UDC 004 **doi: 10.32620/reks.2025.3.13**

Andrii NIKITENKO, Yevhen BASHKOV

Donetsk National Technical University, Drohobych, Ukraine

USING A HYBRID ATTENTION MECHANISM AS A METHOD TO IMPROVE THE EFFICIENCY OF NETWORK INTRUSION DETECTION SYSTEMS

*The **subject matter** of this article is a HybridAttention mechanism integrated into a deep neural architecture for Network Intrusion Detection Systems (NIDS). This study **aims** to develop and study a HybridAttention mechanism based on a combination of global (self-attention) and local (dynamic local attention) models to improve the quality of traffic classification in real-time NIDS. The **tasks** to be solved are as follows: analyzing the applicability of existing attention mechanisms in network intrusion detection; integrating various attention types into a CNN-BiGRU architecture; developing a HybridAttention mechanism based on dynamic window alignment; optimizing the model using Optuna; and experimentally evaluating its performance on benchmark datasets using standard classification metrics. The **methods** used are: deep learning modeling with CNN-BiGRU architecture, integration of various attention mechanisms, including a novel HybridAttention, hyperparameter optimization using Optuna, and performance evaluation based on standard classification metrics. The **results** of this study show that the proposed HybridAttention mechanism demonstrates superiority over individual types of attention in all key metrics. The model achieved up to 99.85% accuracy on the NSL-KDD dataset training data and demonstrated strong generalization on the UNSW-NB15 dataset, achieving up to 98.06% accuracy in multi-class classification and up to 99.20% in binary classification. The proposed model also outperformed state-of-the-art approaches for processing unbalanced data and detecting various types of attacks. **Conclusions**. The scientific novelty of the results obtained is as follows: a HybridAttention mechanism combining self-attention and dynamic local attention was developed to enhance sequential pattern recognition in network traffic; the CNN-BiGRU architecture was improved by integrating multiple attention modules; systematic hyperparameter optimization using Optuna improved generalization on imbalanced data; and the proposed model outperformed existing approaches on benchmark datasets in detecting both known and novel cyberattacks.*

Keywords: *deep learning; attention mechanism; network intrusion detection system; HybridAttention mechanism; dynamic local attention.*

1. Introduction

1.1 Motivation

The rapid growth of cybercrime, coupled with the increasing complexity and volume of network traffic, presents critical challenges to the protection of information systems. Traditional signature-based network intrusion detection systems (NIDS) are no longer sufficient to handle zero-day attacks, evolving threat patterns, or the need for real-time response [1]. Consequently, the cybersecurity domain is shifting toward adaptive, intelligent detection systems that can dynamically analyze traffic and recognize both known and novel intrusions.

Deep learning (DL) approaches have emerged as powerful tools in this field due to their capacity for automated feature extraction and sequence modeling. Architectures that integrate Convolutional Neural Networks (CNNs) and Bidirectional Gated Recurrent Units (BiGRUs) have shown particular promise in capturing both spatial and temporal characteristics of traffic data [5]. However, these models may struggle to prioritize important features in noisy, high-dimensional inputs, which can hinder their detection performance.

Attention mechanisms offer a compelling solution to this limitation by allowing the model to selectively focus on relevant parts of the input data during processing. Initially developed for natural language processing tasks, attention mechanisms have since been applied in intrusion detection to improve classification accuracy and model interpretability. Their integration into NIDS has shown promising results, enabling better detection of complex attack patterns and reducing false positives.

Nevertheless, many existing studies evaluate only a single attention mechanism and often use a limited number of datasets or classification types. Comparative analysis across different attention variants (e.g., global, local, and self-attention) and minimal exploration of hybrid approaches are lacking. This constrains the adaptability and robustness of the proposed models under real-world network conditions.

Creative Commons Attribution NonCommercial 4.0 International

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

This study investigates a *HybridAttention* mechanism that combines self-attention with dynamically adjustable local attention within a CNN-BiGRU architecture to address these challenges. The proposed method aims to enhance the detection of sophisticated and emerging threats while maintaining generalizability across different datasets and attack classes.

1.2 State of art

In recent years, attention mechanisms have been widely used in deep learning tasks, including image captioning, text classification, and speech recognition, and they have been increasingly applied to intrusion detection.

Attention is an important mechanism that can be used in a variety of deep learning models in different fields and tasks. Different attention mechanisms are explained using a general model of attention, a unified notation, and a comprehensive taxonomy of attention mechanisms [1]. The attention mechanism addresses information overload by allocating resources to process critical data and optimizing limited computing power. It is widely used in tasks such as image captioning, text classification, translation, action recognition, speech recognition, recommendations, and graph analysis [2]. Recently, the use of attention mechanisms along with DL methods to detect network intrusions has been increasingly popular.

In [3], a model for intrusion detection based on CNN-BiLSTM-Attention is proposed. A v2 loss function is introduced to address class imbalance by prioritizing minority class data during training. The model was tested on the NSL-KDD, UNSW-NB15, and CIC-DDoS2019 datasets, and it achieved accuracy of 99.79%, 88.84%, and 99.84%, recall of 99.83%, 98.52%, and 99.99%, and FPR of 0.17%, 1.82%, and 0.00%, respectively. However, it lacks diverse evaluation metrics and comprehensive comparisons for UNSW-NB15.

In [4], an intrusion detection model based on a combination of bidirectional long-short-term memory (BiLSTM) and an attention mechanism based on a distributed ensemble architecture for IoT network security for traffic data classification is proposed and tested on the UNSW-NB15 test dataset. The model is evaluated using the metrics of accuracy 99.05%, precision 98.96%, recall 99.36%, and F1-score 99.15%. The disadvantage of this study is the lack of results of multi-class classification and testing the model on only one dataset.

Paper [5] proposes the SSAE-BiGRU-Att intrusion model based on a combination of a stacked sparse autoencoder (SSAE), BiGRU, and an attention mechanism for traffic classification was proposed in the study [5]. Experiments were conducted on the UNSW-NB15 dataset, with results in terms of accuracy 98.68 %, precision 99 %, recall 99 %, and FPR 0.0132 %. The disadvantage, as in the previous work, is the lack of demonstration of the model's multi-class classification and the availability of only one dataset.

A new hierarchical CNN-Attention network (CANET) is proposed in [6]. The model combines CNN and Attention into a CA unit for spatio-temporal feature extraction, with a v2 loss function for balanced training. Tested on NSL-KDD and UNSW-NB15 for multi-class and binary classification, it achieves 99.77% accuracy, 99.72% recall, and 0.18% FPR on NSL-KDD and 89.39% accuracy, 98.93% recall, and 0.87% FPR on UNSW-NB15. However, it lacks precision and F1-score metrics for a more detailed evaluation.

In [7], a hierarchical attention mechanism called HAGRU (Hierarchical Attention Gated Recurrent Unit) is proposed. The model enhances detection by leveraging features from three hierarchies and optimizing resource use by using attention to focus on malicious data flows. Tested on NSL-KDD, CIC-IDS2017, and CES-CIC-IDS2018, it evaluates individual classes but lacks a comparative analysis with the HAGRU model.

In [8], an Enhanced Hybrid Intrusion Detection System (EHID-SCA) that integrates channel and spatial attention within a CNN-based deep learning framework is proposed. The model targets intrusion detection in wireless sensor networks (WSNs), aiming to extract spatial-temporal features efficiently while improving detection accuracy and interpretability. The comparison was performed on the UNWS-NB15, NSL-KDD, and KDDcup99 datasets. The results show an increase in accuracy for all datasets. However, the proposed method focuses primarily on WSNs and may not be well suited to diverse NIDS scenarios, especially those involving high-speed or heterogeneous network environments.

In [9], an improved BiGRU-Inception-CNN model (NIDS-BAI) is proposed for IIoT intrusion detection. It incorporates an attention mechanism, BiGRU for bidirectional temporal feature learning, and Inception-CNN for multi-scale spatial features. This study addresses class imbalance using a hybrid sampling strategy (ADASYN + RENN + LOF) and applies Pearson correlation with Random Forests for feature selection. A comparison of the model's performance was made using the CIC-IDS0217, CIC IoT 2023, and Edge-IIoTset datasets, which showed that the model performed better on the CIC IoT 2023 and Edge-IIoTset datasets. Although the model shows strong performance across multiple datasets, its complexity and high computational overhead may hinder real-time deployment in resource-constrained IIoT environments.

In [10], the SA-DCNN model combines self-

attention with deep convolutional neural networks to detect intrusions in IIoT networks. The model emphasizes the significance of individual features and reduces redundancy via two-step data cleaning and feature filtering based on mutual information. Despite achieving superior results on IoTID20 and Edge-IIoTset datasets, heavy preprocessing and the absence of recurrent components might limit the model's ability to model long-term dependencies in network traffic.

In [11], an optimized intrusion detection system called AIMHCNN-IDS-VANET is proposed, which integrates a multi-head convolutional neural network enhanced with attention mechanisms for detecting various types of cyberattacks in Vehicular Ad Hoc Networks (VANETs). The proposed approach uses the CAN_HCRL_OTIDS dataset and applies tanh-based normalization before classification into DoS, fuzzy, impersonation, and normal classes. The Capuchin Search Optimization Algorithm is employed to fine-tune the model's weights to improve classification accuracy. Although the method achieves improved accuracy, precision, and specificity over existing techniques, it lacks in-depth comparative analysis with alternative optimization algorithms and does not explore the model's adaptability to real-time or resource-constrained VANET environments.

Thus, the reviewed publications focus on combining architectures with attention mechanism variations to improve the efficiency of network intrusion detection. Some papers [4], [5] do not provide results of multiclass classification, and studies [3], [4] use only one dataset to test the model. It should be noted that there is an urgent need to find new solutions to improve the effectiveness of network intrusion detection, particularly with the help of attention mechanisms, which makes this topic relevant for further research.

1.3. Objectives and the methodology

This study aims to improve the performance of NIDS by developing and evaluating a HybridAttention mechanism that combines dynamic local attention and self-attention within a CNN-BiGRU architecture. The objectives of this study are as follows:

1. To review existing attention mechanisms and justify their application in the field of network intrusion detection systems;

2. To enhance a previously proposed CNN-BiGRU-Attention [12] model by integrating various attention mechanisms (global, local, and self-attention) and compare their performance using the NSL-KDD and UNSW-NB15 datasets;

3. To evaluate the proposed HybridAttention mechanism and compare its effectiveness with modern state-of-the-art approaches in terms of classification accuracy, precision, recall, and F1-score.

The research methodology comprises the following steps:

Dataset selection and preparation. The research uses two public datasets—NSL-KDD and UNSW-NB15—chosen for their relevance in NIDS performance benchmarking. Data preprocessing includes the normalization of continuous features and categorical variable label encoding. Stratified train-test splits were applied to ensure class balance for binary and multiclass classification tasks.

Model configuration and design of the attention mechanism The baseline model architecture combines convolutional layers for spatial feature extraction and BiGRU for temporal sequence learning. Four types of attention mechanisms — global, local with monotonic and predictive alignment, and self-attention—were implemented. Additionally, a HybridAttention mechanism was proposed, combining MorphingLocalAttention with self-attention, enabling dynamic adjustment of the attention window during training.

The experimental procedure and evaluation. The experiments were conducted in the Google Colab Pro environment using Python 3.10.12, TensorFlow 2.15.0, and Keras 2.15.0. Hyperparameter optimization was performed using the Optuna framework. Each model was evaluated under identical conditions using the following standard metrics: accuracy, precision, recall, and F1-score. Performance was compared across training and testing datasets for both classification modes.

Analysis and interpretation of data The results were organized into tables and figures that highlight the performance of each attention mechanism. The generalization ability of the models and their robustness across datasets were given special attention. A comparative analysis with recent models was performed to demonstrate the practical relevance of the proposed hybrid approach.

The article is structured as follows. Section 2 presents the materials and methods, including the design of the attention mechanisms and model configuration. Section 3 presents the results and their comparative analysis. Section 4 presents a case study demonstrating the practical deployment of the proposed approach. Section 5 concludes the study, summarizing key contributions and outlining directions for future research.

2. Materials and methods of research

The mechanisms of deep attention can be divided into soft attention (global attention), hard attention (local attention) and self-attention [13].

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

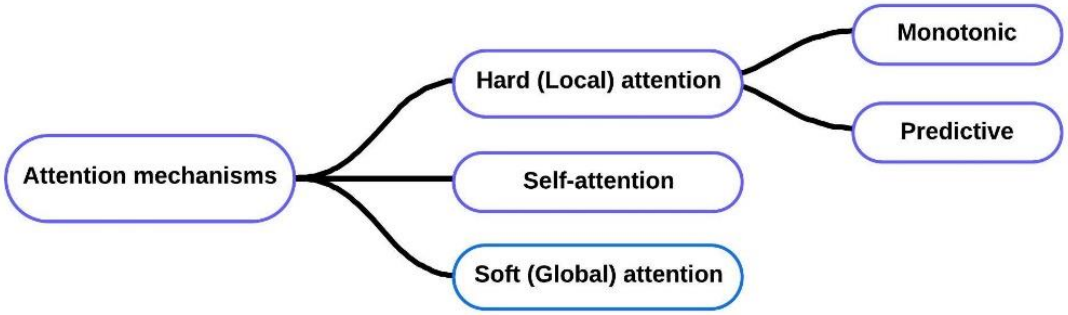

Fig. 1. Types of attention mechanisms

Self-attention. Self-attention, also known as scaled dot-product attention [14], is a fundamental concept in DL and natural language processing. It plays a key role in tasks such as machine translation, text summarization, and sentiment analysis.

The basic idea of self-attention is to calculate similarity scores between each input sequence element and all other elements. These scores are then used as weights to be applied to the input sequence representation. This allows the model to automatically focus on the most relevant elements of the input sequence and consider their relationship when learning the representation.

Self-attention acts as a conductor, providing contextual insight to the model, allowing it to understand individual elements in a sequence and adjust their influence on the final outcome. This type of organization is invaluable in language processing tasks, where the meaning of a word depends on its counterparts in a sentence or document. Self-attention is based on the quartet of queries Q, keys K, values V and self-attention itself. Mathematically, self-attention can be expressed as vector $X = [x_1, x2, \ldots, x_n]$, where x_i – is a vector representing the i-th element in the sequence. The output of self-attention is calculated as follows:

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

where Q, K, V – query, key, and value matrices, respectively. The similarity function calculates the dot product between the query and each key to obtain the weight, which is much faster and more compact in practice [14], i.e., fewer training parameters are needed. Finally, the softmax function is applied to normalize and assign these weights along with the corresponding values to obtain the final attention score.

Self-attention can improve the interpretability of detected features and reduce the semantic gap between artificial intelligence detectors and security analysts. In addition, this mechanism can help security analysts obtain attention scores to select important features for correlation analysis, thus filtering out false alarms to effectively identify and respond to genuine attacks on time. It should also be noted that by utilizing the self-attention mechanism, the model can offer better capabilities to remember long-term dependencies existing in the record to mitigate the problem of gradient vanishing and performance degradation, thereby achieving higher accuracy [15].

Soft (global) attention. Soft attention is a common technique in machine learning, particularly in computer vision and natural language processing, to focus on relevant parts of the input data. First introduced in [16], it uses the weighted average of all keys to construct a context vector. For soft attention, the attention module is differentiated with respect to the inputs, so that the entire system can be trained using standard backpropagation methods.

Soft attention works in both spatial and temporal contexts. Spatially, it highlights or weights relevant features, whereas temporally, it adjusts sample weights in rolling time windows based on their varying contributions. Despite being deterministic and differentiable, soft mechanisms have a high computational cost for large input data [13].

Simultaneously, [17] proposed a global attention identical to the soft attention, with the differences only in the simplification of the calculations. The main idea of global attention is to consider all the encoder's hidden states when deriving the context vector c_t. In this type of model, the alignment vector is of variable length a_t, whose size is equal to the number of time steps on the source side. The alignment vector is derived by comparing the current hidden state of the target h_t with each hidden state of the source $\bar{h}_s$:

$$a_t(s) = \text{align}(h_t, \bar{h}_s) = \frac{\exp(\text{score}(h_t, \bar{h}_s))}{\sum_{s'} \exp(\text{score}(h_t, \bar{h}_s))}$$

where score – a content-based function for which four different alternatives can be considered:

$$\text{score}(h_t, \overline{h_s}) = \begin{cases} h_t^T \bar{h}_s \text{ (dot)} \\ \dfrac{h_t^T \overline{h_s}}{\sqrt{H}} \text{ (scaled} - \text{dot)} \\ h_t^T W_a \bar{h}_s \text{ (general)} \\ v_a^T \tanh(W_a [h_t; \bar{h}_s]) \text{ (concat)} \end{cases}, \quad (1)$$

The score function is an important part of the attention model because it determines the matching or combination of keys and queries. In (1), some common score functions are listed, the most common of which are the additive (concat) score function (as an alignment model in recurrent neural network search) [17] and the less computationally expensive multiplicative (dot product) score function [17]. In [18], an empirical comparison between these two evaluation functions was made and it was found that the parameterized additive attention mechanism slightly but consistently outperforms the multiplicative one. Moreover, in [14], a variant of the multiplicative evaluation function was proposed by adding a scaling factor $\frac{1}{\sqrt{H}}$, where H – the dimensionality of the hidden state of the source, motivated by the fear that the softmax function may have a very small gradient with large input data, which prevents effective learning. Also, in [17], a general score function was presented. The general score function extends the concept of multiplicative attention by introducing a matrix parameter W, a learnable system that can be applied to keys and queries with different representations.

Early attempts to build attention-based models used a location-based function [17], in which alignment estimates are computed only based on the hidden state of the target h_t, as shown below:

$$a_t = \text{softmax}(W_a h_t) \text{(location)}, \quad (2)$$

Given the alignment vector as a weight, the context vector c_t is calculated as a weighted average over all hidden source states.

However, a previous study [12], in which a network intrusion detection system based on the attention mechanism was created, used soft attention [7]:

$$u_t = \tanh(W_w h_t + b_w), \quad (3)$$

$$a_t = \frac{\exp(u_t^T u_w)}{\sum_t \exp(u_t^T u_w)}, \quad (4)$$

$$V = \sum_t a_t h_t, \quad (5)$$

where h_t – hidden state, W_w – matrix of attention weights b_w – attention bias, a_t – weighting matrix, V – attention vector weighted by the attention mechanism.

The differences between the two approaches described above are as follows:

1. In formula (3) to calculate u_t the hidden state vector is calculated using the weighting matrix W_w input hidden state h_t and the added bias b_w. This is similar to the preprocessing step to obtain an intermediate hidden state, but the abovementioned score function depends on two hidden states, not just one h_t;

2. In formula (4), the alignment vector a_t is calculated using the softmax function of the vector $u_t^T u_w$. This is similar to the attention alignment calculation, at the same time the article above describes the alignment calculation a_t by comparing the current hidden state h_t with each hidden state $\overline{h_s}$ using leveling weights a_t .

3. In formula (5), the global context vector is calculated V as a weighted sum of all hidden states h_t using leveling weights a_t, This is similar to how the article describes the calculation of the global context vector c_t, which is also calculated as a weighted sum of all hidden states $\bar{h}_s$ з using leveling scales a_t.

Hard (local) attention. Unlike widespread soft attention, which considers all elements of an input sequence, hard attention [19] selects only a certain part of the elements and discards the rest. Hard attention focuses the model on key elements, making it effective for tasks with limited sequences. However, it is inefficient in sampling and optimization due to its combinatorial nature, which often necessitates reinforcement learning methods. While reducing computational costs, these approaches rely on stochastic processes due to the lack of a clear selection policy.

The local attention discussed in [17] is inspired by a compromise between the soft and hard attention models proposed in [18], where soft attention refers to a global attention approach in which weights are “softly” placed over all regions of the original image. Although the hard attention model is less expensive in terms of inference time, it is not differentiated and requires more sophisticated training methods, such as variance reduction or reinforcement learning.

The local attention concept eliminates the cost of global attention by focusing on a small subset of tokens in the hidden state set obtained from the input sequence. This window is proposed as $[p_t - D, p_t + D]$, where D –

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

empirically selected window width, which ignores positions that cross the sequence boundaries [19]. Aligned position p_t, is determined by:

a) Monotonic alignment – assumes that the source and target sequences are approximately aligned in a monotonic manner. Monotonic alignment is identical to global attention, except that the vector a_t has a fixed length and is shorter:

$$p_t = t;$$

b) Predictive alignment – instead of assuming monotonic alignment, the proposed model predicts the aligned position. It is similar to monotonic alignment except that it dynamically computes p_t and a truncated Gaussian distribution of alignment weights is used:

$$p_t = S \cdot \text{sigmoid}(v_p^T \tanh(W_p h_t)),$$

where W_p and v_p are the model parameters that will be studied to predict positions. S – is the length of the original sentence. Because of the sigmoid, $p_t \in [0, S]$. To promote alignment points in the vicinity p_t, is a Gaussian distribution centered around p_t. The alignment weights are defined as follows:

$$a_t(s) = \text{align}(h_t, \overline{h_s}) \exp\left(-\frac{(s - p_t)^2}{2\sigma^2}\right),$$

The same alignment function is used as in equation (1), and the standard deviation is empirically defined as $\sigma = \frac{D}{2}$. It should be noted [17], that p_t is a real number, while *s* is an integer within the window centered at the point p_t.

We propose a HybridAttentionmechanism (Fig. 2) that combines two different approaches: modified local attention and self-attention.

Modified (dynamic) local attention. The window width D is treated as a hyperparameter that is optimized throughout the model training process using MorphingLocalAttention. This means that the window width changes dynamically after each epoch, allowing the model to better adapt to the specific task requirements. This approach differs from traditional methods in which the window width is empirically chosen. The model determines the best value during training instead of manually selecting the optimal window width, which leads to better generalization and simplifies the model training process.

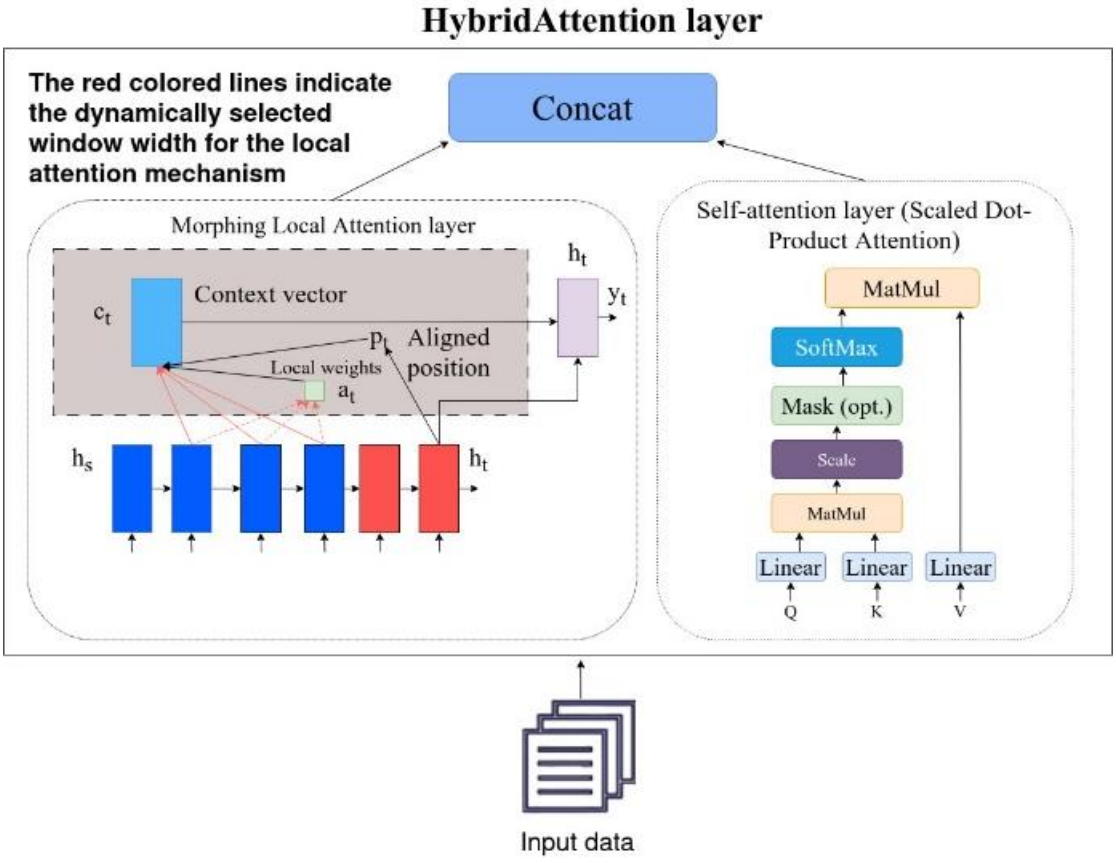

Fig. 2. HybridAttention mechanism

This is accomplished using a sigmoid activation function that limits the output values to a range of 0 to 1. Subsequently, the value is scaled by adding a minimum and maximum window width to ensure that the window width stays within the desired range, and the absolute value of the integer of this result is taken. This ensures that the window width is always a positive integer.

The disadvantage of this modification is that it can only be used with predictive alignment operations because the window width is treated as a tensor, which cannot be used in monotonic alignment operations.

Self-attention. Using self-attention, the model focuses on different parts of the input data and calculates weights for each element based on the relationships between queries, keys, and values, which significantly increases the efficiency of detecting complex patterns in the data.

3. Results and Discussion

AA series of experiments were conducted on the NSL-KDD [20] and UNSW-NB15 [21] datasets to test the effectiveness of different attention mechanisms in the CNN-BiGRU-Attention model proposed earlier [12].

The choice of NSL-KDD and UNSW-NB15 were chosen as benchmark datasets because of their complementary strengths in terms of representativeness and quality. NSL-KDD, an improved version of KDD Cup 1999, removes redundant records and balances class distribution, thereby reducing bias toward majority classes. It provides training (KDDTrain+) and testing (KDDTest+) subsets covering normal traffic and four attack types (DoS, Probe, R2L, U2R). Although it is widely adopted due to its structure and accessibility, its attack scenarios are outdated and less representative of modern threats. Conversely, UNSW-NB15 was created in the Cyber Range Lab at UNSW Canberra using IXIA PerfectStorm to emulate contemporary user activity and synthetic attacks. It contains 49 attributes and nine attack categories (e.g., Fuzzers, DoS, Exploits, Reconnaissance, Shellcode) captured with tcpdump and processed with Argus, Bro-IDS, and feature extraction algorithms. While based on controlled experiments and not fully reflective of real-world variability, UNSW-NB15 offers richer feature diversity and improved balance. Together, the two datasets provide a comprehensive basis for evaluating the proposed model's accuracy and generalization.

Experimental environment. The Google Colab Pro cloud environment was used to conduct the experiments as part of the study. This environment provides access to powerful computing resources, making it ideal for performing resource-intensive tasks. Python version 3.10.12, Tensorflow version 2.15.0, and Keras version 2.15.0 were used in all experiments. The Optuna library was chosen to optimize the hyperparameters. This open-source library, written in Python, uses Bayesian optimization algorithms to find the best hyperparameter sets for machine learning. The other basic parameters of the model are as follows: dropout 0.5, number of epochs 50, and Adam as the optimizer for parameter training. The model uses the loss function categorical_crossentropy for multiclass classification and binary_crossentropy for binary classification. Table 1 shows the hyperparameters selected for the model.

Table 1

Ranges for optimizing the hyperparameters of the CNN-BiGRU-Attention model

Parameter	**Value**
Cnn_layer_filters	[16,128]
Cnn_layer_kernel_size	[2,5]
Gru_layer_units	[8,128]
Dense_layer_units	[16,256]
Activation_dense_layer	['tanh', 'relu']
Activation_cnn_layer	['tanh', 'relu']
Pool_size	[2,6]
Batch_size	[32,512]
Learning_rate	[0.0001, 0.001, 0.01]
Sequence_length	[1,10]
Self_attention_units	[32,256]
Window_width	[1,4]

Evaluation metrics. Attention mechanisms are evaluated using the accuracy, reliability, recall, and F1-score metrics [22]. Accuracy determines the proportion of items correctly identified in the total amount of data. Reliability measures the accuracy of the model's positive predictions and is calculated as the ratio of the number of correctly predicted positive observations to the total number of predicted positive outcomes. Recall measures the ability of the model to capture all positive outcomes by calculating the ratio of correctly predicted positive observations to actual positive observations. The F-measure expresses the harmonic mean of accuracy and reproducibility, which is a balanced measure that considers both FPs and FNs.

First, we consider the performance of the model using attention mechanisms on the NSL-KDD and UNSW-NB15 datasets for multiclass classification (Tables 2-5) and binary classification (Tables 6-9). Experiments were conducted for 21 variants of attention mechanisms and score functions. In each table, only the top 5 options are shown, including global attention, local attention with monotonic alignment, predictive alignment, and a HybridAttention mechanism with the scaled_dot function, as it showed the best results among the other alignment functions.

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

Table 2

Comparison of attention mechanisms by training accuracy, precision, recall and f1-score metrics based on the NSL-KDD dataset (multiclass classification)

Attention	Accuracy	Precision	Recall	F1-score
Global (location)	99.79	99.79	99.79	99.79
Local (Monotonic + dot)	99.82	99.82	99.82	99.82
Local (Predictive + scaled_dot)	99.81	99.81	99.81	99.81
Self-attention	99.83	99.83	99.83	99.83
HybridAttention	**99.85**	**99.85**	**99.85**	**99.85**

Table 3

Comparison of attention mechanisms by testing accuracy, precision, recall and f1-score metrics based on the NSL-KDD dataset (multiclass classification)

Attention	Accuracy	Precision	Recall	F1-score
Global (location)	78.83	81.68	78.84	78.84
Local (Monotonic + dot)	80.50	84.21	79.83	79.83
Local (Predictive + scaled_dot)	81.48	82.80	81.48	81.48
Self-attention	78.86	81.90	78.86	78.86
HybridAttention	**81.91**	**85.12**	**81.91**	**81.91**

Table 4

Comparison of attention mechanisms by training accuracy, precision, recall and f1-score metrics based on the UNSW-NB15 dataset (multiclass classification)

Attention	Accuracy	Precision	Recall	F1-score
Global (scaled_dot)	97.91	97.66	97.91	97.78
Local (Monotonic + concat)	97.98	97.93	97.98	97.95
Local (Predictive + concat)	98.02	98.03	98.02	98.02
Self-attention	97.86	97.75	97.86	97.77
HybridAttention	**98.06**	**98.08**	**98.05**	**98.06**

Table 5

Comparison of attention mechanisms by testing accuracy, precision, recall and f1-score metrics based on the UNSW-NB15 dataset (multiclass classification)

Attention	Accuracy	Precision	Recall	F1-score
Global (scaled_dot)	97.87	97.58	97.87	97.72
Local (Monotonic + concat)	97.94	97.87	97.94	97.90
Local (Predictive + concat)	97.92	97.95	97.67	97.80
Self-attention	97.82	97.55	97.82	97.68
HybridAttention	**98.05**	**98.09**	**98.05**	**98.07**

Analyzing the data in tables 2-5 shows that the HybridAttention mechanism outperforms others on the NSL-KDD dataset, with better results on both training and test datasets, particularly with local attention and the scaled_dot function. Similarly, on the UNSW-NB15 dataset, the hybrid mechanism outperformed all alternatives on both datasets.

Table 6

Comparison of attention mechanisms by training accuracy, precision, recall and f1-score metrics based on the NSL-KDD dataset (binary classification)

Attention	Accuracy	Precision	Recall	F1-score
Global (scaled_dot)	99.79	99.79	99.79	99.79
Local (Monotonic + location)	99.79	99.79	99.79	99.79
Local (Predictive + concat)	99.81	99.81	99.81	99.81
Self-attention	99.79	99.79	99.79	99.79
HybridAttention	**99.85**	**99.85**	**99.85**	**99.85**

Table 7

Comparison of attention mechanisms by testing accuracy, precision, recall and f1-score metrics based on the NSL-KDD dataset (binary classification)

Attention	Accuracy	Precision	Recall	F1-score
Global (scaled_dot)	80.47	83.40	80.47	81.90
Local (Monotonic + concat)	84.12	85.84	**84.13**	84.97
Local (Predictive + concat)	**86.21**	82.38	82.34	82.36
Self-attention	82.01	84.37	82.02	83.17
HybridAttention	83.47	**86.96**	83.65	**85.27**

Table 8

Comparison of attention mechanisms by accuracy, precision, recall and f1-score metrics based on the UNSW-NB15 dataset (binary classification)

Attention	Accuracy	Precision	Recall	F1-score
Global (scaled_dot)	98.98	98.98	98.98	98.98
Local (Monotonic + general)	99.18	99.18	99.18	99.18
Local (Predictive + scaled_dot)	99.18	99.18	99.18	99.18
Self-attention	99.15	99.15	99.15	99.15
HybridAttention	**99.20**	**99.20**	**99.20**	**99.20**

Table 9

Comparison of attention mechanisms by testing accuracy, precision, recall and f1-score metrics based on the UNSW-NB15 dataset (binary classification)

Attention	Accuracy	Precision	Recall	F1-score
Global (scaled_dot)	98.98	98.98	98.98	98.98
Local (Monotonic + general)	99.17	99.18	99.17	99.17
Local (Predictive + scaled_dot)	99.17	99.17	99.17	99.17
Self-attention	99.14	99.14	99.14	99.14
HybridAttention	**99.19**	**99.19**	**99.19**	**99.19**

Tables 4-9 show that the HybridAttention mechanism was the best for all metrics and datasets for the UNSW-NB15 dataset. For the NSL-KDD dataset, the HybridAttention mechanism is the best option for the training dataset. However, for the test dataset, the local

attention mechanism with predictive alignment and concat alignment function showed the best accuracy of 86.21%, and the local attention mechanism with monotonic alignment with concat alignment function showed the best recall of 84.02%. Similarly, the global attention mechanism showed worse results for all the previous tables.

Tables 10 and 11 present the results of comparing the HybridAttention mechanism with modern developments. Note that the tables contain the value "N/A", which indicates the absence of data on these metrics in the work.

Table 10

Comparison of the model with a HybridAttention mechanism with modern developments (NSL-KDD)

Model name	Dataset	Accuracy	Precision	Recall	F1-score
CNN-BiLSTM-Attention [3]	NSL-KDD (multi-class)	99.79	N/A	99.83	N/A
CANET [6]	NSL-KDD (multi-class)	99.77	N/A	99.72	N/A
CANET [6]	NSL-KDD (binary)	99.79	N/A	99.72	N/A
CNN-BiGRU-Attention[12]	NSL-KDD (multi-class)	99.81	99.81	99.81	99.81
CNN-BiGRU-Attention [12]	NSL-KDD (binary)	99.83	99.83	99.83	99.83
The proposed model	NSL-KDD (multi-class)	**99.84**	**99.84**	**99.84**	**99.84**
The proposed model	NSL-KDD (binary)	**99.85**	**99.85**	**99.85**	**99.85**

Table 11

Comparison of the model with a HybridAttention mechanism with modern developments (UNSW-NB15)

Model name	Dataset	Accuracy	Precision	Recall	F1-score
SSAE-BiGRU-Att [5]	UNSW_NB15 (binary)	98.68	99	99	N/A
SSAE-BiGRU-Att [5]	UNSW_NB15 (binary)	98.68	99	99	N/A
CANET [6]	UNSW-NB15 (multi-class)	89.39	N/A	98.93	N/A
CANET [6]	UNSW-NB15 (binary)	96.43	N/A	96.97	N/A
CNN-BiGRU-Attention[12]	UNSW-NB15 (multi-class)	97.80	97.61	97.81	97.71
CNN-BiGRU-Attention[12]	UNSW-NB15 (binary)	99.18	99.17	99.17	99.17
CNN-BiLSTM-Attention [3]	UNSW-NB15 (multi-class)	88.83	N/A	98.51	N/A
Bi-LSTM [22]	UNSW_NB15 (binary)	99.05	98.96	99.36	99.15
The proposed model	UNSW-NB15 (multi-class)	**98.06**	**98.08**	**98.05**	**98.06**
The proposed model	UNSW-NB15 (binary)	**99.20**	**99.20**	**99.20**	**99.20**

The data presented in Tables 10 and 11 demonstrate that the proposed approach with a HybridAttention mechanism outperforms the results of known studies. In a previous study [12], only one attention mechanism was considered, whereas alternative options were evaluated in this study. The results of the training datasets were selected for comparison. The CNN-BiLSTM-Attention [3] and SSAE-BiGRU-Att [5] architectures have shown competitive results, but their effectiveness often depends on the specific dataset. Compared with these, the proposed HybridAttention mechanism demonstrates more stable accuracy and generalization ability on different NSL-KDD and UNSW-NB15 datasets. More complex models, such as CANET [6] and HAGRU [7], use hierarchical attention mechanisms to account for multi-level dependencies, but this increases computational costs and does not guarantee stable generalization. Hybrid approaches, such as EHID-SCA [8], which combine spatial and channel attention, have shown improvements in wireless sensor networks, but remain domain-specific. The Bi-LSTM model [22], which combines multidimensional feature processing with long-term dependency learning, has shown high performance on the UNSW-NB15, NSL-KDD, and CIC-IDS2017 datasets, but its effectiveness also depends on data sampling and preprocessing. In contrast, the proposed HybridAttention mechanism, which combines self-attention and dynamic local attention, demonstrates greater versatility, high accuracy, and generalization ability on the NSL-KDD and UNSW-NB15 datasets.

Although the proposed HybridAttention mechanism demonstrates clear advantages over individual attention variants, certain limitations should be acknowledged. First, the proposed approach increases computational requirements during training and inference, which may pose challenges for deployment in highly resource-constrained environments, such as IoT or edge devices. Nevertheless, this limitation is partly mitigated by the increasing availability of hardware accelerators and optimization frameworks. Second, while dynamic local attention enhances adaptability, the model's interpretability remains limited, which is a common issue for deep learning–based NIDS. Third, the system's performance is inherently dependent on the quality and diversity of the training data, as insufficient representation of novel attacks could reduce robustness; however, the hybrid design improves generalization compared to single-mechanism baselines. Finally, the model's scalability to high-throughput backbone networks requires further evaluation because additional latency may occur under heavy traffic. Despite these considerations, the proposed HybridAttention mechanism is a promising solution that balances detection accuracy and adaptability.

4. Practical Implementation Scenario for Critical or Corporate Networks

While the experimental evaluation presented in Section 3 was conducted in the Google Colab Pro cloud environment, this section focuses on how the trained model can be practically deployed within critical or corporate network infrastructures. The proposed architecture can be implemented in both critical networks, where real-time detection and high availability are essential (e.g., industrial control systems, healthcare, or energy infrastructures), and corporate environments,

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

which emphasize scalability, interoperability, and integration with existing SOC/SIEM systems. In critical networks, stringent latency and reliability requirements require lightweight deployment on dedicated nodes, whereas in corporate infrastructures, multi-layer security and centralized monitoring are prioritized.

As illustrated in Fig. 3, the proposed approach based on a HybridAttention mechanism can be effectively integrated into the security infrastructure of corporate or critical networks.

In this context, network traffic is duplicated at key points within the network, including core switches and aggregation routers, using standard mirroring techniques. The mirrored traffic is transmitted to a dedicated analysis node or an isolated containerized environment, where real-time inspection and classification are performed.

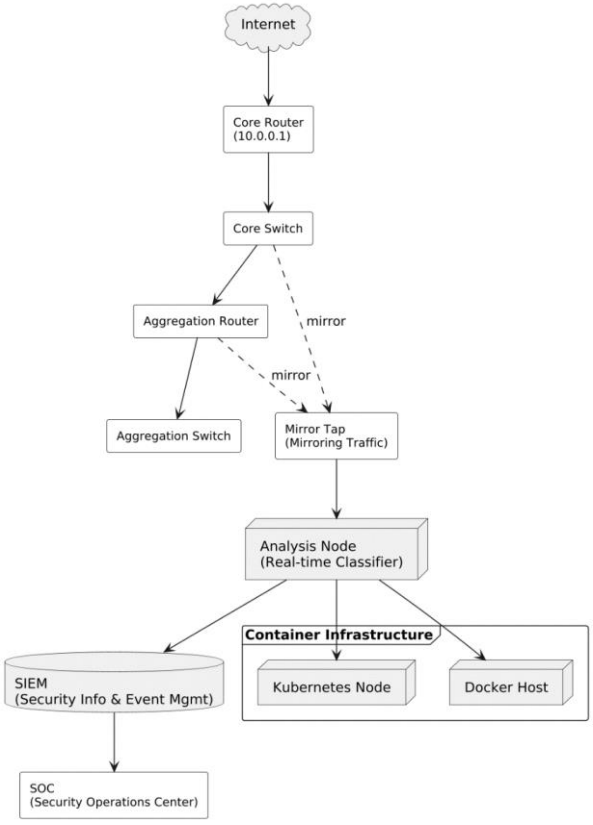

Fig. 3. HybridAttention-Based NIDS Architecture

The system can operate in near real-time and classify observed network flows into benign or potentially malicious categories. The attention mechanism implemented in the classification module allows the system to dynamically focus on significant features and structural patterns of traffic, thereby enabling the detection of both known and previously unseen threats. This adaptability enhances the resilience of the system in changing operational environments.

For practical deployment, the solution may be integrated as a detection module within a Security Operations Center (SOC). Detected anomalies are converted into structured alerts and transmitted centralized monitoring systems, such as Security Information and Event Management (SIEM) platforms. These alerts can initiate automated incident response procedures based on predefined policies, including dynamic modification of firewall rules, traffic redirection, or initiation of in-depth forensic investigation.

The scalability and reliability of the system can be ensured using container orchestration platforms (e.g., Docker and Kubernetes), which allow for horizontal scaling and load balancing depending on the traffic volume and infrastructure constraints. The modular design of the solution supports flexible integration into existing cybersecurity architectures, regardless of the size or technical stack of the organization.

Thus, the proposed method is not limited to theoretical evaluation but demonstrates its applicability in operational conditions. It ensures timely detection of complex threats while maintaining adaptability, transparency, and compatibility with modern infrastructure management practices.

5. Conclusions

1. Based on the analyzed studies, a review of the existing attention mechanisms is carried out. The use of attention mechanisms in NIDS can significantly improve their effectiveness in recognizing and tracking anomalies or attacks in network traffic. Local attention can focus on the key characteristics of network traffic, thereby reducing computational complexity. Global attention can consider the overall context and detect complex anomalies. Self-attention can dynamically determine the importance of different parts of the input data, adapting to different types of traffic and attacks.

2. Different attention mechanisms, such as local, global, and self-attention, are added to the previously developed CNN-BiGRU-Attention model instead of the Attention layer. Their effectiveness in intrusion detection is compared using traditional evaluation metrics. The experimental results show that local and self-attention work better than global attention in the network intrusion detection context.

3. Comparison of the use of a HybridAttention mechanism that includes dynamic local attention and self-attention with the results of known studies. In combination with the HybridAttention mechanism, the CNN-BiGRU model demonstrated better results for both multiclass and binary classification.

These results confirm that the HybridAttention mechanism effectively enhances both spatial and temporal feature extraction within network flows. The mod-

el efficiently prioritizes relevant traffic patterns by combining self-attention with dynamic local attention, thereby improving detection accuracy for both known and emerging threats. In addition, the adaptive window mechanism contributes to better generalization across datasets and reduces false-positive rates, further strengthening the model's robustness.

Despite these promising findings, several limitations remain. The interpretability of the model's decisions is limited, which may hinder human understanding and results' reproducibility in practical applications. Furthermore, the data preprocessing procedure and the evaluation metrics range require further refinement. Future research will focus on addressing these shortcomings and optimizing the model through dimensionality reduction and metaheuristic algorithms, such as the genetic algorithm, to improve both efficiency and transparency.

Contributions of authors: conceptualization, methodology – **Andrii Nikitenko;** formulation of tasks, analysis **– Yevhen Bashkov;** development of model, software, verification **– Andrii Nikitenko;** analysis of results, visualization **– Andrii Nikitenko;** writing – original draft preparation, writing – review and editing **– Yevhen Bashkov.**

Conflict of Interest

The authors declare that they have no conflict of interest in relation to this research, whether financial, personal, authorship or otherwise, that could affect the research and its results presented in this paper.

Financing

This study was conducted without any financial support.

Data Availability

Data will be made available upon reasonable request.

Use of Artificial Intelligence

The authors confirm that they did not use artificial intelligence methods while creating the presented work.

Acknowledgments

All the authors have read and agreed to the published version of this manuscript.

References

1. Brauwers, G., & Frasincar, F. A general survey on attention mechanisms in deep learning. *IEEE Transactions on Knowledge and Data Engineering*, 2022, vol. 35, no. 4, pp. 3279-3298. DOI: 10.1109/TKDE.2021.3126456.

2. Niu, Z., Zhong, G., & Yu, H. A review on the attention mechanism of deep learning. *Neurocomputing*, 2021, vol. 452, pp. 48–62. DOI: 10.1016/j.neucom.2021.03.091.

3. Dai, W., Li, X., Ji, W., & He, S. Network Intrusion Detection Method Based on CNN-BiLSTM-Attention Model. *IEEE Access*, vol. 12, pp. 53099-53111, 2024. DOI: 10.1109/ACCESS.2024.3384528.

4. Yang, Y., Tu, S., Hashim Ali, R., Alasmary, H., Waqas, M., & Nouman Amjad, M. Intrusion detection based on bidirectional long short-term memory with attention mechanism. *Computers, Materials & Continua*, 2023, vol. 74, iss. 1, pp.801–815. DOI: 10.32604/cmc.2023.031907.

5. Wang, J., Chen, N., Yu, J., Jin, Y., & Li, Y. An efficient intrusion detection model combined bidirectional gated recurrent units with attention mechanism. *7th International Conference on Behavioural and Social Computing (BESC)*, Bournemouth, United Kingdom, 2020, pp. 1-6. DOI: 10.1109/BESC51023.2020.9348310.

6. Yuan, S., Ren, K., Zhang, C., Shi, Y., & Huang, Z. CANET: A hierarchical CNN-Attention model for network intrusion detection. *Computer Communications*, 2023, vol. 205, pp 170-181. DOI: 10.1016/j.comcom.2023.04.018.

7. Liu, X., & Liu, J. Malicious traffic detection combined deep neural network with hierarchical attention mechanism. *Scientific Reports*, 2021, vol. 11, iss. 1. DOI: 10.1038/s41598-021-91805-z.

8. Chavan, P., Hanumanthappa, H., Satish, E. G., Sunil, M., Supreeth, S., Rohith, S., & Ramaprasad, H. C. Enhanced Hybrid Intrusion Detection System with Attention Mechanism using Deep Learning. *SN COMPUT. SCI.*, 2024, vol. 5. DOI: 10.1007/s42979-024-02852-y.

9. Yang, K., Wang, J., & Li, M. An improved intrusion detection method for IIoT using attention mechanisms, BiGRU, and Inception-CNN. *Sci Rep*, 2024, vol. 14, no. 19339. DOI: 10.1038/s41598-024-70094-2.

10. Alshehri, M. S., Saidani, O., Alrayes, F. S., Abbasi, S. F., & Ahmad, J. A Self-Attention-Based Deep Convolutional Neural Networks for IIoT Networks Intrusion Detection. *IEEE Access*, 2024, vol. 12, pp. 45762-45772. DOI: 10.1109/ACCESS.2024.3380816.

11. Gupta, N., Malladi, R., Naganjaneyulu, S., & Balhara, S. Optimized Attention Induced Multi Head Convolutional Neural Network for Intrusion Detection Systems in Vehicular Ad Hoc. *IEEE Transactions on Intelligent Transportation Systems*, 2025, pp. 1-10. DOI: 10.1109/TITS.2025.3561545.

12. Nikitenko, A., & Bashkov, Y. Construction of a network intrusion detection system based on a convolutional neural network and a bidirectional gated recurrent unit with attention mechanism. *Eastern-European Journal of Enterprise Technologies*, 2024, vol. 3, pp.6–15. DOI: 10.15587/1729-4061.2024.305685.

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

13. Correia, A. D. S., & Colombini, E. L. Attention, please! A survey of neural attention models in deep learning. *arXiv preprint*, 2021. DOI: 10.48550/arXiv.2103.16775.

14. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. Attention is all you need. *arXiv preprint*, 2023. DOI: 10.48550/ariv.1706.03762.

15. Yang, S., Wu, P., & Guo, H. DualNet: Locate then detect effective payload with deep attention network. *arXiv preprint*, 2020. DOI: 10.48550/arXiv.2010.12171.

16. Bahdanau, D., Cho, K., & Bengio, Y. Neural machine translation by jointly learning to align and translate. *arXiv preprint*, 2016. DOI: 10.48550/arXiv.1409.0473.

17. Luong, M. T., Pham, H., & Manning, C. D. Effective approaches to attention-based neural machine translation. *arXiv preprint*, 2015. DOI: 10.48550/arXiv.1508.04025.

18. Britz, D., Goldie, A., Luong, M. T., & Le, Q. Massive exploration of neural machine translation architectures. *arXiv preprint*, 2017. DOI: 10.48550/arXiv.1703.03906.

19. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R., & Bengio, Y. Show, attend and tell: Neural image caption generation with visual attention. *arXiv preprint*, 2016. DOI: 10.48550/arXiv.1502.03044.

20. *University of New Brunswick*, 2023. *NSL-KDD dataset*. Available at: https://www.unb.ca/cic/datasets/nsl.html. (accessed 12.10.2024).

21. *UNSW, 2023. The UNSW-NB15 Dataset*. Available at: https://research.unsw.edu.au/projects/unsw-nb15-dataset (accessed 12.10.2024).

22. Cao, B., Li, C., Song, Y., & Fan, X. Network intrusion detection technology based on convolutional neural network and BiGRU. *Computational Intelligence and Neuroscience*, 2022, vol. 2022. DOI: 10.1155/2022/1942847.

Received 17.03.2025, Accepted 25.08.2025

ВИКОРИСТАННЯ ГІБРИДНОГО МЕХАНІЗМУ УВАГИ ЯК МЕТОДУ ПІДВИЩЕННЯ ЕФЕКТИВНОСТІ СИСТЕМ ВИЯВЛЕННЯ МЕРЕЖЕВИХ ВТОРГНЕНЬ

А. О. Нікітенко, Є. О. Башков

Предметом дослідження є гібридний механізм уваги, інтегрований у глибоку нейронну архітектуру для мережевих систем виявлення вторгнень (NIDS). **Метою** роботи є розробка та дослідження гібридного механізму уваги на основі комбінації глобальної (самоуваги) та локальної (динамічної локальної уваги) моделей для підвищення якості класифікації трафіку в NIDS, що працюють в режимі реального часу. **Завдання**: аналіз існуючих механізмів уваги на предмет їх застосовності для виявлення мережевих вторгнень; інтеграція різних типів уваги в архітектуру CNN-BiGRU; розробка гібридного механізму уваги на основі динамічного вирівнювання вікон; оптимізація моделі за допомогою Optuna; експериментальна оцінка її продуктивності на тестових наборах даних з використанням стандартних метрик класифікації. Використані **методи**: моделювання глибокого навчання з архітектурою CNN-BiGRU, інтеграція різних механізмів уваги, включаючи нову гібридну увагу, оптимізація гіперпараметрів за допомогою Optuna та оцінка продуктивності на основі стандартних метрик класифікації. **Результати** роботи показують, що запропонований механізм гібридної уваги демонструє перевагу над окремими типами уваги за всіма ключовими метриками. Модель досягла точності до 99,85% на навчальних даних набору даних NSL-KDD і продемонструвала сильне узагальнення на наборі даних UNSW-NB15, досягнувши точності до 98,06% в багатокласовій класифікації і до 99,20% в бінарній класифікації. Модель також перевершила сучасні аналогічні підходи для обробки незбалансованих даних та виявлення різних типів атак. **Висновки**. Наукова новизна отриманих результатів полягає в наступному: розроблено гібридний механізм уваги, що поєднує самоувагу та динамічну локальну увагу для покращення послідовного розпізнавання образів у мережевому трафіку; покращено архітектуру CNN-BiGRU за рахунок інтеграції декількох модулів уваги; систематична гіперпараметрична оптимізація з використанням Optuna покращила узагальнення на незбалансованих даних; запропонована модель перевершила існуючі підходи на тестових наборах даних при виявленні як відомих, так і нових кібератак.

Ключові слова*:* глибоке навчання; механізм уваги; система виявлення мережевих вторгнень; гібридний механізм уваги; динамічна локальна увага.

Нікітенко Андрій Олександрович – асп. каф. прикладної математики та інформатики ДВНЗ «Донецький національний технічний університет», Дрогобич, Україна

Башков Євген Олександрович – д-р техн. наук, проф., проф. каф. прикладної математики та інформатики ДВНЗ «Донецький національний технічний університет», Дрогобич, Україна

Andrii Nikitenko — PhD Student of the Department of Applied Mathematics and Informatics, Donetsk National Technical University, Drohobych, Ukraine,
e-mail: andrii.nikitenko@donntu.edu.ua, ORCID: 0009-0006-1363-2324.

Yevhen Bashkov — Doctor of Technical Sciences, Professor, Professor at the Department of Applied Mathematics and Informatics, Donetsk National Technical University, Drohobych, Ukraine,
e-mail: yevhen.bashkov@donntu.edu.ua, ORCID: 0000-0001-6974-4882, Scopus Author ID: 6602250825.