Explainable-AI for DoS Attacks Detection in 5G Network using Deep Learning Models

Amjad Albashayreh, Yahya Tashtoush, Abdallah Aldosary, Omar Darwish and Firas Albalas
Department of Computer Science, *The University of Jordan*, Amman, Jordan
amalbashayreh20@cit.just.edu.jo
Department of Computer Science, *Jordan University of Science and Technology*, Irbid, Jordan
yahya-t@just.edu.jo
The Department of Computer Engineering, Prince Sattam bin Abdulaziz University, Ar Riyadh, Saudi Arabia
ab.aldosary@psau.edu.sa
Information Security and Applied Computing Department, *Eastern Michigan University, Ypsilanti*, MI 48197, USA
odarwish@emich.edu
Department of Computer Science, *Jordan University of Science and Technology*, Irbid, Jordan
faalbalas@just.edu.jo

***Abstract*—With the emergence of the fifth-generation (5G) network, numerous revolutionary applications are enabled, including low-latency and machine-type communications. This great increase creates a broader security threat concerns, such as denial-of-service (DoS) attacks, which can disrupt network functionality. The complexity and decentralization of 5G networks create new vulnerabilities for adversaries, necessitating comprehensive security procedures to identify, mitigate, and prevent DoS attacks in 5G networks. This paper introduces a novel approach for DoS detection in 5G networks, utilizing deep learning and machine learning models, along with Local Interpretable Model-Agnostic Explanations (LIME), to interpret model predictions and identify the significant role of data features in detecting DoS attacks. The results revealed that the random forest model demonstrated superior recall of 99.98, while BiLSTM demonstrated exceptional performance with a recall of 98.02.**

***Index Terms*—DoS, Explainable-AI, DL, LIME,5G.**

I. Introduction

In recent years, the number of network devices has increased, which has led to an expansion of potential cyber-attacks, prompting the development of intrusion detection systems to ensure system security amid growing cybersecurity concerns [1]. There are many different types of cyberattacks that can lead to system breakdowns, such as denial-of-service (DoS) attacks. DoS attacks are highly disruptive cyberattacks that overwhelm network resources, making services unavailable to legitimate users. The detection and mitigation of DoS attacks are crucial for ensuring network security and service availability. The security of digital infrastructure is critical, especially with the emergence of 5G networks. 5G technology promises extraordinary speed, huge connectivity, and low latency, changing industries including smart cities, healthcare, and transportation. The concept of a global network of intelligent devices with processing, communication, and sensing capabilities holds immense potential for various industries, including industrial automation and healthcare [2]. However, this advancement creates new issues and obstacles in network security [3] [4]. Prior work used artificial intelligence techniques for DoS attack detection in 5G networks, frequently employing the same datasets due to the limited availability of data sources that reflect the unique characteristics of this type of network. AI approaches, including machine learning (ML) and deep learning (DL), help improve detection by spotting aberrant patterns and making faster, more accurate conclusions [5]. Network Intrusion Detection Systems (NIDS) are becoming increasingly popular due to their usage of ML and DL approaches and availability from third-party vendors, allowing networking enterprises to cut expenses and focus on core goods [6]. Yet, the application of AI poses questions concerning openness and interpretability. Explainable AI (XAI) methods such as LIME address these problems by simplifying AI decision-making procedures for human operators. This is critical in network security since understanding the reasoning behind AI-driven detections allows for more timely and effective responses to threats. Advancements in AI, particularly machine learning, have significantly transformed data analysis and decision-making processes in various real-world applications, enabling automation and intelligent decision-making through data combination and analysis [7], [8]. As a result, using deep learning models in conjunction with explainable AI approaches to novel datasets, including a wide range of network flows captured in 5G environments, will be a new addition to that field. The main contribution of this paper is proposing a new approach that utilizes deep learning and machine learning models to detect DoS attacks in 5G network flows. Also, the paper employs the explainable AI technique Local Interpretable Model-Agnostic Explanations (LIME) to interpret model predictions, providing a comprehensive un-

979-8-3503-5469-0/24/$31.00 ©2024 IEEE

derstanding of the model learning process. This approach not only interprets the decision-making process but also identifies the significant contribution of data features in detecting DoS attacks. The reset of this paper is organized as follows: Section II demonstrates the related work; Section III explains the methodology; Section IV discusses the experimental results; and Section V concludes this paper.

II. Related Work

In recent years, the researchers have conducted significant efforts to propose new techniques for DoS attack detection, using machine learning and deep learning algorithms. Kalutharage et al. [9] presented a new explainable AI technique to detect DDoS attacks by analizing network traffic at the network layer. The authors extracted the model using different datasets and compared its performance with random fields, decision trees, and deep neural networks. The model outperformed all the other models with an accuracy of 98%. Yang Xiang et al. [10] developed an entropy-based approach for detecting low-rate DDoS attacks. The approach detects low-rate attacks by comparing the normalized entropy of regular and low-rate traffic probability distributions. Tcydenova et al. [11] designed a novel framework for detecting adversarial attacks using machine learning and explainable AI techniques. The framework comprises two phases: initial operation and detection. During the initialization phase, the training process involves the use of a support vector machine model and LIME technique. In the detection step, the generated explanations are evaluated to determine whether an adversarial attack occurred. The authors gort an accuracy of 96.84%. Sauka et al. [12] proposed a deep learning-based network intrusion detection system that employs adversarial training and AI approaches. Their experiments revealed that the PGD adversarial-trained model was more resilient than the DeepFool and FGSM models, with a ROC-AUC of 87%. The FGSM assault had no effect on the PGD model, however the DeepFool and PGD attacks both lowered the ROC-AUC of the FGSM model. Keshk et al. [13] designed a new intrusion detection system for IoT networks based on a Short-Term Long Memory (LSTM) model. The framework trains and evaluates the LSTM model using a unique SPIP framework, with input features extracted from the NSLKDD, TON_IoT, and UNSWNB15 datasets. The framework outperformed other similar methodologies with an accuracy of 87.30%.

Siganos et al. [14] developed an AI IoT IDS using ML, DL, SHapley Additive Explanations, achieving 99.99% accuracy in performance detection using random forest. Rao et al. [15] proposed a zero-shot strategy for classifying new threats based on feature influence. The system effectively distinguishes attack traffic from normal flow and creates labels for attacks based on contributing characteristics. These labels are straightforward for SIEM analysts and can help them identify the type of assault. The technique was tested on a network flow dataset, yielding results for specific attack types. The authors got an accuracy of 92.00%. Arreche et al. developed an explainable AI framework to improve the interpretability of AI models in network intrusion detection applications. They compare seven models on three different real-world datasets, each with its own set of features and problems. The system generates local and global explanations, recognizes model- and intrusion-specific aspects, and detects overlapping features that affect various AI models. It also detects common patterns across detection methodologies and has a low computational overhead, making it suitable for real-time applications. The authors achieved an accuracy of 99.00%. Mallampati et al. [16] created a data preprocessing approach to increase a model's generalizability. The authors used k-Means SMOTE to address class dissimilarity, proposed a hybrid feature selection approach, and examined a Light Gradient Boosting Machine with hyperparameters tweaked. The experiments on the UNSW-NB15 and CICIDS-2017 datasets produced an accuracy of 90.71% and 99.98%, respectively. Alzu'bi et al. [17] proposed a deep learning method for detecting distributed denial of service (DDoS) attacks in IoT environments. The authors applied deep transfer learning and evaluated the models' accuracy and time complexity using two different datasets. Also, they used multiple deep architectures and explainable artificial intelligence (XAI) techniques to conduct binary and multiclass experiments. The results showed the method's effectiveness, with a recall of 99.39% achieved using the XAI-BiLSTM model. Jiyad et al. [18] introduced a novel ensemble model for detecting DDoS attacks using machine learning algorithms. The authors utilized SHAP and LIME tools to enhance model readability and transparency. The XGBoost ensemble model surpasses conventional classifiers, with an exceptional accuracy rate of 97%.

III. Methodology

A. 5GC PFCP Dataset

The 5GC PFCP Intrusion Detection Dataset was created utilizing an experimental 5G testbed with Open5GS as the cellular core and UERANSIM as the NG-RAN. The testbed used a variety of 5G network functions, such as Network Slice Selection, Network Repository, Policy Control, User Data Management,Network Exposure, Access Management, and Authentication Server. Suspicious behaviors were carried out to create PFCP attacks with a virtualized UE, a virtualized gNodeB (gNb), and an attacker instance. This paper uses network flows obtained from transport layer and contains four types of DoS attacks: establishment DoS, deletion DoS, modification flood (DROP), and modification flood (DUPL). Establishment DoS floods UPF with valid session establishment and heartbeat requests, depleting resources. Delete DoS disconnects a single UE from the DN by focusing on PDU sessions. The modification flood (DROP) evaluates packet handling rules to dissociate a specific UE from the DN. The modification flood (DUPL) causes UPF to reproduce session rules, resulting in duplicated communication over the N6 interface. Table I summarizes the dataset statistics.

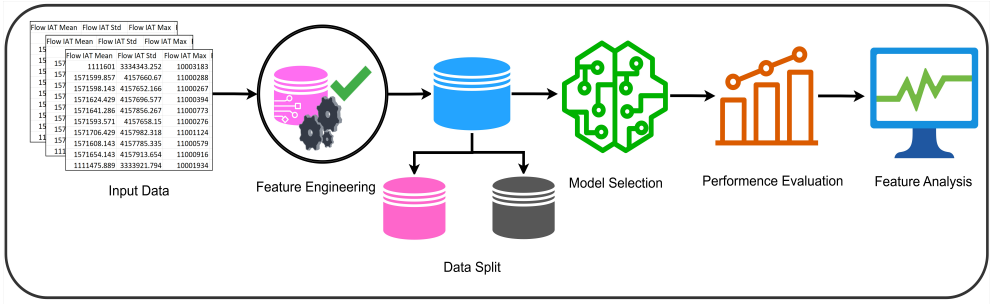

Figure 1. A visual representation of the proposed pipeline for DoS attacks Detection.

Table I
A SUMMARY OF THE OF 5GC PFCP DATASET STATISTICS

Type	Training	Validation	Testing	Total
Estab	4,659	518	1,295	6,472
Mod1	4,660	518	1,294	6,472
Mod2	4,660	517	1,295	6,472
Deletion	4,660	518	1,294	6,472
Benign	4,660	518	1,295	6,473
Total	23,299	2,589	6,473	32,361

B. Data Preprocessing

This paper follows a systematic approch for extensive data preparation to format the data to be fed into machine learning (ML) and deep learning (DL) algorithms. Figure 1 presents a visual representation of the followed pipeline to detect DoS attacks. The pipeline starts with collecting network flows from the transport layer and storing them in a single CSV file. Then several data preprocessing steps were conducted, starting with handling missing values (fillNaNs) and performing feature engineering techniques.Feature engineering encompassed both feature selection and feature standardization. The random forest (RF) classifier has been used to select the top 20 most significant features for usage in the training process. The RF selects features for a model based on their significance in predicting the target variable. Multiple decision trees are used to ensure the model's diversity. The relevance of each feature is determined using permutation importance, which involves shuffled features to test their impact on model correctness. The aggregate significance score from all trees is then used to identify the most significant characteristics for the model. Figure 2 displays the selected features and their impotence determined using RF. During the feature standardization step, we scaled the data features using the standard scaler technique to ensure they were appropriate for the model. The standard scaler converts each feature to a zero-centered, one-standard deviation distribution. It computes the mean and standard deviation for each feature in the dataset, then subtracts the mean and divides by the standard deviation. This ensures that all features have the same scale, which is critical in machine learning and deep learning methods. This technique accelerates algorithm convergence during training and keeps larger numerical ranges from dominating smaller ranges, hence boosting model performance and stability. However, the input data has been rearranged into the ideal format required by the models. The data was then divided into three subsets: 70% for training, 10% for validation, and 20% for testing. These strategies guaranteed that the data was appropriately processed for efficient ML and DL model training, allowing for accurate and interpretable DoS attack detection in 5G networks.

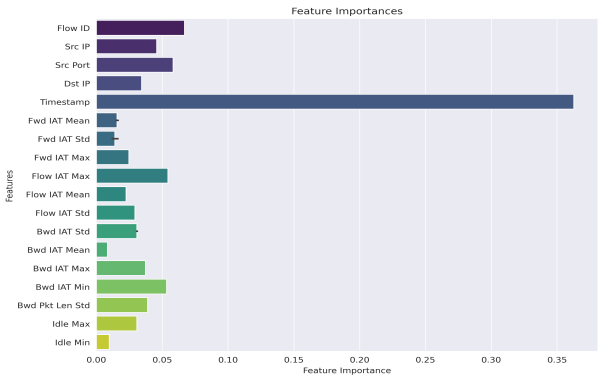

Figure 2. A visual representation of the feature importance using RF classifier.

C. ML and DL Models

This research uses six different machine learning and deep learning techniques to detect Denial-of-Service (DoS) attacks. The machine learning models used include Decision Trees (DT) [19], Random Forests (RF) [20], and Multilayer Perceptron (MLP) [21], which were chosen for their ability to successfully categorize and identify DoS attack patterns based on network traffic data attributes. Furthermore, the deep learning models used include Bidirectional Long Short-Term Memory (BiLSTM) [22], Bidirectional Gated Recurrent Unit (BiGRU) [23], and Bidirectional Recurrent Neural Network (BiRNN) [24], which were chosen for their ability to capture complex temporal dependencies and intricate relationships within sequential data, which are frequently observed in network traffic during DoS attacks. The study aims to enhance detection accuracy and robustness against various types and intensities of DoS attacks using these algorithms, contributing to the evolution of cybersecurity measures.

D. Evaluation Metrics

The performance of the used machine learning and deep learning algorithms is evaluated using metrics such as accuracy, precision, recall, and F1-score. Accuracy measures the ratio of correct predictions over the total number of prediction instances.

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP} \tag{1}$$

Recall is the true positive rate, which measures the ratio of true-positive results out of all actual true positive and false negative results [25].

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

$$Precision = \frac{TP}{FP + TP} \tag{3}$$

F1-score is the harmonic mean of precision and recall [25].

$$F1 - score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{4}$$

E. LIME Analysis

The Local Interpretable Model-agnostic Explanations (LIME) [26] is a machine learning technique that leverages a dataset of altered cases around a certain prediction to provide interpretable and locally faithful explanations. It approximates the behavior of a black-box model in a specific location while monitoring how the model's predictions evolve. A simpler model, such as linear regression, is then used to the modified dataset, with each instance weighted by its closeness to the original occurrence. This simpler model offers information about the most influential aspects in the prediction. LIME improves transparency and confidence in machine learning models by assisting researchers in understanding model decisions and potential biases on a local scale. The LIME method can be used in different fields with various types of data including text, image, and tabular data.

IV. Experimental Results and Discussion

In this paper, several machine learning and deep learning algorithms have been used to detect malicious network flows. The machine learning algorithms are RF, DT, and MLP. On the other hand, the deep learning models that are used are BiLSTM, BiGRU, and BiRNN. All the experiments are conducted on the Colab platform. Several experiments are performed to select the best hyper-parameters for the deep learning models. The categorical cross entropy is used as a loss function, and the Adam optimizer is used with a learning rate equal to 0.01. Each model is trained for seven epochs with a batch size of 16. Table II summarizes the hyper-parameters for deep learning models. However, the performance of machine learning and deep learning algorithms was very close. The random forest classifier outperformed all the other models with an accuracy of 99.98. The RF surpassed the other models that were considered to have more complexity, such as BiLSTM, BiGRU, and BiRNN, due to its robust ability to handle overfitting owing to its ensemble nature. Also, its outstanding performance makes it suitable to use it in the feature selection step, where the best features can be handled by it. Therefore, it is important to consider the performance of the deep learning models, which will not be affected by feature selection steps such as RF and DT. The BiLSTM outperformed the other deep learning models with an accuracy of 98.02, precision of 98.09, recall of 98.02, and f1score of 98.03. Figure IV presents the confusion matrix for BiLSTM model. Figure IV displays a comparison between validation and testing accuracies for all the applied models. Furthermore, The study utilized explainable AI LIME to analyze the performance of random forest and the role of each feature in the decision-making process after obtaining model predictions. The most effective role was for timestamp, flow id, and source port features. On the other hand, features such as source IP, flow ID, and Bwd Pkt Len Std have negative impact and slightly distract the model from the correct class. However, the model in each class demonstrated high confidence in the final prediction, indicating a positive impact of the chosen features. Figure IV shows the LIME explanation for each class using random forest classifier.

Table II
A SUMMARY OF THE DEEP LEARNING MODELS' HYPERPARAMETERS

HyperParameter	Value
Loss Function	Categorical Cross-Entropy
Optimizer	Adam
Epochs	7
Activation Function	Softmax
Learning Rate	0.001
Batch Size	16
Early Stopping (Patience)	3

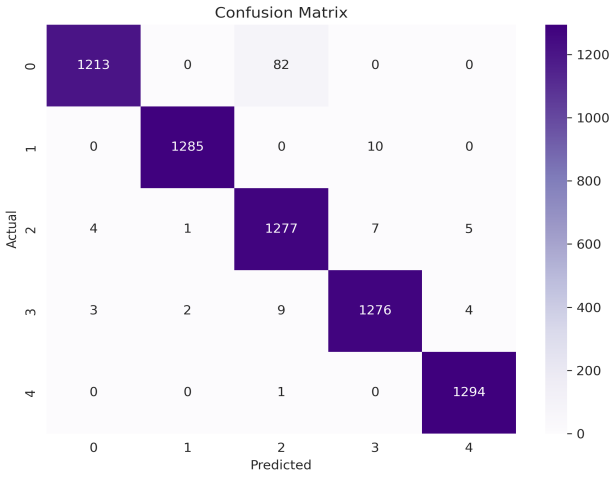

Figure 3. A visual representation of the confusion matrix for the BiLSTM model.

Table III
A SUMMARY OF THE DOS ATTACKS CLASSIFICATION RESULTS.

Model	Val Acc	Test Acc	P	R	F
DT	99.96	99.96	99.96	99.96	99.96
RF	**99.98**	**99.98**	**99.98**	**99.98**	**99.98**
MLP	97.16	97.12	97.23	97.12	97.11
BiLSTM	97.80	98.02	98.09	98.02	98.03
BiGRU	97.80	97.81	97.90	97.81	97.81
BiRNN	97.80	97.67	97.72	97.67	97.66

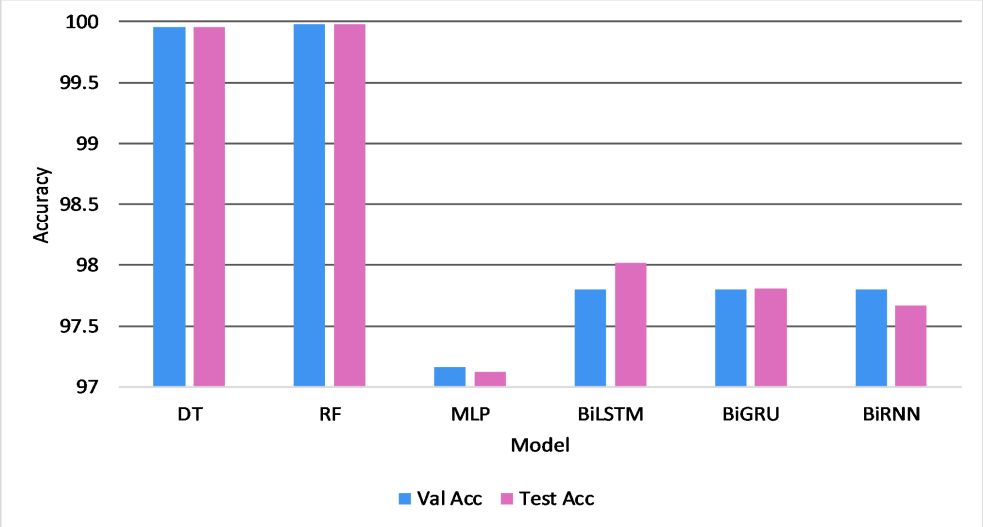

Figure 4. A visual representation of validation and testing accuracies.

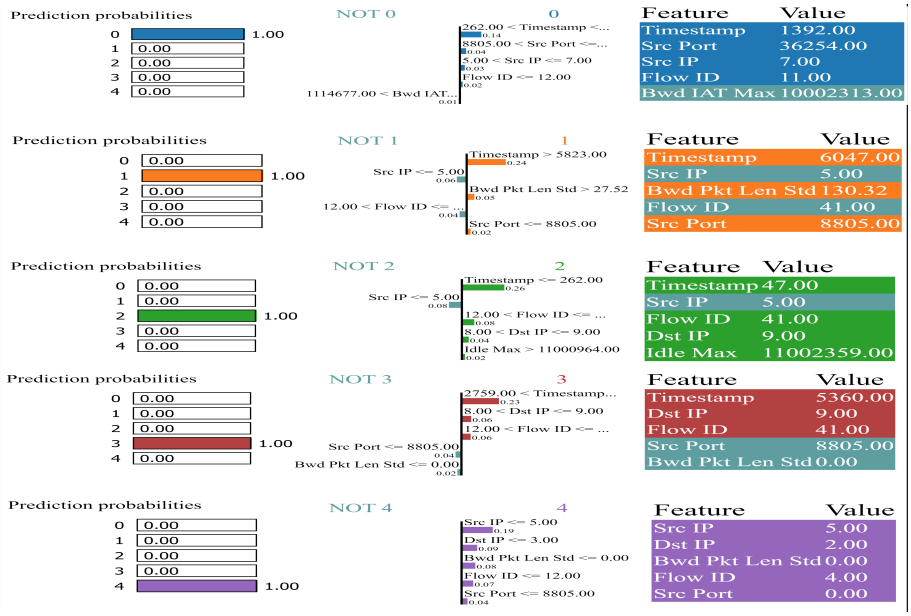

Figure 5. A visual representation of the LIME explanation for features contribution in DoS attacks detection.

V. Conclusion

This paper proposed a new approach for DoS detection in the 5G network using several deep learning and machine learning models. Also, the paper uses the explainable AI LIME to interpret model predictions, providing a comprehensive understanding of the model learning process. This approach not only interprets the decision-making process but also identifies the significant contribution of data features in detecting DoS attacks. The RF model outperformed the other models, demonstrating the significant impact of feature selection in the prediction results. The BiLSTM achieved a recall of 98.02, which reflects the model's high ability to capture malicious traffic in the network. Future work will utilize diverse datasets to detect various types of attacks in 5G networks.

References

[1] P. B. Udas, M. E. Karim, and K. S. Roy, "Spider: A shallow pca based network intrusion detection system with enhanced recurrent neural networks," *Journal of King Saud University-Computer and Information Sciences*, vol. 34, no. 10, pp. 10 246–10 272, 2022.

[2] V. E. Quincozes, S. E. Quincozes, J. F. Kazienko, S. Gama, O. Cheikhrouhou, and A. Koubaa, "A survey on iot application layer protocols, security challenges, and the role of explainable ai in iot (xaiot)," *International Journal of Information Security*, vol. 23, no. 3, pp. 1975–2002, 2024.

[3] A. Rahman, M. S. I. Khan, A. Montieri, M. J. Islam, M. R. Karim, M. Hasan, D. Kundu, M. K. Nasir, and A. Pescapè, "Blocksd-5gnet: Enhancing security of 5g network through blockchain-sdn with ml-based bandwidth prediction," *Transactions on Emerging Telecommunications Technologies*, vol. 35, no. 4, p. e4965, 2024.

[4] A. Albashayreh and M. B. Yassein, "Performance analysis of trickle timer parameters on dio messages in high-density network," in *2024 15th International Conference on Information and Communication Systems (ICICS)*, 2024, pp. 1–6.

[5] M.-D. Nguyen, V. H. La, R. Cavalli, and E. M. De Oca, "Towards improving explainability, resilience and performance of cybersecurity analysis of 5g/iot networks (work-in-progress paper)," in *2022 IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW)*. IEEE, 2022, pp. 7–10.

[6] T. Senevirathna, B. Siniarski, M. Liyanage, and S. Wang, "Deceiving post-hoc explainable ai (xai) methods in network intrusion detection," in *2024 IEEE 21st Consumer Communications & Networking Conference (CCNC)*. IEEE, 2024, pp. 107–112.

[7] I. H. Sarker, *AI-driven cybersecurity and threat intelligence: cyber automation, intelligent decision-making and explainability*. Springer Nature, 2024.

[8] I. H. Sarker, H. Janicke, A. Mohsin, A. Gill, and L. Maglaras, "Explainable ai for cybersecurity automation, intelligence and trustworthiness in digital twin: Methods, taxonomy, challenges and prospects," *ICT Express*, 2024.

[9] C. S. Kalutharage, X. Liu, C. Chrysoulas, N. Pitropakis, and P. Papadopoulos, "Explainable ai-based ddos attack identification method for iot networks," *Computers*, vol. 12, no. 2, p. 32, 2023.

[10] Y. Xiang, K. Li, and W. Zhou, "Low-rate ddos attacks detection and traceback by using new information metrics," *IEEE transactions on information forensics and security*, vol. 6, no. 2, pp. 426–437, 2011.

[11] E. Tcydenova, T. W. Kim, C. Lee, and J. H. Park, "Detection of adversarial attacks in ai-based intrusion detection systems using explainable ai," *Human-Centric Comput Inform Sci*, vol. 11, 2021.

[12] K. Sauka, G.-Y. Shin, D.-W. Kim, and M.-M. Han, "Adversarial robust and explainable network intrusion detection systems based on deep learning," *Applied Sciences*, vol. 12, no. 13, p. 6451, 2022.

[13] M. Keshk, N. Koroniotis, N. Pham, N. Moustafa, B. Turnbull, and A. Y. Zomaya, "An explainable deep learning-enabled intrusion detection framework in iot networks," *Information Sciences*, vol. 639, p. 119000, 2023.

[14] M. Siganos, P. Radoglou-Grammatikis, I. Kotsiuba, E. Markakis, I. Moscholios, S. Goudos, and P. Sarigiannidis, "Explainable ai-based intrusion detection in the internet of things," in *Proceedings of the 18th International Conference on Availability, Reliability and Security*, 2023, pp. 1–10.

[15] D. Rao and S. Mane, "Zero-shot learning approach to adaptive cybersecurity using explainable ai," *arXiv preprint arXiv:2106.14647*, 2021.

[16] S. B. Mallampati and H. Seetha, "Enhancing intrusion detection with explainable ai: A transparent approach to network security," *Cybernetics and Information Technologies*, vol. 24, no. 1, pp. 98–117, 2024.

[17] A. Alzu'bi, A. Albashayreh, A. Abuarqoub, and M. Alfawair, "Explainable ai-based ddos attacks classification using deep transfer learning," *Computers, Materials Continua*, pp. 1–10, 01 2024.

[18] Z. Masud, A. Maruf, M. Haque, S. Mrityunjoy, A. Ahad, and Z. Aung, "Ddos attack classification leveraging data balancing and hyperparameter tuning approach using ensemble machine learning with xai," 01 2024.

[19] B. Mahbooba, M. Timilsina, R. Sahal, and M. Serrano, "Explainable artificial intelligence (xai) to enhance trust management in intrusion detection systems using decision tree model," *Complexity*, vol. 2021, no. 1, p. 6634811, 2021.

[20] H. He, G. Huang, B. Zhang, and Z. Zheng, "[retracted] research on dos traffic detection model based on random forest and multilayer perceptron," *Security and Communication Networks*, vol. 2022, no. 1, p. 2076987, 2022.

[21] M. Wang, Y. Lu, and J. Qin, "A dynamic mlp-based ddos attack detection method using feature selection and feedback," *Computers & Security*, vol. 88, p. 101645, 2020.

[22] Y. Zhang, Y. Liu, X. Guo, Z. Liu, X. Zhang, and K. Liang, "A bilstm-based ddos attack detection method for edge computing," *Energies*, vol. 15, no. 21, p. 7882, 2022.

[23] Y. Song, N. Luktarhan, Z. Shi, and H. Wu, "Tga: a novel network intrusion detection method based on tcn, bigru and attention mechanism," *Electronics*, vol. 12, no. 13, p. 2849, 2023.

[24] Y. Wu, "A ddos attack detection method based on recurrent neural network," in *International Conference on Internet of Things and Machine Learning (IoTML 2023)*, vol. 12937. SPIE, 2023, pp. 29–36.

[25] O. Darwish, Y. Tashtoush, A. Bashayreh, A. Alomar, S. Alkhaza'leh, and D. Darweesh, "A survey of uncover misleading and cyberbullying on social media for public health," *Cluster computing*, vol. 26, no. 3, pp. 1709–1735, 2023.

[26] M. T. Ribeiro, S. Singh, and C. Guestrin, "" why should i trust you?" explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.