2025 Joint European Conference on Networks and Communications & 6G Summit (EuCNC/6G Summit) | 979-8-3503-9180-0/25/$31.00 ©2025 IEEE | DOI: 10.1109/EUCNC/6GSUMMIT63408.2025.11037121

Robust Intrusion Detection System with Explainable Artificial Intelligence

Betül Güvenç Paltun
Ericsson Research
Istanbul, Turkey
betul.guvenc.paltun@ericsson.com

Ramin Fuladi
Ericsson Research
Istanbul, Turkey
ramin.fuladi@ericsson.com

Rim El Malki
Ericsson Standards and Technology
Paris, France
rim.el.malki@ericsson.com

***Abstract*—Machine Learning (ML) models are widely adopted for threat detection and mitigation, but their susceptibility to adversarial inputs presents new vulnerabilities, particularly in time-sensitive environments like 6G and Open Radio Access Network (O-RAN). Existing defenses, such as adversarial training, are resource intensive and often fail in real-time scenarios. To address this, we propose a novel adversarial detection and mitigation framework that leverages eXplainable Artificial Intelligence (XAI) to provide real-time insights and automated zero-touch responses. Our method is integrated into Intrusion Detection Systems (IDS) and validated through extensive testing in the Radio Resource Control (RRC) layer of the O-RAN framework. Experimental results demonstrate improved detection accuracy and reduced response time compared to baseline approaches, showcasing the effectiveness of XAI-enhanced zero-touch security in dynamic network environments.**

***Index Terms*—Adversarial Attacks, Explainable AI, Intrusion Detection System, Open Radio Access Network.**

I. Introduction

Maintaining the confidentiality, integrity, and availability of Artificial Intelligence (AI) models and their data is critically important. Threats appear in various forms, including adversarial attacks on AI models, data manipulation, and unauthorized access to confidential data [1]. To mitigate these risks, security strategies involve deploying strong authentication processes, encrypting data both in transit and at rest, and continuously monitoring for anomalies to safeguard AI elements within network architectures. Addressing cybersecurity risks posed by adversarial AI methods remains a significant challenge. Meanwhile, security measures are increasingly reliant on AI/ML models to identify and counteract evolving, sophisticated threats. For example, IDSs are critical for analyzing network activities and detecting suspicious actions that may indicate possible attacks. Advances in ML have improved the IDS functionality, allowing them to identify anomalies more effectively. Their effectiveness is particularly notable given the large volume of data processed in 5G and beyond-mobile network environments. However, attackers continuously seek to diminish the efficiency of ML-based IDSs, thereby intensifying the impact of network attacks such as RRC signaling storms. For example, adversarial techniques can enable adversaries to bypass AI-powered IDS and gain unauthorized access to essential systems or data. To address this increasing threat, we intend to propose an agnostic robust strategy to detect and mitigate such attacks using XAI for real-time evaluation.

To further explore the vulnerability mentioned above, this paper outlines our significant contribution to developing a comprehensive framework focused on identifying and countering adversarial threats in IDSs. We present an agnostic methodology that utilizes XAI techniques to evaluate how adversarial samples impact the interpretations of machine learning models. In addition, we have developed a zero-touch detection strategy specifically aimed at improving IDS capabilities to strengthen defenses against attackers using network-related attacks and other advanced techniques to evade detection mechanisms. By proactively addressing these security gaps, we aim to enhance both the security and resilience of IDSs. Our approach enables IDS not only to detect emerging attack vectors, but also to respond immediately to potential threats even when the number of features is limited. This proactive approach ensures that security measures are effective and flexible, protecting confidential data from adversary actions. As a test setup, the O-RAN infrastructure has been chosen since it focuses heavily on the use of AI/ML to optimize the functionality and capabilities of the network [2]. O-RAN has incorporated 3rd Generation Partnership Project (3GPP) standards to enhance its interfaces and protocols, addressing the expanded attack surface that comes with virtualization, open interfaces, and multi-vendor settings. Despite these improvements, the architecture is insufficient to address the specific vulnerabilities of ML/AI capabilities within O-RAN components, such as the Non-Real-Time RAN Intelligent Controller (Non-RT RIC) and the Real-Time RAN Intelligent Controller (Near-RT RIC) [3].
Our main contributions are as follows.

This work was funded by The Scientific and Technological Research Council of Turkey, under 1515 Frontier R&D Laboratories Support Program with project (Grant number: 5169902), and in part by the ROBUST-6G Project through the Smart Networks and Services Joint Undertaking (SNS JU) under the European Union's Horizon Europe Research and Innovation Program (Grant number: 101139068).

- Developing a novel approach by incorporating an XAI feature into the ML-based detection framework for real-time assessment.
- The detection model gains deeper insights into the importance of features, leading to better selection of relevant attributes to identify adversarial attacks and reduce false positives.
- The idea of XAI feature centers around understanding the behavior of the unseen data, ensuring that its distribution aligns with the standard behavior range observed in training data. The essential part is to assess the distribution of SHAP values within the training data, rather than the training data itself, which was not previously proposed.
- The proposed XAI-integrated ML-based detection approach improves the overall effectiveness of the detection mechanism, ensuring better adaptability to evolving attack patterns.
- We introduce a straightforward but impactful mitigation technique that relies on anticipated adversarial examples.
- In addition to offering detection and mitigation capabilities, this approach introduces a zero-touch strategy specifically designed to enhance IDS capabilities, thus strengthening defenses against adversaries.

II. Background

This section provides essential background on the proposed framework. We start with an overview of adversarial AI, followed by an introduction to XAI. Lastly, we give a concise description of RRC signaling storm attacks within O-RAN, which we consider as a use case study in this paper.

A. Adversarial AI

Adversarial AI is the field of AI that focuses on investigating how adversaries might take advantage of AI systems by fooling them into providing incorrect results. These modifications, known as adversarial attacks, can lead machine learning models to produce inaccurate predictions or classifications by gradually changing the input data in a way that is often imperceptible to humans but has significant impacts on how well the model is able to make decisions. Attackers may, for example, craft inputs that evade IDS, spam filters, or malware detection algorithms, causing them to mistakenly identify malicious activity as normal or vice versa. It is possible to weaken AI systems using well-crafted adversarial examples, which require strong defenses and a thorough understanding of how these attacks operate. Adversarial attacks can be broadly divided into white-box and black-box attacks. In white-box attacks, the attacker has complete information about the target model, including its architecture, parameters, and training information. This information allows an attacker to create very effective countermeasures. Black-box attacks, on the other hand, assume that the attacker has access only to the inputs and outputs and is not informed about the inner workings of the model.

B. Explainable AI

XAI is an important subfield of AI and addresses the need for transparency in AI systems by explaining internal mechanisms and offering explanations of results [4]. This is vital to promoting trust and fairness through AI- and ML-driven decision insights. Some traditional ML algorithms, such as decision trees and linear regression, are inherently interpretable but limited in predictive power [5]. In contrast, while deep learning has achieved exceptional results in many tasks, its complex nature requires XAI techniques to improve the transparency and understanding of its intricate decision making. Explanation methods can be classified as model-specific and model-agnostic. Model-specific techniques are applied to specific models or groups of models, allowing for an understanding of decisions by investigating their underlying mechanisms, such as explaining coefficient weights in neural networks. Model-agnostic methods, on the other hand, study the relation between input and output variables without having structure of the model, allowing for generalization across multiple models. The two well-known model-independent explanation strategies in the literature are Local Interpretable Model-Agnostic Explanations (LIME) [6] and Shapley Additive Explanations (SHAP) [7]. LIME provides localized explanations by analyzing model predictions with varying input data. The procedure entails generating a new data set consisting of perturbed samples and the corresponding predictions of a black-box model. Alternatively, SHAP is based on the concept of game-theoretically optimal Shapley values as a method to understand the reasoning behind individual predictions.

C. RRC Signaling Storm Attacks

RRC signaling storms present a significant threat to cellular networks by causing excessive signaling activity that overwhelms the control plane. These storms can occur due to malicious activities, such as attacks by malware or improperly configured applications, or unintentionally from high device density and rapid re-registration attempts. The RRC protocol, particularly in 4G and 5G networks (including those adopting the O-RAN architecture), is vulnerable to such disruptions that compromise network operations [8].

Detecting RRC signaling storms in real time is essential to enable timely mitigation, especially since response and mitigation strategies differ depending on whether it is a malicious attack or a legitimate high-load scenario. In the event of a malicious attack, targeted actions must be taken to prevent resource waste and protect the network from service degradation. In contrast, during a high-load scenario, the focus is on managing capacity and ensuring that legitimate traffic is prioritized. Differentiating between these two scenarios allows operators to apply appropriate mitigation techniques, preserving network availability, ensuring continuous connectivity, and safeguarding overall network stability and QoS. This proactive approach helps avoid unnecessary resource consumption and ensures that the network performs optimally under both attack and high-load conditions [9].

III. Related Work

Various approaches have been explored in the literature to protect IDSs against such adversarial threats. For example, in [10], the authors propose adversarial training to enhance the detection of adversarial samples. In [11], researchers advocate for a deep learning-based approach specifically tailored for detecting adversaries. However, existing solutions face several challenges. First, many methods require integration during the training phase and direct application within the ML model used by the IDS. This makes the pre-trained models susceptible to analysis by attackers, who can then adapt their tactics to exploit the weaknesses of the model. Moreover, many solutions ignore the specific features of the attack vectors, missing key characteristics. The survey study [12] highlights the critical importance of explainability in the context of IDS. Although previous studies have made important contributions to the field of IDS by improving explainability, there was no comprehensive framework that combines the strengths of ML robustness against adversarial attacks, real-time applicability, and ensuring the robustness of IDS without human intervention. In our earlier investigation [13], we showed that XAI is highly effective in identifying adversaries through the comparison of explanations derived from preprocessed network traffic data with the decisions reached by ML-based IDS. However, this method demands a substantial amount of features for the analysis of adversarial entities, indicating the necessity of a more robust strategy for comprehensive applications. Bridging these gaps with a unified approach has the potential to significantly enhance the efficiency and reliability of IDS systems. To address the above challenges, this paper presents a robust IDS to detect and defend against adversarial attacks incorporating an XAI feature by running the proposed detection model in run time, even with a limited number of features, including white-box and black-box scenarios. We provide the details in the following sections.

IV. Proposed Framework

We introduce an agnostic adversarial detection approach to enhance IDS robustness by integrating the XAI feature to perform a real-time assessment. This method is applicable regardless of the IDS model or use case. It aims to strengthen IDS, improving its ability to counteract attackers who use a combination of network and adversarial attacks to bypass detection. The core strategy of the framework involves integrating XAI to identify and highlight features modified by attackers, effectively flagging manipulated network traffic for IDS intervention. Fig. 1 presents a general structure of the proposed approach for IDS. Incorporating the XAI feature into the ML model of an IDS assists in determining whether the IDS is vulnerable to adversarial attacks. This is achieved by characterizing the normal behavior pattern of the data through the distribution of SHAP importance values.

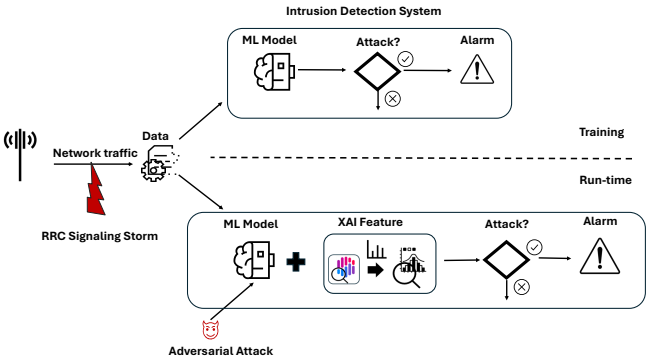

Fig. 1. General structure of the problem and proposed XAI-based adversarial attack detection framework.

A. Adversarial attacks against IDS

We introduce adversarial instances designed to deceive the IDS. These instances are generated with the best possible results to mislead the IDS by mislabeling malicious traffic as normal. Using techniques such as FGSM [14] and projected gradient descent (PGD) [15] can effectively generate adversarial examples. While FGSM is a fast, single-step method that applies a small perturbation in the direction of the gradient of the loss function, PGD is a multistep iterative extension of FGSM that applies successive perturbations and projects the perturbed inputs back onto the valid data domain. In addition to FGSM and PGD, Gaussian attack and Basic Iterative Method (BIM) [16] are other notable techniques for creating adversarial examples. BIM serves as a variation of the PGD approach, iteratively applying small perturbations while also ensuring that each modified input remains within a certain constraint. Meanwhile, the Gaussian attack introduces noise drawn from a Gaussian distribution into the input data. This method can be particularly effective in scenarios where traditional gradient-based approaches may be less successful.

B. Detection of Adversarial attacks

We propose a novel XAI-based adversarial detection framework designed to determine whether unseen data have been manipulated. The learning process is divided into the training and run-time phases.

During the training phase, SHAP feature importance values of training are gathered for each input to characterize the normal behavioral pattern of the data by examining the distribution of these SHAP importance values.

- Let $X = \{x_1, x_2, \ldots, x_n\}$ be the training data set and let $f(X)$ be the ML model trained on X. The SHAP importance values for each input x_i are indicated as $S(x_i) = \{S_1(x_i), S_2(x_i), \ldots, S_m(x_i)\}$, where m is the number of features.
- Let us assume that the distribution of SHAP values for each characteristic j follows a normal distribution:

$$S_j(x) \sim \mathcal{N}(\mu_j, \sigma_j^2)$$

where μ_j and σ_j^2 are the mean and variance computed from the training data.

During run-time, the ML model is used to evaluate unseen data. To assess the behavior of unseen data, we verify if the SHAP feature importance values for this data fall within the normal behavior distribution range of the training data.

- Let $X' = \{x'_1, x'_2, \ldots, x'_k\}$ be the test data. The SHAP values for a test sample x'_i are given by $S(x'_i) = \{S_1(x'_i), S_2(x'_i), \ldots, S_m(x'_i)\}$.
- The behavior of the test data is verified by checking whether each $S_j(x'_i)$ falls within the normal behavior range:

$$\mu_j - \lambda\sigma_j \leq S_j(x'_i) \leq \mu_j + \lambda\sigma_j$$

where λ is a threshold parameter (e.g., $\lambda = 2$ for a 95% confidence interval assuming normality).

If SHAP values of unseen data maintain the same distribution as those of the training data, the performance of the model is considered similar to its training behavior.

- If $S_j(x'_i)$ falls within this range for all j, the input x'_i is considered Normal.
- If there exists at least one feature j such that:

$$S_j(x'_i) < \mu_j - \lambda\sigma_j \quad \text{or} \quad S_j(x'_i) > \mu_j + \lambda\sigma_j$$

then x'_i is considered Attack.

V. Experimental Evaluation

The primary objective of the experimental setup is to show how our proposed approach is capable of detecting and mitigating adversarial instances that cause a significant degradation in the IDS performance which is developed for RRC signaling storm attacks in real-time. We leverage XAI to detect and prevent degradation of IDS performance. The proposed framework seeks to detect potential adversarial attacks by observing significant deviations in the distribution of SHAP importance values for each input in real-time, and to establish a straightforward yet effective zero-touch mitigation strategy.

A. O-RAN Setup

The OpenAirInterface (OAI) setup was utilized to implement the RRC signaling attack and extract relevant features for detection. The key components of the setup include OAI-UE (User Equipment), OAI-gNB (a software-based 5G base station), FlexRIC (an open source Near-RT-RIC that extracts RRC-related features from the OAI-gNB and forwards them to an xApp), an xApp (which processes the extracted features to perform detection tasks), and OAI-core. The OAI-UE has been modified to perform a malicious RRC attack by intentionally exploiting vulnerabilities in the RRC signaling process, enabling the generation of abnormal signaling patterns [17]. The details of data collection are given in Fig. 2.

FlexRIC acts as a real-time RAN Intelligent Controller (RIC) platform to enable communication between the xApp and the gNB using the E2 interface. The gNB is configured with an E2 agent that collects RAN metrics and handles control messages. The xApp subscribes to specific features

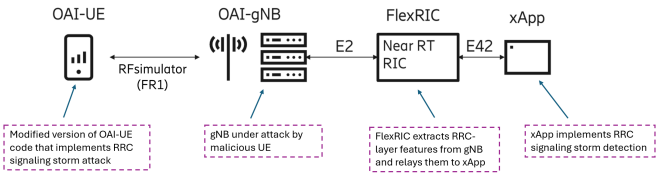

Fig. 2. Steps of xApp feature retrieval from a gNB within OAI setup using FlexRIC.

or metrics (e.g., RRC measurements, CQI, traffic statistics) through FlexRIC. The E2 agent collects the requested features and sends them to FlexRIC via the E2 interface. FlexRIC processes the data received and delivers it to the xApp using its internal communication framework.

B. Experiment Settings

1) Data: In the experiments, there is 1 malicious User Equipment (UE) and no benign UE. The resource reservation delay by the gNB is about 2.7 seconds, which allocates resources to UE. The Open Air Interface gNB can handle up to 16 resources. The attack rate is around 132 messages per second, exceeding the 90-message-per-second threshold, which overloads the gNB. Consequently, the gNB becomes blocked in approximately 157 milliseconds.For the detection of RRC signaling storms, five key characteristics are collected. Detailed descriptions of these features are given in Fig. 3. A decrease in Msg5s, seen in both attacks and high-load situations, causes $R1$ to drop, which helps to confirm the presence of an anomaly. In an attack, the malicious UE does not respond to Msg4s with Msg5s, while in a high-load scenario, not all UEs will receive Msg4, leading to fewer Msg5s in return. These features, when analyzed together, help create a detection system that identifies signaling storms and distinguishes between attacks and high-load scenarios.

Feature	Value	Definition	Purpose	Trend
Msg3	# of Msg3s	Number of incoming Msg3s at the gNB	Detecting abnormal activity	Decreasing for both attack and high-load
R1	$\frac{\#\text{ of Msg5s}}{\#\text{ of Msg3s}}$	Ratio of completed RRC setup procedures, considering all sent Msg3s	Helper feature (optional) to detect abnormal activity. It can't be used to distinguish b/w high-load and attack	Decreasing for both attack and high-load
R2	$\frac{\#\text{ of Msg5s}}{\#\text{ of Msg4s}}$	Number of Msg5s the gNB gets as response to sent Msg4s. Ideally should be = 1	Differentiator b/w attack and high-load	Decreasing for attack. Constant at value equal to 1 for high-load

Fig. 3. The key features used for the detection of RRC signaling storms.

2) Intrusion Detection System: In this paper, IDS is developed using auto-encoders, unsupervised learning neural networks, which are often used to identify anomalies. An auto-encoder comprises an encoder that compresses the input $\mathbf{x} \in \mathbb{R}^n$ to a latent representation $\mathbf{z} \in \mathbb{R}^m$ (with $m < n$) and a decoder that aims to reconstruct the original data. The key benefit lies in its training on normal data, enabling the capture of typical patterns. During testing, the auto-encoder tries to reconstruct the input based on previous learning. If anomalies exist in test data, the reconstruction error is characterized as

$$E(\mathbf{x}) = \|\mathbf{x} - \hat{\mathbf{x}}\|$$

(where $\hat{\mathbf{x}}$ is the reconstructed output), tends to be significantly higher, indicating the existence of unexpected patterns, which can be interpreted as potential anomalies. In this research, the auto-encoder model is trained with a normal traffic data set to ensure it outputs features similar to regular traffic. The similarity between input and output vectors is assessed using vector similarity metrics such as Euclidean distance:

$$d(\mathbf{x}, \hat{\mathbf{x}}) = \sqrt{\sum_{i=1}^{n} (x_i - \hat{x}_i)^2}$$

If attack traffic features are introduced into the model, we expect to obtain larger distance values compared to those for normal traffic inputs.

3) Scenarios: There are three different experimental setups.

In the initial scenario, we assess the performance of the IDS in two scenarios: one with only an RRC signaling storm attack to disrupt network traffic and another that combined this attack with an adversarial attack on the IDS. Various techniques are used to assess the effectiveness of adversarial attacks.

- Initially, we simulate an RRC signaling storm attack characterized by excessive signaling that overloads the control plane. The IDS is then evaluated for its detection performance.
- To establish a baseline, IDS performance is evaluated under normal conditions without adversarial attacks. An auto-encoder, described in Section V-B2, is employed for this purpose.
- We present adversarial examples designed to mislead IDS, followed by performance evaluation to determine the level of degradation.

In the second scenario, our suggested adversarial attack detection strategy, integrating an XAI feature, helps determine whether the unseen data provided to the IDS fall within the typical behavior distribution of the training data.

- During training time, the importance values of each input are calculated using the SHAP method.
- Kernel density estimation is used to create a smooth estimate of the distribution from importance values of training data and used to define the normal behavior pattern of the IDS data.
- During run-time, the autoencoder is run to evaluate unseen data. To determine whether a test input is an anomaly based on a distance, we compute a threshold (defined as a Z-Score) based on our known normal behavior (which could be derived from the training data) and then check if the distance of the unseen input exceeds this threshold.

Finally, we apply a simple yet effective mitigation approach that observes the predicted attack instances.

- If the unseen data follow the same distribution as the training data (i.e. normal), it is considered ”Normal”.
- If an input falls outside the usual distribution range and is classified as an outlier, it is considered manipulated, and we adjust its label accordingly.

C. Performance Results

Fig. 4 shows the accuracy of IDS under different adversarial attack methods. A higher epsilon value indicates stronger perturbations in the X-axis. Y-axis represents the accuracy of the IDS in correctly classifying inputs as either normal or malicious. No attack shows the baseline accuracy of the IDS when there are no adversarial attacks. The accuracy remains constant and high accuracy across all epsilon values indicates the model performs well without adversarial interference. As the epsilon increases, the accuracy drops for the rest of the models, showing that the performance of IDS degrades with stronger adversarial perturbations. The significant drop in accuracy with increasing epsilon suggests that FGSM is highly effective at exploiting vulnerabilities, demonstrating the need for stronger defenses against gradient-based attacks. Although accuracy decreases, the model maintains better performance compared to FGSM, indicating some resilience. This suggests that iterative attacks like PGD are less effective but still significant, highlighting the partial robustness of the model. With slightly better accuracy than PGD, the model shows better handling of iterative attacks, suggesting that existing defenses might be more effective against methods like BIM. The drop in accuracy under Gaussian attack shows high sensitivity to random noise, highlighting the necessity for robust noise features and effective preprocessing.

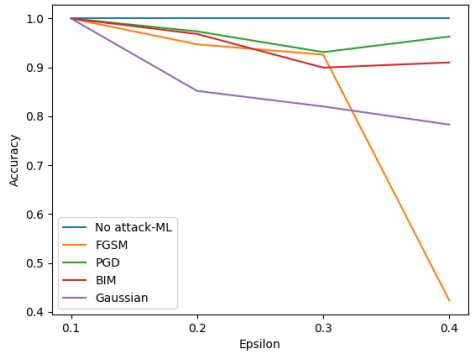

Fig. 4. The accuracy of IDS under different adversarial attack methods.

In the second experiment, to evaluate our proposed approach, we also implemented it with individual feature importance methods such as LIME and permutation importance. Table I shows the comparison of the performance metrics of four approaches that reflect different detection scenarios under the BIM attack. The AE-BIM attack does not apply any adversarial detection method, with a precision of 0.7619 and indicating that it does not detect any false positive, highlighting a challenge in recognizing adversarial examples. The permutation approach shows an improvement with both true negatives

and some true positives, indicating better performance in recognizing adversarial perturbations.The LIME method further improves upon Permutation, with a higher accuracy of 0.8307 and a better balance between precision and recall. Finally, the SHAP method demonstrates the best performance across all metrics, achieving a significant accuracy of 0.9259 and an F1-score of 0.8654. In particular, it correctly identifies all positive cases, while maintaining a high precision of 0.7627. This suggests that SHAP is the most effective, providing both high accuracy and robust anomaly detection capabilities.

TABLE I
DETECTION PERFORMANCE OF FOUR DIFFERENT METHODS.

Method	Accuracy	Precision	Recall	F1-Score
AE-BIM Attack	0.7619	0.0000	0.0000	0.0000
Permutation	0.8095	0.6364	0.4667	0.5385
LIME	0.8307	0.6383	0.6667	0.6522
SHAP	0.9259	0.7627	1.0000	0.8654

In the mitigation scenario, Fig. 5 illustrates the mitigation performance of the proposed approach in various scenarios. The permutation importance approach improves accuracy, suggesting detrimental effects from the attack or transformation. Although LIME helps explain predictions, its effectiveness is comparable to that of the permutation method, highlighting its limitations in mitigating attack impacts. In particular, SHAP achieves the highest level of accuracy among the individual feature importance methods, reflecting its superior capability to manage adversarial influences and improve the robustness of the model. These results are consistent with those obtained in earlier experiments.

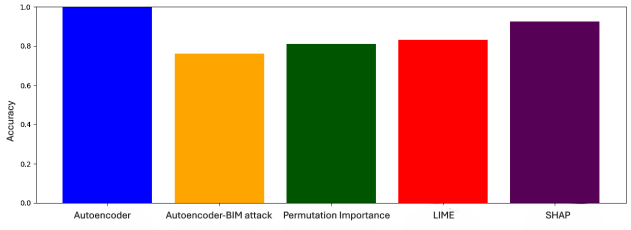

Fig. 5. Method comparison with XAI-based mitigation

VI. CONCLUSION

In conclusion, the proposed system integrates the XAI feature into the adversarial detection method, identifies unusual behaviors indicative of adversarial activities accurately and efficiently. This method demonstrated its effectiveness by improving the performance of IDS in an O-RAN environment, representing notable progress in strengthening cybersecurity measures.The suggested method introduces zero-touch functionality and, when implemented on IDS, provides prompt responses to new threats, significantly reducing potential risks to network infrastructures. However, its reliance on a limited set of features and the focus on a single threat type may limit its effectiveness in more complex real-world scenarios. Future work will focus on improving detection capabilities by incorporating richer features, ensuring scalability in large-scale O-RAN environments, and extending threat coverage to include a wider range of attacks for greater applicability. Furthermore, with the ongoing evolution of O-RAN, the collaboration of XAI and IDS will be critical to adapt to new vulnerabilities and refine the efficiency of XAI-enabled cybersecurity measures, ensuring that network defenses remain strong within the constantly changing landscape of cyber threats.

REFERENCES

[1] A. Chakraborty, M. Alam, V. Dey, A. Chattopadhyay, and D. Mukhopadhyay, "Adversarial attacks and defences: A survey," *arXiv preprint arXiv:1810.00069*, 2018.
[2] S. Niknam, A. Roy, H. S. Dhillon, S. Singh, R. Banerji, J. H. Reed, N. Saxena, and S. Yoon, "Intelligent o-ran for beyond 5g and 6g wireless networks," in *2022 IEEE Globecom Workshops (GC Wkshps)*, pp. 215–220, IEEE, 2022.
[3] V. P. Illiano and E. C. Lupu, "Detecting malicious data injections in wireless sensor networks: A survey," *ACM Computing Surveys (CSUR)*, vol. 48, no. 2, pp. 1–33, 2015.
[4] D. Gunning and D. Aha, "Darpa's explainable artificial intelligence (xai) program," *AI magazine*, vol. 40, no. 2, pp. 44–58, 2019.
[5] D. V. Carvalho, E. M. Pereira, and J. S. Cardoso, "Machine learning interpretability: A survey on methods and metrics," *Electronics*, vol. 8, no. 8, p. 832, 2019.
[6] 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*, pp. 1135–1144, 2016.
[7] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.
[8] A. Tabiban, H. A. Alameddine, M. A. Salahuddin, and R. Boutaba, "Signaling storm in o-ran: Challenges and research opportunities," *IEEE Communications Magazine*, 2023.
[9] A. Kapoor, A. Ganiyu, and V. K. Shah, "Signaling storming attack detection and mitigation in open radio access networks," *Journal of Student-Scientists' Research*, vol. 6, 2024.
[10] A. Abusnaina, A. Khormali, D. Nyang, M. Yuksel, and A. Mohaisen, "Examining the robustness of learning-based ddos detection in software defined networks," in *2019 IEEE Conference on Dependable and Secure Computing (DSC)*, pp. 1–8, IEEE, 2019.
[11] B. Nugraha, N. Kulkarni, and A. Gopikrishnan, "Detecting adversarial ddos attacks in software-defined networking using deep learning techniques and adversarial training," in *2021 IEEE International Conference on Cyber Security and Resilience (CSR)*, pp. 448–454, IEEE, 2021.
[12] 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. 112392–112415, 2022.
[13] B. G. Paltun and R. Fuladi, "Introspective intrusion detection system through explainable ai," in *2024 8th Cyber Security in Networking Conference (CSNet)*, pp. 28–32, IEEE, 2024.
[14] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.
[15] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *arXiv preprint arXiv:1706.06083*, 2017.
[16] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," in *Artificial intelligence safety and security*, pp. 99–112, Chapman and Hall/CRC, 2018.
[17] F. Kaltenberger, A. P. Silva, A. Gosain, L. Wang, and T.-T. Nguyen, "Openairinterface: Democratizing innovation in the 5g era," *Computer Networks*, vol. 176, p. 107284, 2020.