An explainable artificial intelligence approach for detecting network attacks

Dmytro Tymoshchuk[1,*,†], Andriy Sverstiuk[2,†], Yurii Klots[3,†], Nataliia Petliak[3,†] and Vira Titova[3,†]

[1]Ternopil Ivan Puluj National Technical University, Ruska str. 56, Ternopil, 46001, Ukraine
[2]I. Horbachevsky Ternopil National Medical University, Maidan Voli St., 1, Ternopil, 46002, Ukraine
[3]Khmelnytskyi National University, 11, Instytuts'ka str., Khmelnytskyi, 29016, Ukraine

Abstract

The proliferation of sophisticated network attacks necessitates advanced cyber defense systems that are not only accurate but also transparent. However, many machine learning-based intrusion detection systems operate as 'black boxes,' hindering trust and practical implementation. This paper addresses this gap by presenting an explainable artificial intelligence (XAI) approach for detecting network attacks. We developed and compared three models—Random Forest, XGBoost, and an MLP neural network—finding that the MLP model achieved the highest accuracy of 0.9508, a log loss of 0.0793, and an AUC of 0.9922. To ensure transparency, we integrated XAI methods like LIME for local and Permutation Feature Importance for global explanations, successfully identifying key traffic characteristics that influence model decisions. The proposed approach was validated in a realistic KVM-based laboratory environment, confirming that the synergy of high-accuracy machine learning with XAI provides a robust and trustworthy framework for modern cybersecurity.

Keywords

IDS, IPS, machine learning, XAI, LIME, permutation feature importance, cybersecurity, hypervisor, operating systems

1. Introduction

The rapid growth in the number of devices connected to the network and the intensity of data transmission leads to an increase in the level of threats in cyberspace. Modern attacks on network infrastructure are characterized by high complexity, multi-vector nature, and the ability to bypass traditional security measures. In these conditions, it is important to improve intrusion detection systems (IDS) and intrusion prevention systems (IPS) that are capable of responding to traffic anomalies in a timely manner.

Traditional cyber security methods based on a signature approach no longer provide sufficient protection against modern attacks. This is due to the emergence of new types of threats that rapidly change their characteristics. In addition, the growth in network traffic data volumes creates additional challenges for effective analysis and requires the implementation of more flexible and scalable solutions.

Recent years have seen significant progress in the application of machine learning methods in many areas. In finance, machine learning is used to improve financial analytics [1], in transportation to develop intelligent systems based on IoT [2], in medicine to predict disease progression and support clinical decision-making [3], in materials science both for classifying composites [4] and for modeling the fatigue lifetime of metals [5], in energy to estimate energy consumption [6], and in cybersecurity to analyze network traffic [7], which confirms the versatility and effectiveness of these methods in prediction tasks.

ExplAI-2025: Advanced AI in Explainability and Ethics for the Sustainable Development Goals, November 07, 2025, Khmelnytskyi, Ukraine

*Corresponding author.
†These authors contributed equally.
dmytro.tymoshchuk@gmail.com (D. Tymoshchuk); sverstyuk@tdmu.edu.ua (A. Sverstiuk); klots@khmnu.edu.ua (Y. Klots); npetlyak@khmnu.edu.ua (N. Petliak); titovav@khmnu.edu.ua (V. Titova)
0000-0003-0246-2236 (D. Tymoshchuk); 0000-0001-8644-0776 (A. Sverstiuk); 0000-0002-3914-0989 (Y. Klots); 0000-0001-5971-4428 (N. Petliak); 0000-0001-8668-4834 (V. Titova)

The use of classification algorithms and intelligent models improves the accuracy and speed of identifying abnormal behavior on the network. At the same time, the widespread implementation of such approaches is accompanied by the "black box" problem, where the results of the model are difficult to interpret and explain. This limits trust in cybersecurity systems, especially in critical infrastructure, where decisions must be understandable to experts and comply with regulatory requirements. In this context, a promising direction for development is the use of Explainable Artificial Intelligence (XAI) methods [8], which allow explaining which network traffic features had the greatest impact on decision-making. This ensures model transparency, increases their reliability, and facilitates faster integration into practice. The development of attack detection systems using XAI also has an indirect but significant connection with achieving the Sustainable Development Goals, as enhancing cybersecurity is a fundamental prerequisite for the stable functioning of modern society and its institutions.

The problem remains finding methods and models that can combine high attack detection accuracy with minimization of false positives, effective real-time operation, and a high level of interpretability. The developed systems must ensure scalability, adaptability to new types of threats, and the ability to explain their decisions in an understandable way. The main contribution of this work lies in combining machine learning models with Explainable AI methods, which simultaneously improves the accuracy of network attack detection and ensures decision-making transparency in IDS/IPS systems.

Next, the paper is organized as follows. Section 2 provides a review of related work and modern approaches to applying machine learning and Explainable AI in cybersecurity. Section 3 describes the dataset used, the selected machine learning methods, and the Explainable AI techniques. Section 4 presents the modeling results, performance analysis of the models, and interpretation of their decisions. Section 5 summarizes the findings.

2. Related Work

The issue of detecting and preventing intrusions into computer networks has attracted the attention of scientists over the past decades. Machine learning methods have proven to be effective in classifying network traffic and detecting various types of attacks with high accuracy [9, 10]. One of the current trends is the integration of XAI methods into attack detection systems. Technologies such as SHAP [11], LIME [12], and Integrated Gradients [13] provide transparency in decision-making and allow researchers and security professionals to better understand the impact of individual traffic features on classification. This increases trust in cyber defense systems.

In [14], the authors address the problem of the growing number of DDoS attacks, which significantly threaten the availability of online services. A new method for detecting attacks based on Explainable AI is proposed, which analyzes network traffic at the network level (L3) and identifies the most influential features for each anomalous behavior. Based on weighting coefficients, a system of threshold values is formed, which allows for the adaptive creation of security policies against different classes of attacks.

The paper [15] discusses the integration of the Internet of Things (IoT) and intelligent transportation systems (ITS), which forms the concept of the Internet of Vehicles (IoV). The combination of IoV with fifth-generation (5G) communication technologies enables the development of intelligent connected vehicles (ICVs). The authors identify five critical security domains: protection of vehicles, intelligent devices, service platforms, V2X communications, and data. Artificial intelligence models are actively used to counter intrusions. The authors show that the role of XAI increases transparency and trust in IDS. A review of XAI solutions has shown their significant potential in the field of ICV cyber security and improving the efficiency of transport systems.

The authors in [16] consider the problem of increasing security risks in the context of IoT development, where a large number of connected devices increases the likelihood of intrusions. A deep learning methodology is proposed for detecting and classifying DDoS attacks in the IoT environment. The approach is based on the use of learning mechanisms with different model architectures, which allowed both binary and multi-class experiments with varying classification complexity to be conducted. To increase transparency, Explainable AI methods are used to explain the contribution of features to

the attack detection process. The results of the experiments demonstrated high efficiency for the XAI-BiLSTM model, confirming the promise of the approach.

Study [17] provides an overview of current research on the application of XAI in cybersecurity. The authors note that artificial intelligence methods are actively used to detect intrusions, malware, and spam, demonstrating higher efficiency compared to classical signature-based or rule-based approaches. However, most of these models function as "black boxes," which reduces the trust of experts and users. This highlights the need to implement XAI to ensure transparency and interpretability without losing accuracy. The authors highlight the lack of systematic reviews dedicated specifically to the application of XAI in cybersecurity and propose their own approach, which outlines a roadmap for further research in this area.

The paper [18] investigates the problem of attacks on the Domain Name System (DNS), which remains one of the key vectors of intrusion. Despite the advantages of the DNS over HTTPS (DoH) protocol, particularly in terms of privacy and security, it makes it difficult for network administrators to detect malicious traffic. To address this issue, the authors proposed an approach based on Explainable AI, implemented in a new ML architecture. Using the CIRA-CIC-DoHBrw-2020 dataset, the authors developed a balanced model with a Random Forest ensemble, which achieved high Precision, Recall, and F1-score metrics. Additionally, the contribution of features to the classification process was demonstrated, which increases transparency and trust in the results.

The authors in [19] investigate the vulnerability of Software-Defined Networking (SDN) architectures to DDoS attacks, which can significantly reduce service performance and availability. The authors proposed a deep learning-based DDoS attack detection system specifically designed for SDN environments. The CIC-DDoS2019 dataset was used for training and testing. Various architectures were investigated, including CNN, LSTM, RNN, GRU, and ANN, with ANN demonstrating the highest results with accuracy, precision, recall, and F1-score at 0.9999. In addition, Explainable AI methods, including SHAP and LIME, were applied to improve transparency and interpretability, which made it possible to explain the impact of features on decision-making and increase trust in the system.

The paper [20] addresses the problem of DoS attacks in 5G networks, which are vulnerable due to the complexity and decentralization of their architecture. To detect and prevent such attacks, the authors proposed an approach that combines deep learning methods with Explainable AI tools. In particular, the LIME method is used to interpret model decisions and identify the most significant features in the classification process. Experiments have shown that the Random Forest model achieved the highest recall rate of 99.98%.

The study [21] examines the problem of DDoS attacks, which disrupt the operation of critical online services and are becoming increasingly sophisticated. The authors note the key role of machine learning in detecting and classifying attacks, especially using modern datasets that cover new types of threats, including UDP-flood, SIDDoS, HTTP-flood, and Smurf. IDS integrated with ML allow analyzing network traffic and separating normal activity from malicious activity. Explainable AI methods (LIME, SHAP) improve the interpretability of models by providing an understanding of the impact of features on results.

The paper [22] examines the vulnerability of IoT systems to cyberattacks caused by limited or typical security measures. To improve security, machine learning methods are proposed, including intelligent IDS, threat detection systems, and behavioral analysis. At the same time, ML-oriented IDS face problems such as high false positives, evolutionary attacks, data quality, and insufficient transparency. The authors propose a new XAI-oriented IDS architecture for detecting IoT vulnerabilities, in particular the Ripple20 family. The framework integrates ML classifiers with XAI to form interpretable decisions. Experiments on proprietary and open datasets (binary and multi-class classification) demonstrated high accuracy and efficiency. The advantage is increased expert confidence and the ability to adapt to a dynamic IoT environment.

Study [23] examines the problem of increasing attacks on IoT devices, particularly DDoS and malware, which reduce network performance and pose threats to data privacy and integrity. Traditional centralized detection mechanisms demonstrate limited scalability and privacy issues, which highlights the need for decentralized solutions. The authors proposed a framework based on Federated Learning (FL), which

implements two-level feature selection using Recursive Feature Elimination and correlation filtering, allowing the most relevant and independent parameters to be identified. To increase transparency, Explainable AI methods are integrated to explain the contribution of features to the classification process. Gradient boosting is used as the model, which provided 99.73% accuracy. Thanks to the distributed architecture of FL, the system is suitable for real-world application in IoT networks, increasing security through scalability, privacy preservation, and efficiency.

The paper [24] examines the impact of AI development on the changing nature of cyberattacks. Modern attackers using AI are able to automate actions, analyze large amounts of data, and detect vulnerabilities with high accuracy. At the same time, despite its numerous advantages, the IoT creates serious security problems that require effective IDS. Traditional ML- and DL-based IDS demonstrate high performance, but problems remain with false positives and a lack of transparency in decisions. The authors proposed an AI-oriented IDS with built-in Explainable AI mechanisms that uses the SHAP method to explain the classification process. Experimental results confirmed the high accuracy and effectiveness of the proposed approach, as well as its ability to increase administrators' trust in IDS systems in the IoT environment.

The study [25] examines new approaches to IDS development, driven by the rapid growth in the number of attacks on network systems. The authors emphasize the need for explainability of artificial intelligence models used in such systems in order to increase the trust and effectiveness of security analysts. A framework for evaluating black-box Explainable AI methods in the context of network IDS is proposed. Both global and local aspects of XAI application, in particular SHAP and LIME methods, are investigated using six metrics. Experiments were conducted on three popular datasets and seven AI models. The results revealed the strengths and weaknesses of XAI methods and provided a basic toolkit for the cybersecurity community.

The paper [26] provides an overview of modern XAI methods in the field of network traffic analysis (NTA). It considers key tasks, including traffic classification, intrusion detection, attack classification, and traffic characteristic prediction. The review covers techniques, practical applications, requirements, challenges, and current projects that illustrate the role of XAI in ensuring security, optimizing performance, and improving network reliability. Particular attention is paid to the transparency of AI model decisions, as in critical environments it is important not only to obtain accurate results, but also to understand the reasons behind them. The proposed study forms a holistic vision of XAI for NTA and can serve as a guide for researchers and practitioners in creating explainable and trusted cybersecurity systems.

The goal of our research is to develop and evaluate machine learning models for network traffic analysis using Explainable AI, aimed at improving threat detection accuracy and ensuring interpretability of results in cybersecurity systems.

3. Materials and methods

3.1. Dataset description

The dataset used in this study was based on the UNSW-NB15 dataset created by the Australian Centre for Cyber Security (ACCS) [27]. It combines characteristics of real normal traffic with malicious scenarios that reflect current attack vectors in networks. The resulting dataset contained 38 input features describing network attributes, time parameters, packet and byte statistics, protocol and service characteristics.

Each entry corresponded to a separate network interaction session and was used to classify traffic as normal or malicious. The dataset covers a wide range of attacks, including DoS, Exploits, Fuzzers, Backdoors, Generic, Analysis, Shellcode, Reconnaissance, and Worms, making it suitable for comprehensive evaluation of intrusion detection systems. Figure 1 shows the distribution of classes in the dataset using a logarithmic scale on the Y-axis.

The output parameter in the research is a binary variable that indicates whether network traffic belongs to the normal or abnormal class. A value of 0 corresponds to legitimate traffic (Normal), while

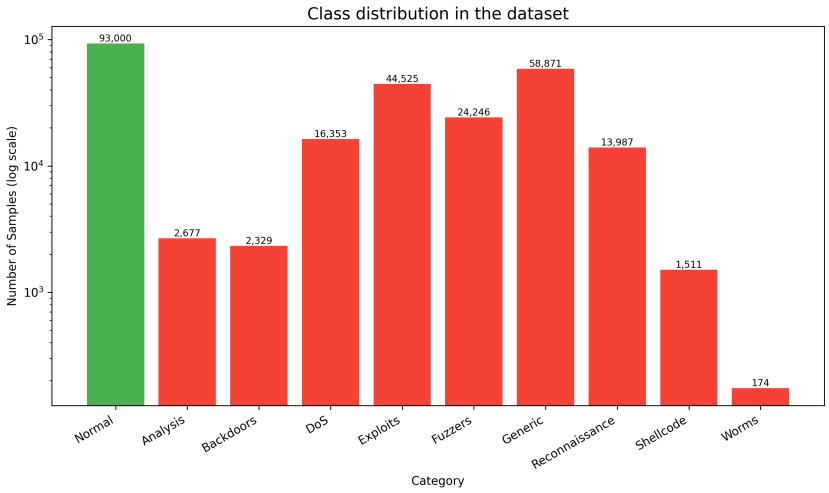


Figure 1: Distribution of normal and attack classes in the UNSW-NB15 dataset. The y-axis is on a logarithmic scale to accommodate the wide range of sample counts across different categories, from 'Normal' traffic to various attack types like 'DoS', 'Exploits', and 'Worms'.

a value of 1 combines all types of attacks (Attack). This approach allowed us to formulate the task as a binary classification problem and simplified the training of machine learning models by generalizing the multi-class structure of the dataset.

To build and evaluate the effectiveness of machine learning models, the formed dataset was divided into training and test sets in a 70/30 ratio, ensuring that the structure of the target variable was preserved. The division was carried out according to the principle of stratified splitting, which means proportional representation of each class of attacks in the training and test subsets. Thus, for each type of attack, the data was distributed in a ratio of 70% for training and 30% for testing, which prevents class distribution bias and ensures the representativeness of both subsets.

3.2. Machine learning methods and explainable AI

In this study, modern machine learning methods were used to analyze network traffic and classify attacks, including Random Forest (RF), Extreme Gradient Boosting (XGBoost), and a multilayer neural network of the MLP type.

Random Forest (RF) is an ensemble machine learning method based on building a large number of decision trees using random subsets of data and features. The key principle is the Bagging (bootstrap aggregating) technique, which involves randomly selecting samples for training each tree and then aggregating their results. This approach reduces the risk of overfitting and increases the stability of the model. In classification tasks, the final prediction is determined by voting trees. An important advantage of RF is its ability to work with high-dimensional datasets that include both numerical and categorical features. Extreme Gradient Boosting (XGBoost) is an optimized implementation of gradient boosting that combines high classification accuracy with computational efficiency. Unlike Random Forest, where trees are built independently, XGBoost forms an ensemble sequentially, and each subsequent tree is aimed at correcting the errors of the previous ones. The use of regularization, parallel processing, and memory optimization makes this method extremely productive for large and complex datasets. XGBoost demonstrates excellent results in anomaly and attack detection tasks. A Multilayer Perceptron (MLP) neural network is a classic artificial neural network with multiple layers of perceptrons, where each layer consists of neurons with nonlinear activation functions. Thanks to this architecture, MLP is capable of approximating any complex functions and modeling nonlinear relationships between network traffic features and attack categories. In IDS/IPS tasks, the use of MLP allows for high accuracy, but requires significant computational resources and hyperparameter optimization to avoid overfitting. Despite these challenges, MLP remains one of the basic and at the same time effective approaches for

processing complex multidimensional data in the field of cybersecurity.

Explainable artificial intelligence is a field of research aimed at improving the transparency and interpretability of machine learning algorithms. In the field of cybersecurity, where it is important not only to obtain an accurate prediction but also to understand the reasons behind it, XAI provides experts with additional knowledge about which characteristics of network traffic influenced the classification as normal or malicious activity. Local Interpretable Model-Agnostic Explanations (LIME) is an approach focused on explaining individual model decisions [12]. Its essence is that for a specific data sample, LIME builds a simplified local model (for example, linear regression) that approximates the behavior of a complex model in the vicinity of this sample. This makes it possible to determine which features contributed most to the decision. This approach is particularly valuable for IDS systems, where it is critical to explain the reasons for triggering suspicious activity. Permutation Feature Importance is a method of global feature importance assessment that determines their impact by randomly permuting the values of a specific feature in a test set and measuring the decrease in model accuracy [28]. If the classification accuracy drops significantly after the permutation, this feature is highly important for the model. This approach is independent of the specific algorithm and allows identifying key characteristics of network traffic that determine the overall performance of the model. Combining the local explanations provided by LIME with the global assessment of feature importance using Permutation Importance creates a comprehensive interpretation system that balances prediction accuracy, decision transparency, and trust in machine learning models in the context of detecting network attacks.

3.3. Model performance indicators

To objectively evaluate the quality of the constructed models, we used a set of classic metrics [29] that are widely used in IDS/IPS research. Accuracy reflects the proportion of correctly classified examples among all observations and characterizes the overall accuracy of the model. However, in cases of class imbalance, it may not be sufficiently informative. Recall measures the model's ability to correctly identify positive examples, i.e., real attacks. A high value of this metric is especially important for cybersecurity tasks, since missed attacks pose the greatest threat. Specificity determines the proportion of correctly classified negative examples, i.e., the model's ability to correctly identify normal traffic. It is critical for reducing the number of false positives. Precision shows what proportion of all predicted attacks are actually attacks. This metric is important for assessing system reliability, as a high Precision score means a low false alarm rate. F1-Score is the harmonic mean between Precision and Recall, combining them into a single balanced metric. It is especially useful in cases where it is important to consider both missed attacks and false positives. G-Mean (geometric mean) is used for a comprehensive assessment of the balance of classification between classes. It takes into account both Recall and Specificity, providing a more reliable measurement in cases of significant data imbalance. The metrics Accuracy, Recall, Specificity, Precision, F1-score, and G-Mean were determined using standard formulas (Table 1) based on four key classification indicators: TP (true positives), TN (true negatives), FP (false positives), and FN (false negatives).

The log loss function is one of the basic metrics for evaluating probabilistic models in binary classification tasks. It measures how well the model's confidence in its predictions matches the actual classes [30]. If the model predicts the correct class with high probability, the contribution to the log loss value will be minimal. If it is wrong and at the same time shows excessive confidence (for example, predicts an "attack" with a probability of 0.99, but in fact it is "normal traffic"), the penalty in the loss function will be very large. For a binary classification task, log loss is defined as follows:

$$\text{LogLoss} = -\frac{1}{N}\sum_{i=1}^{N}(y_i \cdot \ln(p_i) + (1 - y_i) \cdot \ln(1 - p_i)), \tag{1}$$

where N denotes the total number of samples in the dataset, $y_i \in \{0, 1\}$ is the true class label for the i-th example, and p_i is the probability assigned by the model to the positive class.

If the model is absolutely confident in its prediction and it coincides with reality, the sum of the

Table 1
Formulas for calculating the key performance metrics used to evaluate the classification models.

Metric	Formula	Eq.
Accuracy	$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$	(1)
Recall	$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$	(2)
Specificity	$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$	(3)
Precision	$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$	(4)
F1-Score	$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$	(5)
G-Mean	$\text{G-Mean} = \sqrt{\text{Recall} \times \text{Specificity}}$	(6)

addends approaches zero. In the case of an incorrect prediction with high confidence, the value of the logarithmic term becomes very large in magnitude, which increases the overall log loss. Thus, log loss is sensitive not only to the correctness of the classification, but also to the distribution of probabilities that the model assigns to classes. This makes it a useful metric for evaluating models focused on probabilistic predictions, especially in the field of anomaly detection and network attacks, where it is important not only to classify correctly, but also to have a measure of confidence in the decision.

In addition to these metrics, additional evaluation tools were used. Visualization using a confusion matrix shows the relationship between true positives, negatives, and false classifications. Another important indicator is the Area Under the ROC Curve (AUC), which characterizes the algorithm's ability to separate the positive class from the negative class at different classification thresholds and determines the overall discriminatory power of the model. Equally important are Precision-Recall curves, which reflect the interdependence between Precision and Recall for different threshold values and are particularly informative in cases of significant class imbalance, which is typical for attack detection tasks. Using these methods in combination with classical metrics allows us to obtain a comprehensive view of model performance, identify their strengths and weaknesses, and make a reasonable comparison of results.

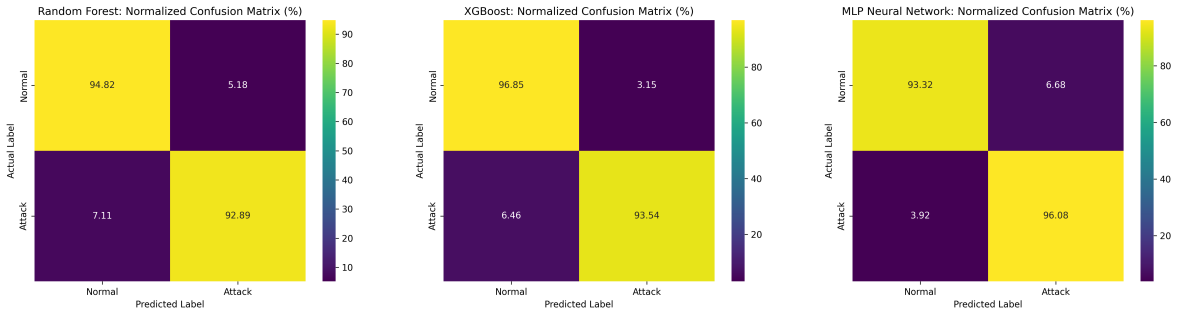
4. Results and discussion

4.1. Performance of machine learning models

For each machine learning method, confusion matrices were constructed and basic classification metrics were calculated. Since neural networks showed the highest accuracy among all the algorithms studied, an extended analysis was performed for the MLP model. Precision-Recall, ROC, and F1-score vs Threshold curves were constructed. In addition, Permutation Importance and LIME analysis were used to explain the model's decisions.

For the Random Forest model, an ensemble of 620 decision trees was used, which were constructed according to the Gini criterion. The model was not limited by the maximum depth of the trees, which allowed complex dependencies to be reproduced, while the number of features at each split was determined by the sqrt rule. To compensate for class imbalance, the class_weight=balanced_subsample strategy was used, with a minimum of one example in the leaf and two for the split. The algorithm ran in multiprocessing mode (n_jobs=-1) with reproducible results thanks to a fixed random state (random_state=42). For this model, the optimal hyperparameters were selected using the GridSearchCV exhaustive search method, which allowed us to find the most balanced parameter values. In the case of XGBoost, the settings were focused on the task of binary classification with a logistic loss function (objective=binary:logistic, eval_metric=logloss). The number of trees was 680 with a maximum depth of six, and the learning rate was 0.05. To reduce the correlation between trees, partial subsampling

parameters were used (colsample_bytree=0.9 and subsample=0.9). Positive class balancing was ensured by the scale_pos_weight=0.5647 coefficient, and regularization was performed with the reg_alpha=0.0 and reg_lambda=1.0 parameters. To speed up training, the tree_method=hist tree construction method was used, and the random state was also fixed (random_state=42). The optimal values of the hyperparameters of this model were also selected by searching GridSearchCV, which ensured maximum classification quality. The multilayer neural network of the MLP type consisted of two hidden layers with 128 and 64 neurons with the ReLU activation function. Network weight optimization was performed using the Adam algorithm with parameters beta_1=0.9, beta_2=0.999, and momentum=0.9. The initial learning rate was set to 0.001, with adaptive learning rate adjustment (learning_rate=adaptive). To increase the stability of the model, early stopping (early_stopping=True) was used with a validation data fraction of 0.1, the maximum number of iterations was 200, and the number of optimization function calculations was limited to 15,000. The input features were scaled using StandardScaler, which ensured better algorithm convergence, and the reproducibility of the results was guaranteed by a fixed random state (random_state=42). For MLP, the optimal values of hyperparameters were selected using the RandomizedSearchCV random search method, which allowed us to quickly explore the parameter space and obtain a high-quality model configuration. Figure 2 shows confusion matrices illustrating the classification results using three machine learning methods.



(a) (b) (c)

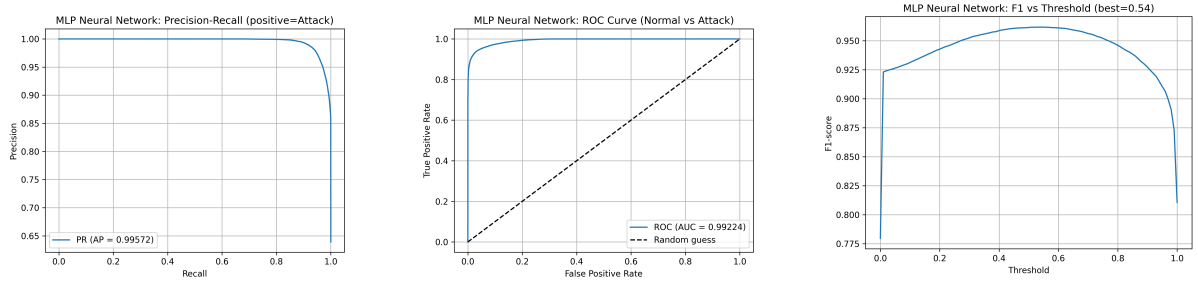
Figure 2: Normalized confusion matrices for the three classification models on the test dataset. The plots show the performance of (**a**) Random Forest, (**b**) XGBoost, and (**c**) MLP, illustrating the percentage of true positives, true negatives, false positives, and false negatives for each model.

Table 2 shows the detailed performance indicators of the models. A comparative analysis demonstrated differences in classification quality between the three methods considered. Random Forest demonstrated the worst result among the algorithms considered. The overall accuracy is 0.9358, and the average precision and recall are significantly lower compared to other models, indicating a higher number of misclassifications. The XGBoost model shows a much better balance, achieving an accuracy of 0.9473 and high precision (0.9813 for attacks) and recall (0.9685 for normal traffic) values. However, its log loss is 0.1131, indicating lower confidence in predictions compared to MLP. The MLP neural network demonstrates the highest overall accuracy of 0.9508 and the lowest log loss of 0.0793, which means more stable and confident predictions. Also, the F1-score for both classes in MLP exceeds the corresponding indicators of other algorithms, and the G-Mean value confirms the balance of classification. Thus, among the methods studied, the MLP model proved to be the most effective for classifying network traffic. In the process of evaluating the performance of the MLP model, several important visualizations were obtained, which allowed for a comprehensive analysis of its work. Figure 3 shows the Precision-Recall curve, the ROC curve with the AUC indicator, and the F1-score vs. Threshold for the MLP model.

The graph showing the dependence of the F1-score on the threshold value indicates that the best balance between Precision and Recall is achieved at a threshold of 0.54. This indicates that the correct choice of classification threshold is critical for maximizing model performance, as too low or too high values lead to a decrease in F1-Score due to an imbalance between false positives and false negatives. Analysis of the Precision-Recall curve showed very high performance with an average AP curve area of

Table 2
Detailed performance indicators for the Random Forest, XGBoost, and MLP models on the test set. The table breaks down metrics for both 'Normal' and 'Attack' classes, with the MLP model showing the best overall performance.

Model	Class	TP	TN	FP	FN	Accuracy	Recall	Specificity	Precision	F1-Score	G-Mean	Log loss
Random Forest	Normal	26456	45890	3512	1444	0.9358	0.9482	0.9289	0.8828	0.9143	0.9385	0.1425
	Attack	45890	26456	1444	3512		0.9289	0.9482	0.9694	0.9487		
XGBoost	Normal	27022	46209	3193	878	0.9473	0.9685	0.9353	0.8943	0.9299	0.9518	0.1131
	Attack	46209	27022	878	3193		0.9353	0.9685	0.9813	0.9578		
MLP	Normal	26037	47467	1935	1863	**0.9508**	0.9332	0.9608	0.9308	0.9320	**0.9469**	**0.0793**
	Attack	47467	26037	1863	1935		0.9608	0.9332	0.9622	0.9615		



(a) (b) (c)

Figure 3: Performance evaluation curves for the MLP model on the test dataset, providing a deeper analysis of its effectiveness. The plots show (**a**) the Precision-Recall curve with an Average Precision (AP) of 0.99572, (**b**) the ROC curve with an Area Under the Curve (AUC) of 0.99224, and (**c**) the F1-score as a function of the classification threshold, with the optimal threshold identified at 0.54.

0.99572, confirming the model's ability to maintain high accuracy. This result is particularly important in attack detection tasks, where an excessive number of false alarms reduces the effectiveness of the security system. The ROC curve showed similarly high classification quality, with an AUC area under the curve of 0.99224, close to the ideal value. This means that the model has excellent discriminatory power and is able to clearly distinguish normal traffic from malicious traffic. An AUC value close to 1.0 confirms that even with variable classification thresholds, the model maintains a consistently high level of accuracy. The combination of high F1-Score, AP, and AUC values indicates the high effectiveness of MLP for attack detection tasks. The visualizations obtained not only confirm the quality of the model but also provide a clear picture of its behavior in conditions of class imbalance, allowing for a reasonable assessment of its suitability for use in cyber defense systems.

4.2. Explanation of the MLP model

Ensuring the interpretability of machine learning results is one of the important tasks in the field of cybersecurity, since the transparency of decisions determines the trust of experts in the attack detection system. To explain the operation of the MLP model, two approaches from the Explainable AI arsenal were used: Permutation Feature Importance and LIME. Figure 4 shows an example of a local prediction explanation for sample #45078, classified as Attack.

The sttl (source time to live) feature made the biggest contribution to the decision, significantly pushing the model to choose the Attack class with a weight of over 0.21. The ct_state_ttl (number of connections with a similar state and TTL) and swin (TCP window size on the source side) indicators also proved to be important, strengthening the model's confidence in classifying the traffic as malicious. The indicators sloss (number of lost packets from the source), is_sm_ips_ports, and sbytes (volume of bytes transferred from source to destination) also had a positive effect on the classification. At the same time, features such as dttl (destination time to live), ct_dst_src_ltm (number of connections between the

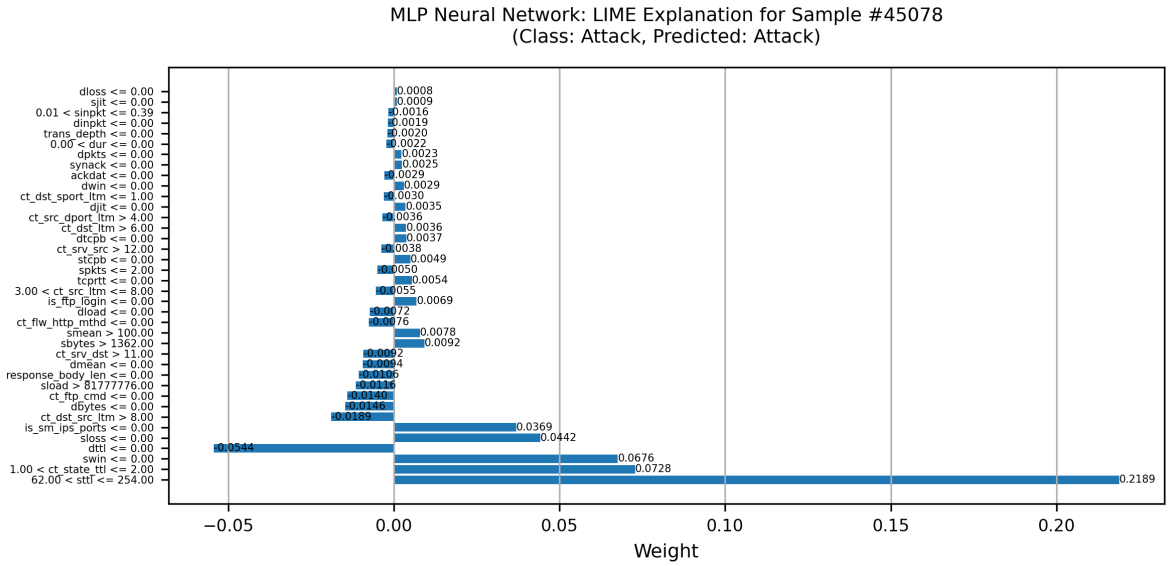


Figure 4: Local interpretable model-agnostic explanation (LIME) for a single instance (sample #45078) classified as an 'Attack'. The chart displays the features that contributed most to the prediction, with 'sttl' (source time-to-live) having the most significant positive impact (weight > 0.21), pushing the prediction towards the 'Attack' class.

same source and destination within a specified time interval), and dbytes (volume of bytes transferred from destination to source) reduced the probability of classifying an example as Attack. However, their impact was less significant. This analysis demonstrates that the model's decision was based on time parameters and protocol characteristics (TTL, TCP window, delays) and statistical properties of flows (number of packets, byte volume, losses), confirming the use of a comprehensive set of features. For global analysis, the Permutation Feature Importance method was used, the results of which are shown in Figure 5.

As can be seen, the most significant parameters for classification are sttl, ct_dst_src_ltm, and sbytes. These features determine the overall behavior of traffic and are key to detecting attacks in a network environment. The Permutation Importance results are consistent with the local explanations provided by LIME, confirming the consistency of conclusions regarding the impact of specific characteristics. The combination of global and local approaches provides a deeper understanding of how the MLP model works, allowing us to identify both general patterns in the data and the reasons for decisions in specific cases. This increases the transparency and reliability of the model, which is a prerequisite for practical use in intrusion detection systems.

4.3. Laboratory environment for evaluating IDS with MLP

The laboratory environment was deployed on a KVM hypervisor, which enabled the creation of an isolated multi-component infrastructure for simulating real network conditions and testing IDS. The network consisted of an external segment connected to the Internet and an internal segment (Figure 6).

Several operating systems functioned within the virtual environment, replicating the real corporate infrastructure. Parrot Security OS and Kali Linux were used as environments for simulating attacks and conducting penetration testing. Ubuntu Linux provided a set of basic services (HTTP, DNS, IMAP, SMTP, SSH). Metasploitable VM acted as a vulnerable machine for testing exploits and attacks. Windows Server simulated a corporate environment by providing Remote Desktop Services. It also supported key corporate functions, including Active Directory services for managing user accounts and security policies. Windows Server also had DNS and DHCP services enabled for network infrastructure, file services for resource sharing, and IIS services for testing web applications. This lab environment setup created conditions that were as close as possible to real-world scenarios. The IDS system, integrated

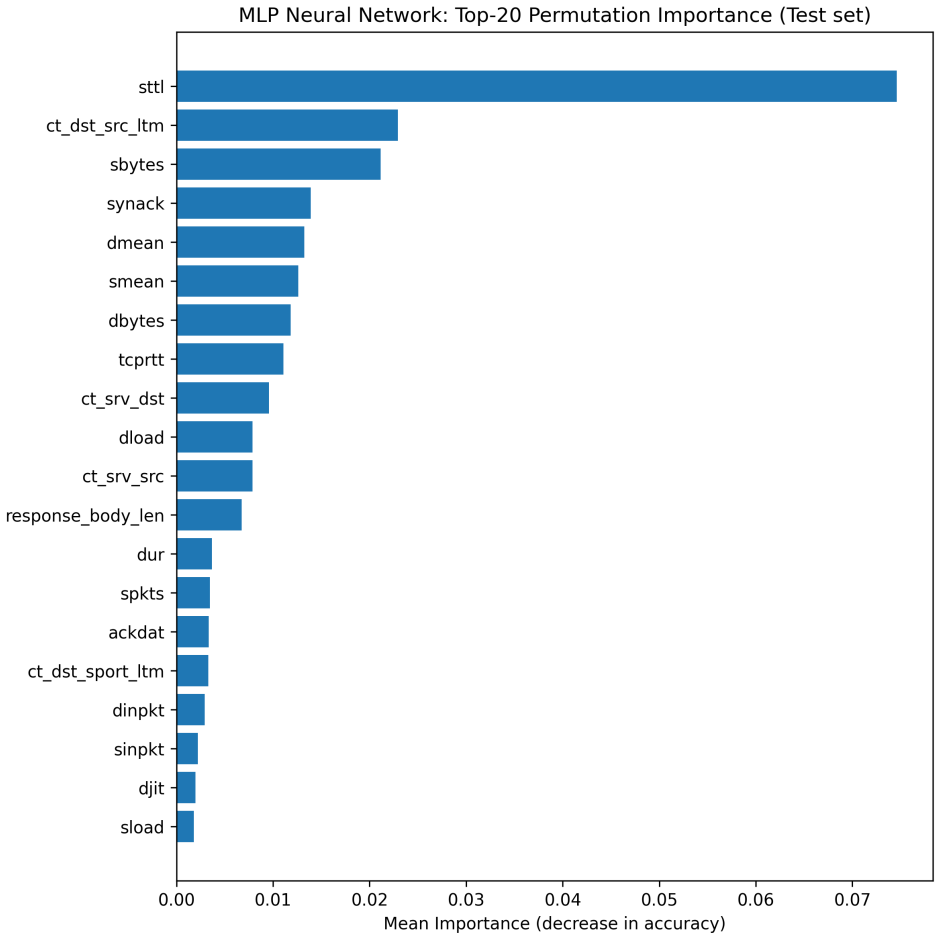


Figure 5: Top-20 most important features for the MLP model as determined by the Permutation Feature Importance method on the test set. The features are ranked by the mean decrease in model accuracy when their values are randomly shuffled. 'sttl', 'ct_dst_src_ltm', and 'sbytes' are identified as the most influential features for the model's overall performance.

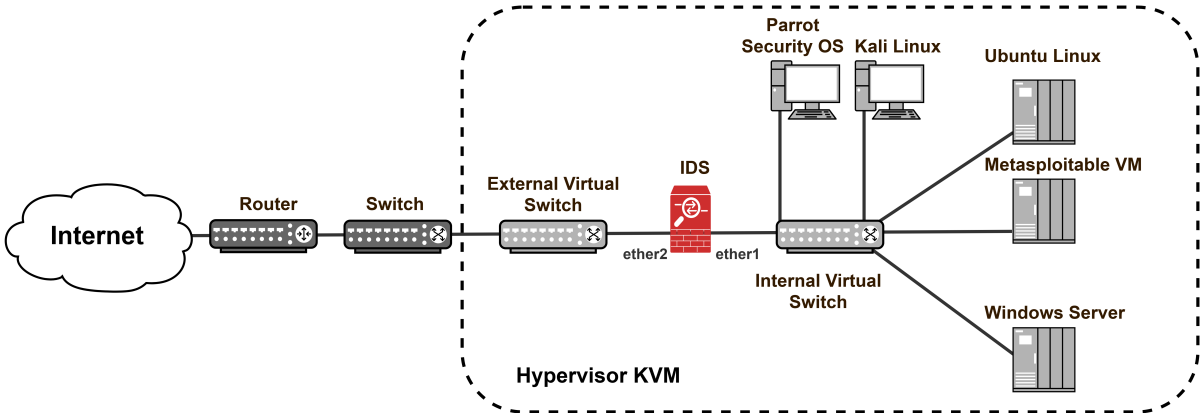


Figure 6: Schematic of the laboratory environment used for testing the Intrusion Detection System (IDS). The setup, deployed on a KVM hypervisor, simulates a corporate network with internal and external segments, various operating systems (Kali Linux, Windows Server), and vulnerable machines (Metasploitable VM) to evaluate the MLP-based IDS in a realistic setting.

into the internal segment, monitored and classified traffic using the MLP model in combination with signature-based and anomaly-based approaches. The laboratory environment made it possible to evaluate the effectiveness of MLP integration into the IDS system.

5. Conclusion

The paper demonstrates that combining machine learning methods with Explainable AI approaches can significantly improve the effectiveness and trustworthiness of intrusion detection systems. Three machine learning models were developed and compared: Random Forest, XGBoost, and MLP. The MLP neural network showed the best results. The overall accuracy was 0.9508, the smallest log loss among the models was 0.0793, the average AP value on the PR curve was 0.9957, and the AUC ROC was 0.9922. This indicates stable and confident probabilistic predictions and high discriminatory power of the model. The integration of XAI made it possible to transform the model's "black box" into a tool whose decisions can be explained. Local LIME explanations for individual samples and a global assessment of feature importance based on Permutation Feature Importance consistently pointed to the important role of TTL (sttl) parameters, connection statistics (ct_dst_src_ltm, ct_state_ttl), and data transfer volumes (sbytes, dbytes). This not only increases the confidence of security experts in automated solutions, but also provides practical guidance for IDS/IPS policy formation (e.g., thresholds, event correlation rules, alert prioritization). The KVM-based lab environment with realistic corporate infrastructure made it possible to recreate application scenarios and verify the proposed approach.

Declaration on Generative AI

During the preparation of this work, the authors used Grammarly in order to grammar and spell check, and improve the text readability. After using the tool, the authors reviewed and edited the content as needed to take full responsibility for the publication's content.

References

[1] S. Ahmed, M. M. Alshater, A. E. Ammari, H. Hammami, Artificial intelligence and machine learning in finance: A bibliometric review, Research in International Business and Finance 61 (2022) 101646. doi:10.1016/j.ribaf.2022.101646.

[2] F. Zantalis, G. Koulouras, S. Karabetsos, D. Kandris, A review of machine learning and IoT in smart transportation, Future Internet 11 (2019) 94. doi:10.3390/fi11040094.

[3] S. O. Nykytyuk, A. S. Sverstiuk, S. I. Klymnyuk, D. S. Pyvovarchuk, Y. B. Palaniza, Approach to prediction and receiver operating characteristic analysis of a regression model for assessing the severity of the course lyme borreliosis in children, Rheumatology 61 (2023) 345–352. doi:10.5114/reum/173115.

[4] O. Yasniy, P. Maruschak, A. Mykytyshyn, I. Didych, D. Tymoshchuk, Artificial intelligence as applied to classifying epoxy composites for aircraft, Aviation 29 (2025) 22–29. doi:10.3846/aviation.2025.23149.

[5] O. Yasniy, D. Tymoshchuk, I. Didych, N. Zagorodna, O. Malyshevska, Modelling of automotive steel fatigue lifetime by machine learning method, in: Proceedings of the 2nd Workshop on AI-driven Education: Innovation, Vision and Best Practices (AIE 2024), volume 3896 of *CEUR Workshop Proceedings*, 2024, pp. 165–172. URL: https://ceur-ws.org/Vol-3896/paper14.pdf.

[6] E. García-Martín, C. F. Rodrigues, G. Riley, H. Grahn, Estimation of energy consumption in machine learning, Journal of Parallel and Distributed Computing 134 (2019) 75–88. doi:10.1016/j.jpdc.2019.07.007.

[7] B. Lypa, I. Horyn, N. Zagorodna, D. Tymoshchuk, T. Lechachenko, Comparison of feature extraction tools for network traffic data, in: Proceedings of the 2nd Workshop on AI-driven Education: Innovation, Vision and Best Practices (AIE 2024), volume 3896 of *CEUR Workshop Proceedings*, 2024, pp. 1–11. URL: https://ceur-ws.org/Vol-3896/paper1.pdf.

[8] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, et al., Explainable artificial intelligence (xai): Concepts,

taxonomies, opportunities and challenges toward responsible ai, Information Fusion 58 (2020) 82–115. doi:10.1016/j.inffus.2019.12.012.

[9] D. Tymoshchuk, O. Yasniy, M. Mytnyk, N. Zagorodna, V. Tymoshchuk, Detection and classification of DDoS flooding attacks by machine learning method, in: Proceedings of the 1st Workshop on Social Media Analysis and Monitoring (SMAM 2024), volume 3842 of *CEUR Workshop Proceedings*, 2024, pp. 184–195. URL: https://ceur-ws.org/Vol-3842/paper16.pdf.

[10] Y. Klots, N. Petliak, S. Martsenko, V. Tymoshchuk, I. Bondarenko, Machine learning system for detecting malicious traffic generated by IoT devices, in: Proceedings of the 4th International Workshop on Intelligent Information Technologies and Systems of Information Security (IntelITSIS 2024), volume 3742 of *CEUR Workshop Proceedings*, 2024, pp. 97–110. URL: https://ceur-ws.org/Vol-3742/paper8.pdf.

[11] S. Lundberg, S.-I. Lee, GitHub - shap/shap: A game theoretic approach to explain the output of any machine learning model, https://github.com/shap/shap, 2017.

[12] InterpretML, Local interpretable model-agnostic explanations, https://interpret.ml/docs/lime.html, 2024.

[13] TensorFlow, Integrated gradients, https://www.tensorflow.org/tutorials/interpretability/integrated_gradients, 2024.

[14] C. S. Kalutharage, X. Liu, C. Chrysoulas, N. Pitropakis, P. Papadopoulos, Explainable AI-based DDoS attack identification method for IoT networks, Computers 12 (2023) 32. doi:10.3390/computers12020032.

[15] C. I. Nwakanma, L. A. C. Ahakonye, J. N. Njoku, J. C. Odirichukwu, S. A. Okolie, C. Uzondu, C. C. Ndubuisi Nweke, D.-S. Kim, Explainable artificial intelligence (XAI) for intrusion detection and mitigation in intelligent connected vehicles: A review, Applied Sciences 13 (2023) 1252. doi:10.3390/app13031252.

[16] A. Alzu'bi, A. Albashayreh, A. Abuarqoub, M. A. M. Alfawair, Explainable AI-based DDoS attacks classification using deep transfer learning, Computers, Materials & Continua 78 (2024) 1649–1668. doi:10.32604/cmc.2024.052599.

[17] Z. Zhang, H. A. Hamadi, E. Damiani, C. Y. Yeun, F. Taher, Explainable artificial intelligence applications in cyber security: State-of-the-art in research, IEEE Access 10 (2022) 89480–89503. doi:10.1109/access.2022.3204051.

[18] T. Zebin, S. Rezvy, Y. Luo, An explainable AI-based intrusion detection system for DNS over HTTPS (DoH) attacks, IEEE Transactions on Information Forensics and Security 17 (2022) 2339–2349. doi:10.1109/tifs.2022.3183390.

[19] M. Z. Raihan, M. S. Islam, Deep learning-based DDoS detection in SDN networks with explainable AI transparency, in: 2024 27th International Conference on Computer and Information Technology (ICCIT), 2024, pp. 1–6. doi:10.1109/iccit64611.2024.11022440.

[20] A. Albashayreh, Y. Tashtoush, A. Aldosary, O. Darwish, F. Albalas, Explainable-AI for DoS attacks detection in 5G network using deep learning models, in: 2024 International Conference on Intelligent Computing, Communication, Networking and Services (ICCNS), 2024, pp. 166–171. doi:10.1109/iccns62192.2024.10776299.

[21] S. K. Jena, A. Ranjan, V. P. Singh, DDoS attack detection using explainable AI in machine learning, in: Communications in Computer and Information Science, Springer Nature Switzerland, 2025, pp. 3–16. doi:10.1007/978-3-031-83796-8_1.

[22] S. B. Hulayyil, S. Li, N. Saxena, Explainable AI-based intrusion detection in IoT systems, Things 31 (2025) 101589. doi:10.1016/j.iot.2025.101589.

[23] S. S. Gonthina, K. Prachodhan Mudumba, M. Rao, fDoS: Explainable AI-based federated learning for DDoS detection in IoT networks, in: 2025 22nd International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), 2025, pp. 1–6. doi:10.1109/ECTI-CON64996.2025.11100775.

[24] M. Siganos, P. Radoglou-Grammatikis, I. Kotsiuba, E. Markakis, I. Moscholios, S. Goudos, P. Sarigiannidis, Explainable AI-based intrusion detection in the internet of things, in: The 18th International Conference on Availability, Reliability and Security (ARES 2023), ACM, 2023, pp.

1–10. doi:10.1145/3600160.3605162.

[25] O. Arreche, T. R. Guntur, J. W. Roberts, M. Abdallah, E-XAI: Evaluating black-box explainable AI frameworks for network intrusion detection, IEEE Access 12 (2024) 23954–23988. doi:10.1109/access.2024.3365140.

[26] A. Nascita, G. Aceto, D. Ciuonzo, A. Montieri, V. Persico, A. Pescapé, A survey on explainable artificial intelligence for internet traffic classification and prediction, and intrusion detection, IEEE Communications Surveys & Tutorials (2024). doi:10.1109/comst.2024.3504955.

[27] N. Moustafa, J. Slay, UNSW-NB15: A comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set), in: 2015 Military Communications and Information Systems Conference (MilCIS), IEEE, 2015, pp. 1–6. doi:10.1109/milcis.2015.7348942.

[28] scikit-learn developers, Permutation feature importance, https://scikit-learn.org/stable/modules/permutation_importance.html, 2024.

[29] A. Correndo, Classification performance metrics and indices, https://adriancorrendo.github.io/metrica/articles/available_metrics_classification.html, 2024.

[30] scikit-learn developers, log_loss, https://scikit-learn.org/stable/modules/generated/sklearn.metrics.log_loss.html, 2024.