ELENIDS: An EnsembLE Network-based Intrusion Detection System

1st Virgilio Cusano
DEIB
Politecnico di Milano
Milano, Italy
virgilio.cusano@mail.polimi.it

2nd Emilio Fattibene
DEIB
Politecnico di Milano
Milano, Italy
emilio.fattibene@mail.polimi.it

3rd Mariagrazia Fugini
DEIB
Politecnico di Milano
Milano, Italy
0000-0002-0692-0153

4nd Fabrizio Amarilli
Business School
Dublin City University
Dublin, Ireland
0000-0002-6307-8353

***Abstract*—To face increasing threats, Intrusion Detection Systems (IDS) demand high accuracy, short response time, and a never seen agility in recognizing evolving threats. This research explores Machine Learning (ML) with Deep Learning (DL) for IDS, and proposes a model based on ensemble voting among several classifiers. We perform testing on real-world data using an unbalanced database under a parallel setting with four classification algorithms: Decision Tree (DT), Random Forest (RF), K-nearest neighbors (KNN), and Multiple Layer Perceptron (MLP). The voting ensemble classification method is used to improve the accuracy of the model and to reduce the number of false positives. We also address the issue of explainability to increase trust in anomaly-based Network-IDS.**

***Index Terms*—IDS, Deep Learning, Explainability, Machine Learning, Ensemble Learning, UNSW-NB15 dataset.**

I. Introduction

Anomaly-based Intrusion Detection Systems (IDS) are one of the most valuable tools to face cyber security risks in organizations [1]. Nowadays, IDS can more effectively detect malicious operations and respond to a threat thanks to Machine Learning (ML) techniques [2]. However, even using ML approaches, anomaly-based IDS can have high False-Alarm-Rate (FAR) [3]. Moreover, they can suffer from low explainability [4].

To balance the above disadvantages, the approach presented in this paper uses both ML and Deep Learning (DL) to classify real-world traffic and achieves near-real time performance. The proposed *ELENIDS* (EnsembLE-based, passive, Network-based Intrusion Detection System) system uses an ensemble system, a commonly used design technique for network-based IDS, made of several classifiers based on voting. It shows capable of lowering the FAR of network-based, hybrid and passive IDS. With the aim of mitigating the limits of existing classifiers and increasing the overall performance,the voting-based ensemble system combines black-box classifiers, such as Neural Networks, with classifiers characterized by a *stronger explainability*. Our aim is to achieve the results granted by a DL approach while maintaining high explainability. Explainable Artificial Intelligence (XAI) is needed to achieve trust in the model, which leads to reduced analysis time for the human operator to verify the results of the classifier.

The novelty of the approach lies in the idea of creating an IDS with two main purposes. Initially, the hybrid model looks for unusual activity by tracking network traffic data. It also looks for patterns that change or diverge from typical behavior, as these could be signs of an attack. Second, it supports personnel in security to look into the situation and take necessary action as soon as an attack is detected. By resolving the above issues, ELENIDS enhances current IDS by increasing generalization, accuracy, and explainability. Our model improves *generalization* using a neural network, which is more flexible than other classifiers, such as for instance Decision Trees.

The paper is *centered on multiclass classification*, which is rarely offered, rather than on the accuracy achieved *per sé*. Accuracy could be improved sufficiently but not significantly, and was not investigated any further. As for XAI, we include *SHAP (SHapley Additive exPlanations)*[1], the commonly used explanation method for deep neural networks that provides detailed information about the contribution of each input feature to a given estimate produced by ELENDIS.

The paper is organized as follows. Section II presents related work and compare our approach to existing ones. Section III illustrates the ELENIDS approach, the employed dataset and the preprocessing phase. Section IV describes the classification algorithms. Section V presents the metrics used to evaluate our approach and discusses the experimental results. Section VI concludes the paper.

II. Related work

Recent research on IDS focuses on improving accuracy and adaptability to new threats, through the implementation of ML and DL techniques [5]. For example, through DL, high accuracy is achieved in the binary classification of Distributed Denial of Service - DDoS attacks. Despite progresses, most Network-IDS are signature-based, leading to a complex network exposed to zero-day attacks and to variations of known attacks. In particular, an increasing number of systems are adopting signature-based IDS rather than anomaly detection IDS due to durability of training data or high costs and error rates related to the dynamic nature of data. Progressively, IDS are evolving as market products, letting companies use them

[1]SALIH, Ahmed M., et al. A perspective on explainable artificial intelligence methods: SHAP and LIME. Advanced Intelligent Systems, 2025, 7.1: 2400304.

979-8-3315-5693-8/25/$31.00 ©2025 IEEE

as COTS, as they are becoming the one of the primary subject and direction of Internet security research. [6] is an interesting overview on IDS, comparing different classifiers performances on three different datasets. This is one of the few studies that compares not only binary vs multi-class classification, but also computation time. It is noted that multi-class classification lacks in precision and is more computationally demanding compared to binary classification. [7] proposes an hybrid DL-based IDS. The model is heavily focused on DL and achieves a 99.64% accuracy. While performance is outstanding, the computational time is not discussed, and the used dataset is not well-known. [8] explores a weighted voting ensemble method using DL-based classifiers. This model, tested using the CICIDS-2017 dataset, is unable to identify Web Attacks but produces outstanding results in all the other categories. Also in this paper, computational time, as well as XAI, is not discussed, which could be problematic when using a DL-based ensemble method. [9] proposes a supervised learning-based voting classifier. The model produces optimal results, but the chosen dataset was collected more than 25 years ago and contains only four categories of attacks. [10] proposes an ensemble technique composed of eleven models with classifiers both ML-based and DL-based. Memory-managing classifiers such as LightGBM are also explored. In a nutshell, DL-based IDS and voting-based IDS are both explored in the literature and provide promising results. On the other hand, their computational cost is seldom explored and discussed, even though is extremely relevant in NIDS. XAI, which is crucial for passive IDS, is beginning to be considered.

III. ELENIDS: Aims, Dataset and Pre Processing Phase

A. Aims

To improve intrusion detection and strengthen network security, the ELENIDS method:

- aims at handling a wider range of attack detection wrt most IDS currently in place. This holds for IDS that do not employ the DL approach;
- is able to perform near-real-time detection;
- handles multiclass classification, which is seldom implemented;
- directly addresses explainability for IDS;
- tackles the problem of high false positive rate seldom addressed by current IDS, which produce many false alerts, so creating problems for security staff. We use multiclass classification to identify the causes of events and reduce the number of false positives.

To address the shortcomings of current IDS, our primary objectives are to reduce the false positive alert rate, to improve accuracy, to generalize to unseen threats, and to improve explainability.

B. The Dataset

ELENIDS employs the UNSW-NB15 dataset [11], one of the (few) publicly available for network-based anomalies, which gained strong acknowledgments in literature. UNSW-NB15 is one of the most recent and rich (in number of different attacks) available systems. The Cyber Range Lab of UNSW was generated and captured using tcpdump 100GB of raw traffic. UNSW-NB15 contains a total of two million and 540,044 records. In ELENIDS, a portion of this dataset has been used as a training and testing set. This portion is divided in 41 features and contains ten distinct categories of attacks. This database was chosen from six different databases found in the literature (KDD99-Cup, UNSW-NB15, CIC-IDS 2017, CSE-CIC-IDS 2018, BOT-IoT, CIDDS-001/CIDDS-002). The main reasons for selection of this database are the following:

1) it contains a collection of both manufactured and real-world traffic, making it realistic;
2) it was collected using publicly available tools (tcpdump), which allows for recreation of the collection process in a real network;
3) the dataset is one of the most recent and contains common attacks.

C. Preprocessing

Data preprocessing, a fundamental step in ML applications, helps reducing the time needed for training, and helps avoiding some of the classifiers' limitations, such as data dimensionality-related problems. Commonly used classification algorithms, such as Decision Tree and Random Forest, provide high accuracy but can be prone to problems such as overfitting. To face these problems, feature selection is used. The main advantages of using feature selection are:

- Avoiding the curse of dimensionality.
- Avoiding overfitting.
- Reducing the comprehensive time required for the model fitting and prediction.
- Facilitating the model interpretability.

The dataset used in this research was already preprocessed. Additional data preprocessing was applied to convert string values to numerical. Moreover, a scaler is applied before every classification algorithm to improve its performance. Each scaler was chosen based on the resulting performances of the model.

IV. Classification algorithms

The classifiers used in the ensemble system were selected based on performances, time-complexity, and explainability. The employed algorithms are Decision Trees (DT), Random Forest (RF), and K-Nearest Neighbor (KNN) with a scaler used to increase its classifying capability. KNN's performances were measured before the scaler application and after applying different scalers, to find the optimal scaler and the optimal number of neighbors. MinMaxScaler was found to be the optimal scaler for our problem. Finally, a Multi-layer Perceptron (MLP) is employed [12]. The time complexity T of a MLP can be approximated following this formula:

$$T = ((n-1)/2)^2 \in O(^2). \tag{1}$$

Given the fact that the number of neurons n for a given problem can be regarded as a constant, the overall complexity is:

$$O(n^2) = O(1). \tag{2}$$

where n is the number of neurons that compose in the network.

A. Ensemble system

We used a parallel ensemble method to maintain the same time complexity of the classifiers. The parallel ensemble method's time complexity can be expressed as

$$T_{ensemble} = \max(\{T_i | i \in [0, ..., n]\}) \tag{3}$$

with n being the number of classifiers used by the ensemble system, and T_i the time complexity of the $i-st$ classifier. Its architecture is depicted in Figure 1.

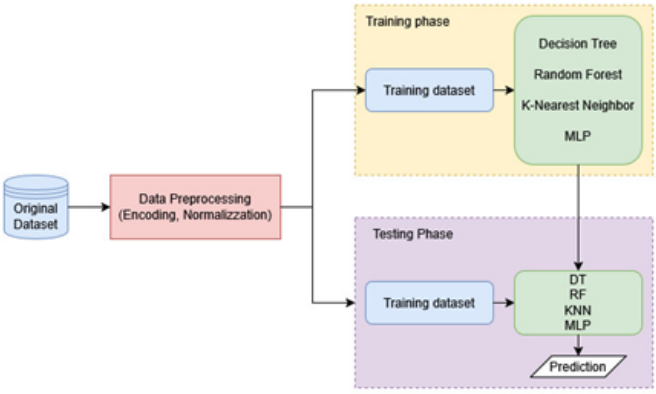

Fig. 1. Architecture of the Ensemble System

Regarding explainability, an issue of strong interest in current research in IDS [13], we follow our approach presented in [14] to provide an understandable explanation of predictions. This issue is crucial in passive IDS for human operators who need to interpret the results since it reduces the overall analysis time and increments the accuracy of the human operations.
To proceed towards model explainability, first, we chose to deliver the output of ELENIDS in a xlsx format, which is relatively easy to interpret, providing only the necessary and most meaningful information about the identified malicious transaction.

Second, the model can provide additional information about its reasoning. To achieve this, model-specific methods, such as decision paths and feature importance graphs, can be produced. Model-agnostic techniques, such as SHAP, are used to explain both the reasoning of the model and the reasoning of the single classifiers. Model-agnostic techniques help to explain the reasoning of complex classifiers, such as MLP. Combining model-specific methods and model-agnostic techniques helps to increase the explainability of the model. Although the output is always provided, producing this additional information may be more time-consuming and is provided on request. We employ SHAP graphs which can be requested for both the model and the single classifiers based on the user request and can be queried for both binary and multiclass classification.

V. Used Metrics and Experimental Results

To evaluate our model, we used the following metrics.

- Accuracy: the proportion of all correct classifications over all the classifications.
- Precision: measures the proportion of correctly predicted positive cases out of all cases predicted as positive.
- Recall: measures the proportion of correctly predicted positive cases out of all actual positive cases.
- F_1: combines precision and recall, providing a balanced measure of how well the model performs.
- FAR (False Alarm Rate): used to determine how many negatives were classified as positives. In our case, it is the amount of attacks that went undiscovered.

Once classified, data are divided into True Positives, True Negatives, False Positives and False Negatives, each computed as usually defined (see [15]).

An additional used metric is *time*. Classification time and detection time, which are, respectively, the duration required to train the model and the duration required to categorize a sample using the trained model, are seldom found in literature. This metric is extremely relevant in our research and for IDS, since the main goal of an IDS is to achieve the highest precision with the lowest time consumption. Long classification and detection time may indicate that the used model complexity is too high.

VI. Results

Results show that the Ensemble method yields significantly better performance than the individual classifiers. The accuracy, recall, and F1 score improve, while FAR, which is crucial for a passive IDS, is reduced. The fitting and testing time follows the worst classifier, which is reasonable considering the time complexity of the ensemble method. Accuracy and Precision are reported in Fig. 2 and 3, respectively. Two distinct models are proposed: the first, ELENIDS, employs all four classifiers; the second model, F-ELENIDS, excludes the KNN.

Table I reports the computed results, divided by metric and classifier. The results are proposed for both binary (B) and multiclass (M) classification.

Figures 4 to 9 report the results in terms of the various considered metrics. Time-related statistics are measured in seconds and reported using a logarithmic scale. The results show an improvement of 0.5 to 1 percentage points in accuracy, recall, precision, and F1-score in multiclass classification between the individual classifiers and the proposed systems, while the FAR is slightly reduced. The ELENIDS model performs worse in terms of FAR while providing a good tradeoff and being more convenient than using a single classifier.

Considering that ML-based systems usually achieve lower accuracy and higher false alarm rates, but are more agile and can discover unknown attacks, and that a downside of

	DT	RF	MLP	KNN	ELENIDS	F-ELENIDS
Accuracy (B)	97.6440	96.5178	97.209	94.7844	97.7656	97.8834
Precision (B)	97.66	96.51	97.20	94.90	97.76	97.88
Recall (B)	97.644	96.517	97.209	94.784	97.765	97.883
F1-score (B)	97.648	96.518	97.208	94.81	97.76	97.884
FAR (B)	2.1	2.2	3.6	5.4	1.5	1.7
Fitting time (s) (B)	3.101364	55.27883	118.72924	0.19008	71.41491	111.06709
Classification Time (s) (B)	0.034102	0.92052	0.10433	19.21020	15.87057	17.43396
Accuracy (M)	84.2744	83.7980	84.3326	79.4161	85.0278	85.2543
Precision (M)	84.32	84.21	83.70	78.21	84.87	84.93
Recall (M)	84.27	83.79	84.33	79.41	85.02	85.25
F1-score (M)	82.98	81.32	83.19	78.39	83.50	83.49
FAR (M)	2.84	4.80	3.46	5.91	2.99	2.69
Fitting time (s) (M)	1.7491	22.102	62.191	0.0887	74.689	70.608
Classification Time (s) (M)	0.0328	0.5425	0.0743	24.310	21.112	0.9809
Macro AUC (M)	0.96	0.97	0.97	0.86	0.97	0.97

TABLE I
PERFORMANCE STATISTICS

these systems is their computational complexity, which is very costly, finding the optimal model is a research challenge which can help in reducing cost. Accordingly, in this study we investigated classifiers that would not overload the system. We could not fully aim at overcoming this problem, which is known to require tools different from the voting ensemble.

The results of the ELENIDS system show the achievement of high precision in near-real time by combining several classifiers aimed at high precision in binary or multiclass classification of malicious traffic.

By combining both DL and ML approaches, our classifier can provide explainability while lowering the FAR and achieving high levels of accuracy and precision, even though lower than the results obtained for binary classification in multi-class classification for attack-type classification.
Explainability is provided in terms of the format of the model output, namely a .xlsx format file containing seven features (IP_{sender}, $IP_{reciever}$, Protocol, Service, Duration, Rate, Attack category). This format can be more easily explored and changed both in the file format and in the features. Regarding the format, a .csv was also explored to reduce the report production time, but the .xlsx format was chosen to improve readability. The use of an ensemble of model-agnostic instruments, such as SHAP, and model-specific instruments, increases the explainability of the model, supporting better trust in the model and additional explainability. The aim of instruments like SHAP is to facilitate the human operator decision making based on the classification results.

VII. Concluding Remarks

This paper has presented an IDS model based on an ensemble of ML approaches, in the line of anomaly-based IDS currently proposed in the literature and used in practice.

Accuracy, FAR, and time are the focus metrics in our research and have been analyzed in the paper. However, accuracy could not be achieved sufficiently, even though we worked with multiclass classification, which is rarely offered.

The paper has presented the approach, the results, and the issue of XAI.

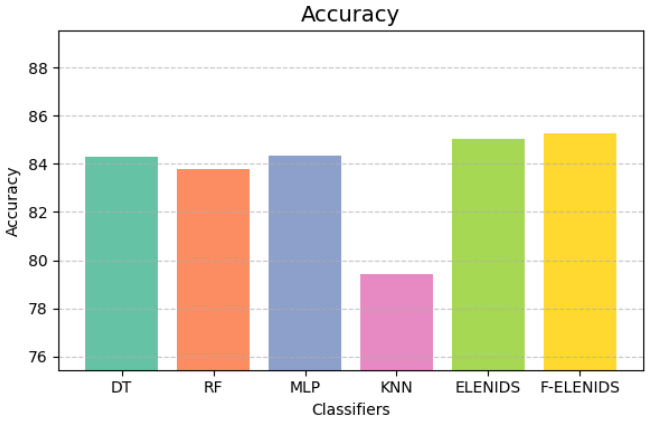

Fig. 2. Accuracy multiclass classification

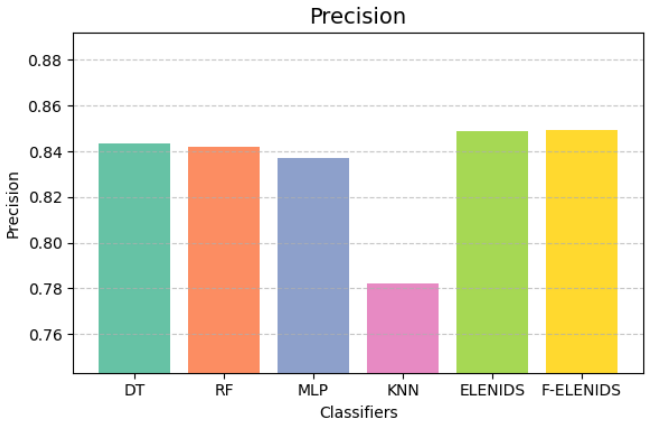

Fig. 3. Precision for multiclass classification

Future research will consider implementing a more sophisticated DL algorithm, compared to the employed MLP, and improving the data preprocessing process. We investigated the optimal MLP for this problem; research has shown that a Convolution Neural Network (CNN) could be more performative for this problem. This is an open debate in the literature (see for instance [16]). Several proposals are currently present [17]. Still, time complexity is scarcely explored as an essential metric.

Future research should also implement more time-efficient classifiers (e.g., XGBoost or Trees) or using CUDA-based

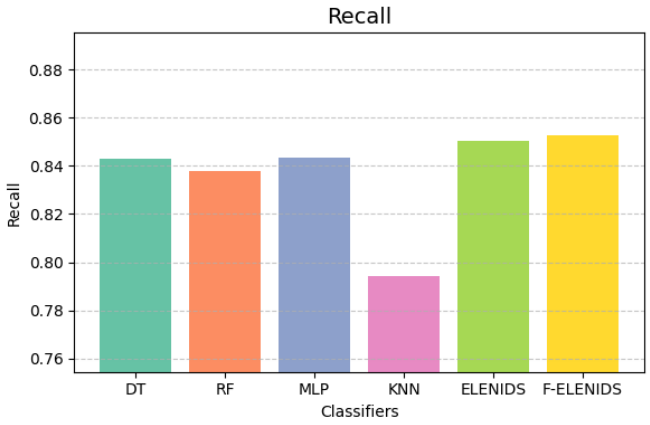

Fig. 4. Recall for multiclass classification

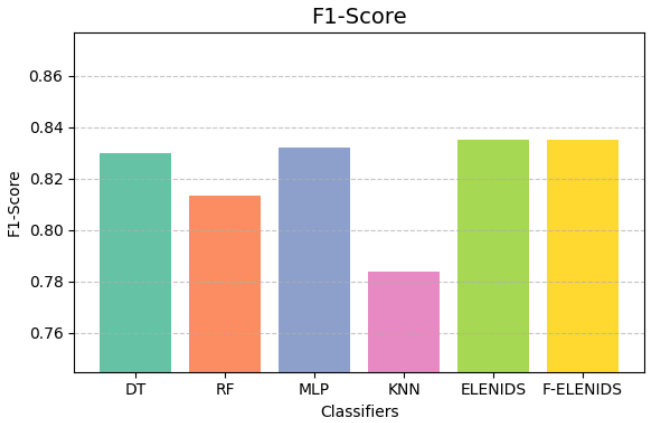

Fig. 5. F1-Score for multiclass classification

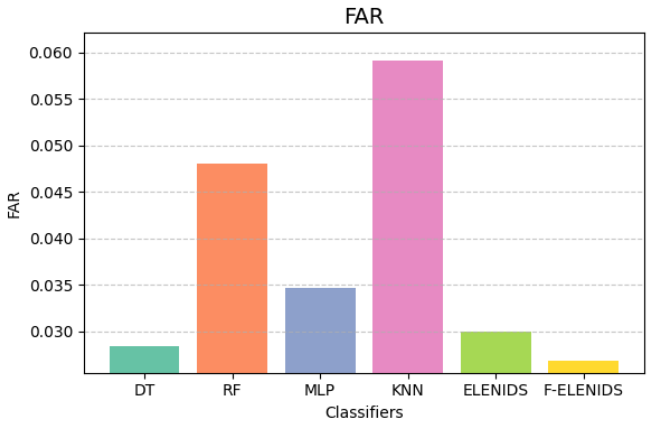

Fig. 6. FAR for multiclass classification

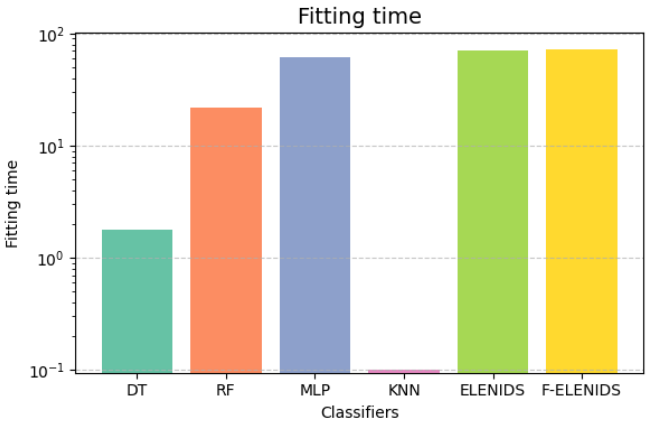

Fig. 7. Fitting time for multiclass classification in seconds

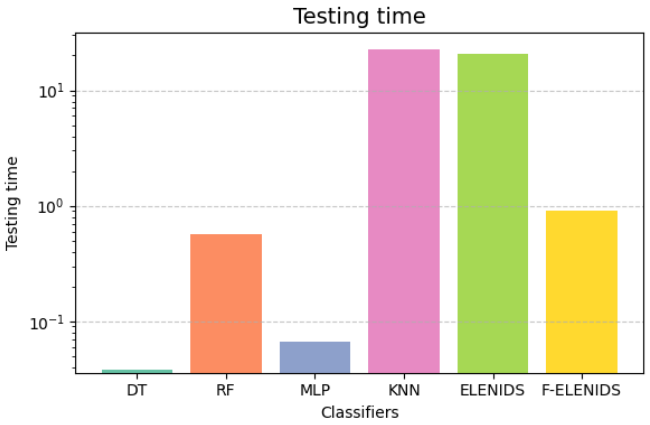

Fig. 8. Testing time for multiclass classification in seconds

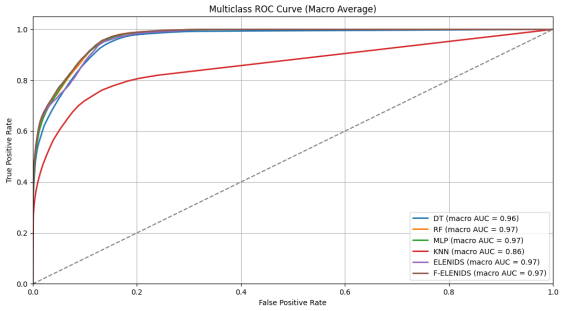

Fig. 9. ROC for multiclass classification

solutions. Finally, the proposed model could be implemented with low effort as a server solution or as an on-device program.

REFERENCES

[1] M. Fugini *et al.*, "Cyber risk and cyber security: cyber access control with data mining," *OPEN ACCESS BIOSTATISTICS & BIOINFORMATICS*, vol. 3, no. 5, pp. 1–21, 2024.

[2] E. Altulaihan, M. A. Almaiah, and A. Aljughaiman, "Anomaly detection ids for detecting dos attacks in iot networks based on machine learning algorithms," *Sensors*, vol. 24, no. 2, p. 713, 2024.

[3] S. M. Hussein and A. M. Ashir, "Machine learning-driven intrusion detection systems: Reducing false alarms and enhancing accuracy," *EURASIAN JOURNAL OF SCIENCE AND ENGINEERING*, vol. 10, no. 3, pp. 85–96, 2024.

[4] A. Kumar and V. L. Thing, "Evaluating the explainability of state-of-the-art machine learning-based online network intrusion detection systems," *arXiv preprint arXiv:2408.14040*, 2024.

[5] S. Racherla, P. Sripathi, N. Faruqui, M. A. Kabir, M. Whaiduzzaman, and S. A. Shah, "Deep-ids: A real-time intrusion detector for iot nodes using deep learning," *IEEE Access*, 2024.

[6] A. Thakkar and R. Lohiya, "A survey on intrusion detection system: feature selection, model, performance measures, application perspective, challenges, and future research directions," *Artificial Intelligence Review*, vol. 55, no. 1, pp. 453–563, 2022.

[7] V. Hnamte and J. Hussain, "Dependable intrusion detection system using deep convolutional neural network: A novel framework and performance evaluation approach," *Telematics and Informatics Reports*, vol. 11, p. 100077, 2023.

[8] V. Ciric, M. Milosevic, D. Sokolovic, and I. Milentijevic, "Modular deep learning-based network intrusion detection architecture for real-world cyber-attack simulation," *Simulation Modelling Practice and Theory*, vol. 133, p. 102916, 2024.

[9] M. A. Khan, N. Iqbal, H. Jamil, D.-H. Kim *et al.*, "An optimized ensemble prediction model using automl based on soft voting classifier for network intrusion detection," *Journal of Network and Computer Applications*, vol. 212, p. 103560, 2023.

[10] A. K. Phulre, S. Jain, and G. Jain, "Evaluating security enhancement through machine learning approaches for anomaly based intrusion detection systems," in *2024 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS)*. IEEE, 2024, pp. 1–5.

[11] N. Moustafa and 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)*, 2015, pp. 1–6.

[12] R. Kruse, S. Mostaghim, C. Borgelt, C. Braune, and M. Steinbrecher, "Multi-layer perceptrons," in *Computational intelligence: a methodological introduction*. Springer, 2022, pp. 53–124.

[13] S. Neupane, J. Ables, W. Anderson, S. Mittal, S. Rahimi, I. Banicescu, and M. Seale, "Explainable intrusion detection systems (x-ids): A survey of current methods, challenges, and opportunities," *IEEE Access*, vol. 10, pp. 112 392–112 415, 2022.

[14] F. Amarilli, S. Uboldi, F. Saraceni, and L. Tencati, "Managing paradoxical tensions in the implementation of explainable ai for product innovation," in *2025 33rd International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE)*, 2025, pp. 1–6.

[15] G. Meena and R. R. Choudhary, "A review paper on ids classification using kdd 99 and nsl kdd dataset in weka," in *2017 International Conference on Computer, Communications and Electronics (Comptelix)*. IEEE, 2017, pp. 553–558.

[16] R. Hegde and S. Soumyasri, "Code-ids: Convolutional neural network based intrusion detection system using deep learning."

[17] G. Nalinipriya, S. Rama Sree, K. Radhika, E. Laxmi Lydia, F. K. Karim, M. K. Ishak, and S. M. Mostafa, "Leveraging explainable artificial intelligence for early detection and mitigation of cyber threat in large-scale network environments," *Scientific Reports*, vol. 15, no. 1, p. 24662, 2025.