z.-H. (2020). An improved closed procedure for testing multiple hypotheses. *Statistics in Medicine*, 39(26), 3772-3786. doi: 10.1002/sim.8692.
[22] Ochieng, D. (2024). Multiple testing of interval composite null hypotheses using randomized p-values. *Statistical Papers*. doi: 10.1007/s00362-024-01591-9.
[23] Ramesh, Bhagyamma, G., & Wasiq, M.R. (2023). Exploring hypotheses in scientific inquiry: Challenges, formulation, and testing. *Vaikunta Baliga College of Law*, 8, 87-120.
[24] Raykov, T., Doebler, P., & Marcoulides, G.A. (2022). Applications of Bayesian confirmatory factor analysis in behavioral measurement: Strong convergence of a Bayesian parameter estimator. *Measurement Interdisciplinary Research and Perspectives*, 20(4), 215-227. doi: 10.1080/15366367.2021.2005959.
[25] Sekulovski, N., & Hoijtink, H. (2023). A default bayes factor for testing null hypotheses about the fixed effects of linear two-level models. *Psychological Methods*. doi: 10.1037/met0000573.
[26] Shportko, O.V., & Mushyn M.M. (2023). Using program testing opportunities on remote servers for comparing the efficiency of combinatory optimization methods. *Automation of Technological and Business Processes*, 15(1). doi: 10.15673/atbp.v15i1.2497.
[27] Tang, J., & Dette, H. (2024). Simultaneous semiparametric inference for single-index models. *ArXiv*. doi: 10.48550/arXiv.2407.01874.
[28] Wang, J.J. (2022). Approximate Bayesian estimator for the random-coefficients model. *Communication in Statistics – Simulation and Computation*, 53(6), 2579-2594. doi: 10.1080/03610918.2022.2093372.
[29] Woo, S. (2023). Design methodology – parametric accelerated life testing. In S. Woo (Ed.), *Design of mechanical systems* (pp. 305-327). Cham: Springer. doi: 10.1007/978-3-031-28938-5_7.
[30] Zhou, H., & Zou, H. (2024). A non-parametric box-cox approach to robustifying high-dimensional linear hypothesis testing. *ArXiv*. doi: 10.48550/arXiv.2405.12816.

Покращені методи пришвидшення A/B тестування для оцінки параметричних гіпотез: порівняння T-test з CUPED, CUPED++ та Bayesian Estimator

Артур Марков
Аспірант
Київський національний університет імені Тараса Шевченка
01033, вул. Володимирська, 60, м. Київ, Україна
https://orcid.org/0009-0000-5222-4397

Анотація. Метою дослідження було порівняння методів статистичного аналізу для покращення тестування альтернатив. У дослідженні оцінювалися чотири основні методи: класичний T-тест, традиційний і вдосконалений метод контрольованих експериментів з використанням передекспериментальних даних (CUPED), а також Байєсівський оцінювач. Основні результати включали демонстрацію процесу A/B тестування, а описані методи статистичного аналізу включали детальні характеристики та приклади використання. Моделювання та практичне застосування показали, що T-тест забезпечує високу точність при невеликих вибірках, але його ефективність знижується зі збільшенням розміру вибірки через високі вимоги до ресурсів. Калькулятор для цього методу продемонстрував ефективність у простих завданнях, але мав обмеження при роботі з великими даними. Традиційний метод CUPED показав підвищену точність завдяки варіаційній корекції, але його ефективність знижується при роботі з великими та складними наборами даних. Написана програма для цього методу показала свою ефективність у випадках, коли попередні дані добре представлені, але її можливості обмежені при обробці великих масивів даних. Вдосконалена версія забезпечила значне покращення як точності, так і швидкості обробки, особливо для великих наборів даних, завдяки вдосконаленому моделюванню та оптимізації. Результати роботи коду підтвердили, що цей метод є високоефективним для складних експериментів, особливо при обробці великих обсягів даних. Крім того, Байєсівський оцінювач продемонстрував високу точність завдяки інтеграції попередніх знань, але вимагав більше обчислювальних ресурсів і часу. Платформа, що використовувалася для цього методу, продемонструвала здатність враховувати невизначеність, але вимагала складних налаштувань моделі. Результати підкреслили важливість вибору відповідного методу статистичного аналізу залежно від масштабу та складності даних для забезпечення оптимальної точності та ефективності тестування

Ключові слова: статистичний аналіз; корекція варіацій; ефективність підходів; обробка даних; моделювання експериментів

ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ ТА КОМП'ЮТЕРНА ІНЖЕНЕРІЯ

Науково-технічний журнал

Том 21, № 3, 2024

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

Редагування англомовних текстів:
С. Воровський, К. Касьянов

Комп'ютерна верстка:
О. Глінченко

Підписано до друку 26.12.2024 р. Формат 60*84/8
Умовн. друк. арк. 15,5
Наклад 50 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. 21, No. 3, 2024

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

Editing English-language texts:
S. Vorovsky, K. Kasianov

Desktop publishing:
O. Glinchenko

Signed for print 26.12.2024. Format 60*84/8
Conventional printed pages 15.5
Circulation 50 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