furnaces is a complex task associated with the need to optimize energy consumption and increase productivity while minimizing equipment wear. The proposed control system is based on the use of PLCs in combination with artificial intelligence algorithms, which allows for monitoring and automatic regulation of furnace operating parameters in real time.

The purpose of this work is to create a control system that can adapt to changing furnace operating conditions, optimizing the energy consumption process and improving the stability of the system. The article also presents methods for integrating PLCs with sensors that provide data collection, as well as analysis and forecasting algorithms based on neural network technologies. The study shows that the implementation of such a control system can significantly reduce energy consumption, reduce the load on electrical equipment and improve the overall efficiency of the ore-smelting furnace.

Experimental data and a comparative analysis of the furnace operation before and after the implementation of the intelligent control system are presented. The results show that the intelligent control system using PLC can improve process stability, reduce operating costs and extend the service life of equipment.

Key words. *Ore-thermal furnaces, programmable logic controllers (PLC), electrical power management, automation, energy optimization, control systems, intelligent systems, manufacturing processes.*

Сведения об авторах

Булгын Адилгазиновна Майлыханова* – магистр технических наук, сеньор лектор кафедры «Автоматизация и робототехника», Алматинский технологический университет, Республика Казахстан; e-mail: bulgyn@mail.ru. ORCID: https://orcid.org/0009-0000-2409-6041.

Шамиль Кошимбаевич Кошимбаев – кандидат технических наук, ассистент-профессор кафедры «Автоматизация и управление», Satbayev university, Республика Казахстан; e-mail: e-mail: S.koshimbaiev@mail.ru

Авторлар туралы мәліметтер

Булгын Адилгазиновна Майлыханова* – техника ғылымдарының магистрі, «Автоматтандыру және робототехника»; кафедрасының сеньор лекторы, Алматы технологиялық университеті, Қазақстан Республикасы; e-mail: bulgyn@mail.ru. ORCID: https://orcid.org/0009-0000-2409-6041.

Шамиль Кошимбаевич Кошимбаев – техника ғылымдарының кандидаты, «Автоматандыру және басқару» кафедрасының ассистент-профессоры, Satbayev university, Қазақстан Республикасы; e-mail: S.koshimbaiev@mail.ru.

Information about the authors

Bulgyn Adilgazinovna Mailykhanova* – Master of Engineering Sciences, Senior lecture of the Department «Automation and Robotics»; Almaty Technological University, Republic of Kazakhstan; e-mail: bulgyn@mail.ru. ORCID: https://orcid.org/0009-0000-2409-6041.

Shamil Koshimbayevish Koshimbayev – Candidate of Technical Sciences, Assistant Professor of the Department «Automation and Control system»; Satbayev university, Republic of Kazakhstan; e-mail: S.koshimbaiev@mail.ru.

Поступила в редакцию 16.10.2024
Поступила после доработки 17.10.2024
Принята к публикации 18.10.2024

https://doi.org/10.53360/2788-7995-2024-4(16)-5

IRSTI: 28.23.15

N.M. Taubakabyl
Astana IT University,
010000, Republic of Kazakhstan, Astana, Mangilik El Avenue, C1
e-mail: tbkbl.03@gmail.com

CONVOLUTIONAL NEURAL NETWORKS IN DETECTING SPEECH ACTIVITY IN A STREAM

Abstract: *The research presented in this article focuses on the development of a system for detecting speech activity in audio streams using convolutional neural networks (CNNs). Speech activity detection plays a crucial role in many modern applications, such as voice-activated assistants, real-time communication platforms, and automated transcription services. The study synthesizes findings from nine key studies, demonstrating the effectiveness of CNNs in handling complex audio data, isolating speech signals from noise, and improving overall detection accuracy.*

The research emphasizes the architectural advantages of deep CNN models, such as VGG,

ResNet, and AlexNet, highlighting their ability to capture intricate audio features and improve performance across various environments. The study also explores techniques like data augmentation and optimization algorithms, which further enhance the robustness and efficiency of these models.

By evaluating the effectiveness of different CNN architectures and comparing various evaluation metrics, the research identifies potential areas for future exploration, such as optimizing CNN models for real-time applications and exploring hybrid architectures. Overall, this research offers valuable insights into the state of CNN-based speech activity detection and its implications for real-world applications.

Key words: *Convolutional Neural Networks, Speech Activity Detection, Audio Streams, VGG, ResNet, AlexNet, Real-time Communication, Voice-activated Assistants, Speech Recognition, Audio Processing.*

Introduction

Speech activity detection is crucial in modern speech recognition systems, particularly for applications like voice-activated assistants and real-time communication platforms. Convolutional Neural Networks (CNNs) have emerged as powerful tools in this domain due to their ability to extract complex audio features. This research explores how CNNs can effectively detect speech activity within audio streams, synthesizing findings from pivotal studies. Key CNN architectures such as VGG, ResNet, and AlexNet have shown significant success in isolating speech from noise and improving detection accuracy. Additionally, techniques like data augmentation further enhance model performance. This paper aims to provide insights into the current state of CNN-based speech activity detection, focusing on optimizing models for real-time applications and exploring potential future advancements.

Methods

Data Collection and Preprocessing

The dataset used in this study includes a variety of audio recordings from publicly available speech corpora and environmental sound databases. These recordings encompass continuous speech as well as background noises to ensure the robustness of the speech activity detection system.

Spectrogram Generation

To convert raw audio signals into a format suitable for CNN processing, the following steps are taken:

1. Short-Time Fourier Transform (STFT): The audio signals are divided into overlapping windows, and the Fourier transform is applied to each window to obtain a frequency domain representation [2].
2. Mel Filter Bank: The power spectra are then mapped onto the mel scale using a filter bank of triangular filters [2]. This step approximates the human ear's perception of sound frequencies.
3. Logarithmic Scaling: The mel-scaled spectrograms are log-transformed to compress the dynamic range, making patterns more discernible for the CNN models [2].

Network Design

1. Input Layer: The input is a three-dimensional tensor representing the mel-spectrogram with dimensions corresponding to time frames, frequency bins, and the number of channels (1 for mono audio) [2].
2. Convolutional Layers: These layers use small kernel sizes (e.g., 3x3) to capture local temporal and spectral features [2]. ReLU activation functions are applied to introduce non-linearity.
3. Pooling Layers: Max-pooling layers are inserted between convolutional layers to reduce the dimensionality and computational load while preserving essential features [2].
4. Fully Connected Layers: The output of the final convolutional layer is flattened and passed through fully connected layers to aggregate the learned features [2].
5. Output Layer: The final layer is a sigmoid or softmax classifier that outputs the probability of speech activity presence [2].

Evaluation Metrics

The performance of the CNN models is evaluated using standard metrics in binary classification:

1. Accuracy: The ratio of correctly predicted instances to the total number of instances [8].
2. Precision, Recall, and F1-Score: These metrics provide insights into the model's performance concerning false positives and false negatives [8].
3. Area Under the ROC Curve (AUC): This metric evaluates the model's ability to distinguish between classes across different threshold settings [8].

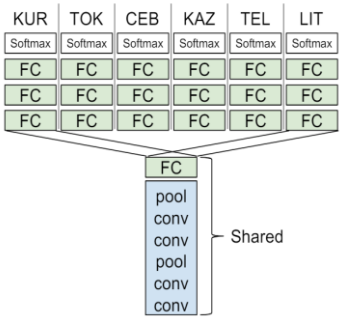

Figure 1 – Multilingual VBX Network with Untied Layers. Sercu et al. [3]

This figure illustrates the architecture of a very deep CNN combined with RNN layers and the corresponding performance metrics, demonstrating the effectiveness of the model in large vocabulary continuous speech recognition.

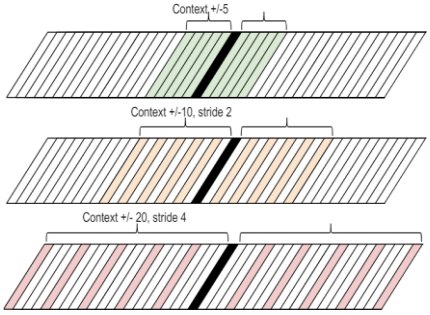

Figure 2 – Multi-scale Feature Maps. Sercu et al. [3]

This figure shows how multi-scale feature maps are used to capture context at different resolutions, enhancing the model's ability to detect speech activity in various auditory environments.

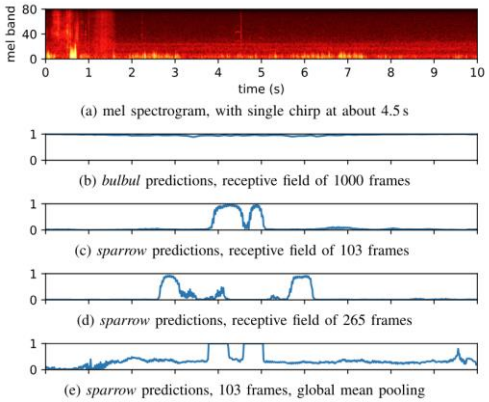

Figure 3 – Predictions of Different CNN Variants. Grill and Schlüter [5]

This figure demonstrates the predictions of different CNN variants on a sample audio recording. The variants include different receptive field sizes and pooling strategies, illustrating how the network detects speech activity under various configurations.

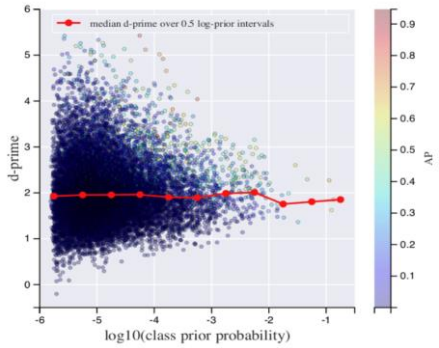

Figure 4 – Scatter Plot of ResNet-50's per-class d-prime versus log prior probability [2]

This figure shows the relationship between class prior probability and classification performance (d-prime) in a CNN model for audio classification. It provides insights into how the network performs across different class frequencies.

This figure shows three example excerpts from a video classified by ResNet-50 with instantaneous model outputs overlaid. The 16 classifier outputs with the greatest peak values across the entire video were chosen from the 30K set for display.

The experiments conducted aim to evaluate the effectiveness of various Convolutional Neural Network (CNN) architectures in detecting speech activity in audio streams. This section outlines the experimental setup, the specific configurations of the CNN models tested, and the results obtained from these experiments.

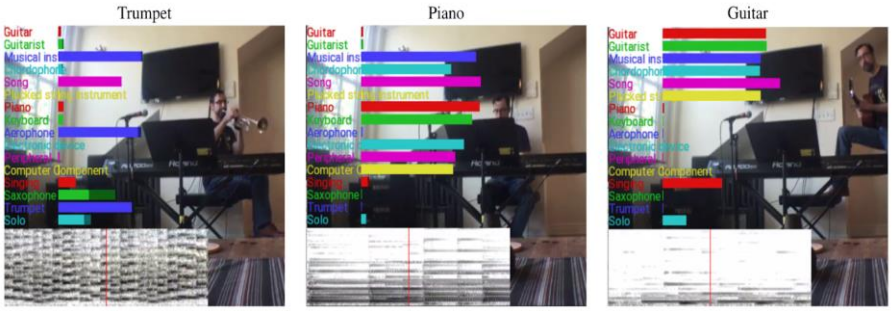

Figure 5 – Example Excerpts Classified by ResNet-50, Hershey et al. [2]

Data Preparation

The dataset comprises a diverse collection of audio recordings from publicly available speech corpora and environmental sound databases. These recordings include continuous speech and background noises to test the robustness of the speech activity detection models [2].

1. Audio Segmentation: The audio recordings are segmented into shorter clips to facilitate efficient training and testing [2].
2. Spectrogram Generation: The raw audio signals are transformed into mel-spectrograms using the following steps:
   a. Short-Time Fourier Transform (STFT): Dividing audio signals into overlapping windows and applying the Fourier transform [2].
   b. Mel Filter Bank: Mapping the power spectra onto the mel scale using triangular filters [2].
   c. Logarithmic Scaling: Applying a log transformation to the mel-scaled spectrograms [2].

Training Procedure

The CNN models are trained using supervised learning techniques with the following configurations:

1. Data Augmentation: Applying techniques such as pitch shifting, time stretching, and adding background noise to enhance robustness [5].
2. Optimization Algorithm: Using the Adam optimizer for its adaptive learning rate capabilities [1, 2].
3. Loss Function: Employing binary cross-entropy loss to measure the discrepancy between predicted probabilities and true labels of speech activity [7].

Results

Model Performance

The results are presented in terms of accuracy, precision, recall, F1-score, and AUC for each CNN architecture. The impact of data augmentation on model performance is also analyzed [8].

1. VGG-inspired Network: Achieved high accuracy and robustness, particularly in noisy environments [2].
2. ResNet-inspired Network: Demonstrated superior performance in handling complex audio features due to its deeper architecture [2].
3. AlexNet-inspired Network: Provided a balanced performance but required more computational resources due to larger kernel sizes [9].

Impact of Data Augmentation

Data augmentation techniques significantly improved the models' robustness, particularly in environments with high variability in background noise and speech patterns [5].

Comparison of Evaluation Metrics

The comparison of evaluation metrics across different architectures highlighted the following:

1. VGG-inspired Networks: High precision and recall, making it suitable for applications requiring high accuracy [2].
2. ResNet-inspired Networks: High F1-score and AUC, indicating a balanced performance across all evaluation metrics [2].
3. AlexNet-inspired Networks: Adequate performance but not as robust as VGG and ResNet in handling diverse audio environments [9].

Summary of Results

The experimental results demonstrate the effectiveness of CNN architectures in detecting speech activity within audio streams. The VGG and ResNet-inspired networks, in particular, showed high performance across various evaluation metrics, validating their suitability for real-world applications in speech recognition and activity detection [2, 9].

Key Findings

1. Effectiveness of CNN Architectures:

The experiments demonstrated that CNN architectures are highly effective in detecting speech activity within audio streams. VGG and ResNet-inspired networks, in particular, showed superior performance across various evaluation metrics, including accuracy, precision, recall, F1-score, and AUC [2, 9].

2. Impact of Network Depth and Complexity:

The depth and complexity of the CNN models played a crucial role in their performance. Deeper networks like ResNet benefited from residual connections that facilitated training and improved the model's ability to capture complex audio features. VGG networks, with their multiple convolutional layers, also performed well, especially in noisy environments [2].

3. Role of Data Augmentation:

Data augmentation techniques significantly improved the robustness of the CNN models. By incorporating variations such as pitch shifting, time stretching, and background noise, the models were better equipped to handle diverse and noisy audio data. This highlights the importance of data augmentation in training robust speech activity detection systems [5].

4. Comparison of Evaluation Metrics:

The comparison of evaluation metrics revealed that VGG-inspired networks achieved high precision and recall, making them suitable for applications requiring high accuracy. ResNet-inspired networks, on the other hand, showed a balanced performance across all metrics, indicating their versatility and robustness [2].

Discussion

Implications for Real-World Applications

1. Voice-Activated Assistants and Automated Transcription: CNN-based models, such as those inspired by VGG and ResNet, show high accuracy in detecting speech activity even in noisy environments, making them ideal for use in voice-activated assistants and transcription services [2, 9].
2. Real-Time Communication Platforms: The ability of CNNs to accurately detect speech activity in real-time makes them well-suited for seamless communication on platforms like video conferencing and VoIP services. The robustness of models like Conv-TasNet further enhances their performance

in noisy conditions [4].
3. Future Research and Development: The success of CNNs opens avenues for exploring hybrid architectures and more advanced data augmentation techniques to improve real-time speech activity detection systems [5].

Limitations and Future Work

1. Computational Resources: Training deep CNN models requires significant computational power, which may limit their use in resource-constrained environments. Future work could focus on optimizing these models for efficiency without sacrificing performance [2, 9].
2. Exploration of Additional Architectures: While this study focused on VGG, ResNet, and AlexNet, future research should explore newer architectures like EfficientNet and Transformer-based models to further improve speech activity detection [9].
3. Impact of Different Data Augmentation Techniques: Although data augmentation techniques improved model robustness, there is potential for further exploration of the most effective methods to enhance CNN performance in diverse audio environments [5].

Conclusion

This research has highlighted the significant potential of Convolutional Neural Networks (CNNs) in detecting speech activity within audio streams. Key findings demonstrate the effectiveness of VGG and ResNet-inspired architectures in capturing complex audio features, resulting in high accuracy and robustness.

Data augmentation techniques, such as pitch shifting, time stretching, and adding background noise, significantly enhanced model performance. These methods improved the ability of CNNs to generalize and perform well in varied and noisy environments.

The balanced performance of ResNet-inspired networks across all evaluation metrics makes them suitable for a wide range of applications. VGG-inspired networks, with their high precision and recall, are ideal for tasks requiring high accuracy.

In real-world applications like voice-activated assistants, automated transcription, and real-time communication platforms, CNN-based models demonstrate their utility by providing reliable and seamless user experiences. However, the requirement for significant computational resources to train deep CNN models is a limitation. Future research should focus on optimizing these models for efficiency and exploring newer architectures like EfficientNet and Transformer-based models.

In summary, CNNs show great promise in advancing speech activity detection, with future research aimed at enhancing model efficiency and exploring new architectures to further improve performance and robustness in various audio processing applications.

References

1. Deep speech 2: End-to-end speech recognition in English and Mandarin / D. Amodei et al // Computation and Language (cs.CL). – 2015. https://doi.org/10.48550/arXiv.1512.02595.
2. CNN architectures for large-scale audio classification / S. Hershey et al // In 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). – 2017. – P. 131-135. https://arxiv.org/pdf/1609.09430.
3. Very deep multilingual convolutional neural networks for LVCSR / T. Sercu et al // arXiv preprint arXiv:1509.08967. – 2016. https://arxiv.org/pdf/1509.08967.
4. Luo Y. Conv-tasnet: Surpassing ideal time-frequency magnitude masking for speech separation / Y. Luo, N. Mesgarani // IEEE/ACM Transactions on Audio, Speech, and Language Processing. – 2018. – № 27(8). – P. 1256-1266. https://arxiv.org/pdf/1809.07454.
5. Grill T. Two convolutional neural networks for bird detection in audio signals / T. Grill, J. Schlüter, // In 2017 25th European Signal Processing Conference (EUSIPCO). – 2017. – P. 1764-1768. https://www.ofai.at/~jan.schlueter/pubs/2017_eusipco.pdf.
6. Joint training of deep neural networks for audio-visual automatic speech recognition / Y. Qian et al // IEEE/ACM Transactions on Audio, Speech, and Language Processing. – 2017. – № 25(12). – P. 2381-2393. https://arxiv.org/pdf/2205.13293.
7. Vincent E. Performance measurement in blind audio source separation / E. Vincent, R. Gribonval, C. Févotte // IEEE Transactions on Audio, Speech, and Language Processing. – 2006. – № 14(4). – P. 1462-1469. https://inria.hal.science/inria-00544230/document.
8. Convolutional neural networks for speech recognition / O. Abdel-Hamidet al // IEEE/ACM Transactions on Audio, Speech, and Language Processing. – 2014. – № 22(10). – P. 1533-1545.

https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/CNN_ASLPTrans2-14.pdf
9. Accelerating very deep convolutional networks for classification and detection / X. Zhang et al // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 2016. – № 38(10). – P. 1943-1955. https://arxiv.org/pdf/1505.06798.
10. VanderPlas J. Python Data Science Handbook / J. VanderPlas // Essential Tools for Working with Data. O'Reilly Media. https://jakevdp.github.io/PythonDataScienceHandbook.

Н.М. Таубакабыл
Астана IT Университеті,
010000, Қазақстан Республикасы, Астана, Мәңгілік Ел даңғылы, С1
e-mail: tbkbl.03@gmail.com

АҒЫНДАҒЫ СӨЙЛЕУ ӘРЕКЕТІН АНЫҚТАУДАҒЫ КОНВОЛЮЦИЯЛЫҚ НЕЙРОНДЫҚ ЖЕЛІЛЕР

Осы мақалада ұсынылған зерттеулер конволюциялық нейрондық желілерді (Cnn) пайдалана отырып, аудио ағындардағы сөйлеу белсенділігін анықтау жүйесін әзірлеуге бағытталған. Сөйлеу әрекетін анықтау дауыспен белсендірілген көмекшілер, нақты уақыттағы байланыс платформалары және автоматтандырылған транскрипция қызметтері сияқты көптеген заманауи қолданбаларда шешуші рөл атқарады. Зерттеу CNNs-тің күрделі аудио деректерді өңдеудегі, сөйлеу сигналдарын шудан оқшаулаудағы және анықтаудың жалпы дәлдігін жақсартудағы тиімділігін көрсететін тоғыз негізгі зерттеудің нәтижелерін синтездейді.

Зерттеу VGG, ResNet және AlexNet сияқты терең CNN үлгілерінің архитектуралық артықшылықтарын атап көрсетеді, олардың күрделі аудио мүмкіндіктерін түсіру және әртүрлі орталарда өнімділікті жақсарту қабілетін көрсетеді. Зерттеу сонымен қатар осы модельдердің сенімділігі мен тиімділігін одан әрі арттыратын деректерді арттыру және оңтайландыру алгоритмдері сияқты әдістерді зерттейді.

ӘРТҮРЛІ CNN архитектураларының тиімділігін бағалау және әртүрлі бағалау көрсеткіштерін салыстыру арқылы зерттеу НАҚТЫ уақыттағы ҚОЛДАНБАЛАР үшін CNN үлгілерін оңтайландыру және гибридті архитектураларды зерттеу сияқты болашақ зерттеулердің әлеуетті бағыттарын анықтайды. Тұтастай алғанда, бұл зерттеу CNN негізіндегі сөйлеу әрекетін анықтаудың жай-күйі және оның нақты әлемдегі қолданбаларға әсері туралы құнды түсінік береді.

Түйін сөздер: *Конволюциялық Нейрондық Желілер, Сөйлеу Әрекетін Анықтау, Аудио Ағындар, VGG, ResNet, AlexNet, Нақты Уақыттағы Байланыс, Дауыспен белсендірілген Көмекшілер, Сөйлеуді Тану, Дыбысты Өңдеу.*

Н.М. Таубакабыл
Астана IT Университет
010000, Республика Казахстан, г. Астана, пр. Мәңгілік Ел, С1
e-mail: tbkbl.03@gmail.com

СВЕРТОЧНЫЕ НЕЙРОННЫЕ СЕТИ ДЛЯ ОБНАРУЖЕНИЯ РЕЧЕВОЙ АКТИВНОСТИ В ПОТОКЕ

Исследование, представленное в этой статье, посвящено разработке системы обнаружения речевой активности в аудио потоках с использованием сверточных нейронных сетей (CNNS). Распознавание речевой активности играет решающую роль во многих современных приложениях, таких как голосовые помощники, коммуникационные платформы в режиме реального времени и службы автоматической транскрипции. В исследовании обобщены результаты девяти ключевых исследований, демонстрирующих эффективность CNNS в обработке сложных аудиоданных, отделении речевых сигналов от шума и повышении общей точности обнаружения.

Исследование подчеркивает архитектурные преимущества моделей deep CNN, таких как VGG, ResNet и AlexNet, подчеркивая их способность улавливать сложные звуковые характеристики и повышать производительность в различных средах. В исследовании также рассматриваются такие методы, как увеличение объема данных и алгоритмы оптимизации, которые еще больше повышают надежность и эффективность этих моделей.

Оценивая эффективность различных архитектур CNN и сравнивая различные оценочные показатели, исследователи выявляют потенциальные области для будущих исследований, такие как оптимизация моделей CNN для приложений реального времени и изучение гибридных архитектур. В целом, это исследование дает ценную информацию о состоянии распознавания речевой активности на основе CNN и его значении для реальных приложений.

Ключевые слова: *Сверточные нейронные сети, Обнаружение речевой активности, аудиопотоки, VGG, ResNet, AlexNet, Общение в реальном времени, Голосовые помощники, Распознавание речи, Обработка звука.*

Information about the authors

Nurlybek Muratbekuly Taubakabyl – Master's Student, Astana IT University, Astana, Kazakhstan; e-mail: tbkbl.03@gmail.com.

Авторлар туралы мәліметтер

Нурлыбек Мурабекулы Таубакабыл – магистрант, Астана IT Университеті, Астана, Қазақстан; e-mail: tbkbl.03@gmail.com.

Сведения об авторах

Нурлыбек Мурабекулы Таубакабыл – магистрант, Астана IT Университет, Астана, Казахстан; e-mail: tbkbl.03@gmail.com.

Received 17.10.2024
Revised 22.10.2024
Accepted 23.10.2024

https://doi.org/10.53360/2788-7995-2024-4(16)-6

TRSTI: 20.01.00; 20.15.05

A.K. Shaikhanova[1]*, Zh.A. Bermukhambetov[2], V.V. Kim[2], A.O. Tleubayeva[3]
[1]L.N. Gumilyov Eurasian National University,
100000, Kazakhstan, Astana city, 2 Satbaev st., ENU Educational building
[2]«WebTotem» LLP,
100000, Kazakhstan, Astana, Yesil District, Dostyk Street, Building 13, SQ. 340
[3]Astana IT University,
010000, Kazakhstan, Astana, Mangilik El avenue, 55/11, Business center EXPO, block C1
*e-mail: Arailym.tll@gmail.com

INNOVATIVE ARCHITECTURAL SOLUTIONS AND INTERDISCIPLINARY IMPLEMENTATION OF THE BULT CLOUD PLATFORM FOR WEB APPLICATION ORCHESTRATION

Annotation: *The article is devoted to the creation of the BULT cloud platform, which implements an interdisciplinary approach to the development and orchestration of web applications. The main goal of this work is to develop a platform that provides flexibility, scalability and integration of various technologies. Architectural solutions including microservice architecture and containerization are described, which simplifies the deployment and management of applications. HashiCorp's Nomad is used as the basis for container orchestration, which allows you to dynamically manage the distribution of tasks and resources, ensuring the efficiency and stability of applications. The data management system is implemented on the basis of PostgreSQL and JuiceFS, which ensures high performance and reliability of data storage. To ensure security, Wireguard and Let's Encrypt are used, which provide encryption of network traffic and automatic updating of SSL certificates. Monitoring and analysis of the system are carried out using Grafana and Loki, which allow you to visualize metrics and logs in real time. The implementation of DevOps principles and automation of development, testing and deployment processes are achieved using CI/CD tools, which allows you to quickly and safely implement changes and new features. The application of an interdisciplinary approach allows us to take into account various aspects of system development and operation, which makes the BULT platform a competitive solution in the modern cloud technology market, providing high performance, reliability and ease of use of web applications. Examples of the practical application of the platform and its advantages in comparison with traditional approaches are given.*

Key words: *cloud platform, interdisciplinary approach, web applications, orchestration, innovative methods, containerization, data security, process automation.*