Explainable Adversarial Mitigation Framework for Zero-Trust Cyber Warfare

Ebuka Chinaechetam Nkoro *, Cosmas Ifeanyi Nwakanma†, Jae-Min Lee *, and Dong-Seong Kim *
*IT Convergence Engineering, † ICT Convergence Research Center, Kumoh National Institute of Technology, Korea.

***Abstract*—The Zero-trust security architecture is a paradigm shift toward resilient cyber warfare. Although Intrusion Detection Systems (IDS) have been widely adopted within military operations to detect malicious traffic and ensure instant remediation against attacks, this paper proposed an explainable adversarial mitigation approach specifically designed for zero-trust cyber warfare scenarios. It aims to provide a transparent and robust defense mechanism against adversarial attacks, enabling effective protection and accountability for increased resilience against attacks. The simulation results show the balance of security and trust within the proposed parameter protection model achieving a high F1-score of 94%, a least test loss of 0.264, and an adequate detection time of 0.34s during the prediction of attack types.**

***Index Terms*—Adversarial Machine Learning, Zero-trust Security, IDS, XAI**

I. INTRODUCTION

Military operations are witnessing a growing demand for battlefield cybersecurity leveraging artificial intelligence (AI). Although the physical war has shone more severity on lives and military operations, cyberwar invasions, as witnessed in Russia vs. Ukraine, have also resulted in numerous cyberwar casualties, affecting the confidentiality, integrity, and availability of critical infrastructure for individuals and corporations [1]. In order to mitigate cyber threats [2], intrusion detection systems (IDSs) have enabled robust defense against various cyberattacks.

This study is motivated by two security mitigation concerns. Firstly, the tolerant access IDS in which bad actors have compromised with diverse perturbation techniques, especially in deep learning (DL) based frameworks. With these perturbations, attackers maliciously learn machine learning (ML) models, forcing them to misclassify attacks. However, the *Zero-trust* security principle-*"never trust"*, *"always verify"*, employs a layered cyber defense strategy that assumes breach to scrutinize untrusted devices, users, and especially IDSs before adoption [3]. Secondly, in an era where current AI development seeks trustworthiness and explainability, various black-box IDS algorithms employed for cyber deterrence have lacked fair, trustworthy, and interpretable human-machine collaboration for increased cyber resilience [4].

In order to address the increasing level of adversarial cyberattacks and lack of transparent IDS models in the military domain, this paper proposes an explainable adversarial mitigation method, using the *parameter protection* (PP) technique for efficient detection of anomalous from benign attacks while offering a layer of defense against adversarial perturbation attacks. It serves as a first layer of zero trust defense for model obfuscation and preventing the actual model gradient from exposure and manipulation by malicious actors. The specific PP approach employs a custom gradient (*SGDOptimizerWithMask*) to ensure robust mitigation of the model gradient against perturbation attacks. Furthermore, this paper adopts the *Explainable AI (XAI)* technique as another pillar of Zero-trust security, which offers security experts and users a global explanation that satisfies the trust and reliability issues of complex black box model predictions of anomalous from benign attacks. Specifically, this study evaluated two explainability methods SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) regarding effectiveness and model impact decisions.

This study proposed two layers of zero-trust security for detecting network anomalies. Firstly the trained DL model is shielded using the PP technique, thereby inhibiting the attacker from the knowledge of the model. According to [5], attackers obtain model parameters to perpetuate malicious adversarial attacks. Therefore, adopting the PP mechanism prevents excessive model gradient exposure to gain control over the original model. Secondly, the SHAP and LIME explainability methods, on the other hand, improve model interpretability, trustworthiness, and visual explainability of its prediction to improve trust and concordance with security experts.

Specifically, this study focuses on the following:

- A zero-trust IDS framework with two rings of layered defense leveraging parameter protection and model explainability to increase cyber defense against network intrusions.
- An extensive dual experimentation to show the model's performance with and without an adversarial mitigation technique using the parameter protection method.
- Presentation of the SHAP and LIME explainability method comparison regarding efficiency and impact decisions for model transparency.

The study arrangement is thus: Following the introduction in Section I, Section II reviews brief security concepts of zero trust for network anomaly detection, adversarial mitigation, and XAI for trustworthy IDS models. Section III discusses the proposed architecture using parameter protection approach with model explainability. Section IV discusses the experimentation environment and results. The study concludes in Section V.

II. BACKGROUND AND RELATED WORKS

This section discusses brief security concepts of zero trust for network anomaly detection, adversarial mitigation, and XAI for trustworthy IDS models.

A. Zero-trust for Network Anomaly Detection

Cyber-warfare activities wreak heavy havoc on national security, public operations, and critical infrastructure connected to the internet (such as transportation [6], government operations, energy plants, and financial institutions). The novel Zero-trust security Architecture (ZTA) enforces the *"never trust"*, *"always verify"* security principle, to combat early detection and classification of malicious from benign network behavior [7]. Various organizational policies and human and ML-based models integrate to form a robust ZTA against cyber threats. Adopting intelligent AI-based IDS models in cybersecurity is a deterrent technique to interrogate network anomalies and reduce surface attacks on critical infrastructure. However, ZTA security principles reveal the compromise of intelligence of IDS models by bad actors to misclassify attacks, leading to high false-positive rates. This concept is known as Adversarial Perturbation [5].

B. Adversarial Perturbation VS Adversarial Mitigation

A scenario where an attacker obtains a model's full knowledge is known as a white box, while an attacker who lacks model knowledge is a black box [8]. Adversarial perturbation attacks aim at forcing perturbation input within a model to misclassify a prediction. On the other hand, adversarial mitigation within the concept of intrusion detection uses a counterattack mechanism to shield IDS models from adversarial perturbations and attacks. The essence of adversarial mitigation in the light of ZTA security aims at counteracting and neutralizing adversarial attacks while adding layers of security that serve as a barrier to attackers [9]. In this way, the model is robust against degraded security issues, ranging from false-positive predictions to undetected vulnerabilities. Some adversarial mitigation defense mechanisms include; model optimization, adversarial training, adversarial detection, and the parameter protection technique. This study focuses on the parameter protection technique for securing model gradients against adversarial perturbations.

Authors in [8] proposed a deep neural network black-box adversarial attack for a binary network intrusion classification of Tor-nonTor network traffic and attaining a 96% and 93.54% respectively on original predictions. After 2644 adversarial samples of Tor traffic were applied, the overall accuracy dropped from 96% to 77% respectively. Their experiment using the UNB-CIC Tor network traffic dataset showed that successful adversarial examples introduced to a DL model could cause severe risks to network infrastructure. Concerning adversarial training, researchers in [10] proposed an adversarial training framework - GADoT against DDoS attacks, by training a Generative Adversarial Network (GAN) to defend against adversarial attacks. This approach is computationally expensive and time-consuming.

An ensemble mechanism (Def-IDS) involving the retraining and smoothing of an adversarial training process proposed by [11] for defense against adversarial attacks. Their works focused on improving the robustness and computation costs of the neural network. The authors propose a multi-class generative adversarial network (MGAN) with a second multi-source adversarial retraining for increased robustness. The distinction between previous studies and our work is that a zero-trust defense strategy combining adversarial threat mitigation with black box model explainability (XAI) is essential for layered defense against network intrusions; hence, the focus of this study.

C. XAI for Model Interpretability

The motivation for XAI adoption in cybersecurity is to interpret, explain, and evaluate the confidence level, usability, reliability, and fairness of the decisions of the ML model [12]. Since users need help understanding most black-box ML models, XAI aims to clarify, interpret, and justify intrinsic deep-learning model predictions. Two categories of explainability methods, *Post-hoc and Ad-hoc*, are used for XAI intrusion detection models. The ad-hoc explainability method provides model explanations during its decision process: rule-based systems, decision trees, rule extraction, and prototype-based explanation. Post-hoc methods offer explainability information after model prediction in terms of intrinsic explanations, such as feature contributions to the model output. Commonly used post-hoc explainability methods in IDS model explainability are the SHAP and LIME. Current research into various cyber-attacks, such as phishing attacks, botnets, and fraud, is gaining better insights, proper visualizations, and deeper forensics into the nature of these attacks and significant features for model training to perform efficient remediation.

A forthright zero-trust cyber-awareness assumes that ML models can be compromised and become a source of cyberattacks. This awareness has led to including the human-in-the-loop interpretation and explanation of IDS models to ensure that bias predictions and cyberattacks are not further propagated [13]. Alongside building frameworks with high accuracy, XAI methods ensure that IDS models are explainable, interpretable, and justifiable to users, security teams, and senior management for increased defense against cyber threats.

III. PROPOSED APPROACH

This section covers a brief overview of Zero-trust for cyber warfare and a systematic process of designing a white-box adversarial deep learning attack, parameter protection, and model explainability predictions to classify normal from anomalous network traffic attack types.

A. ZTA for Cyber-warfare

For optimal resilience against cyberattacks and successful military missions, the US Department of Defense (DoD) adopts the National Institute of Standards and Technology (NIST) Zero-trust framework [8] for integrated threat intelligence and remediation. The ZTA assumes no permission of

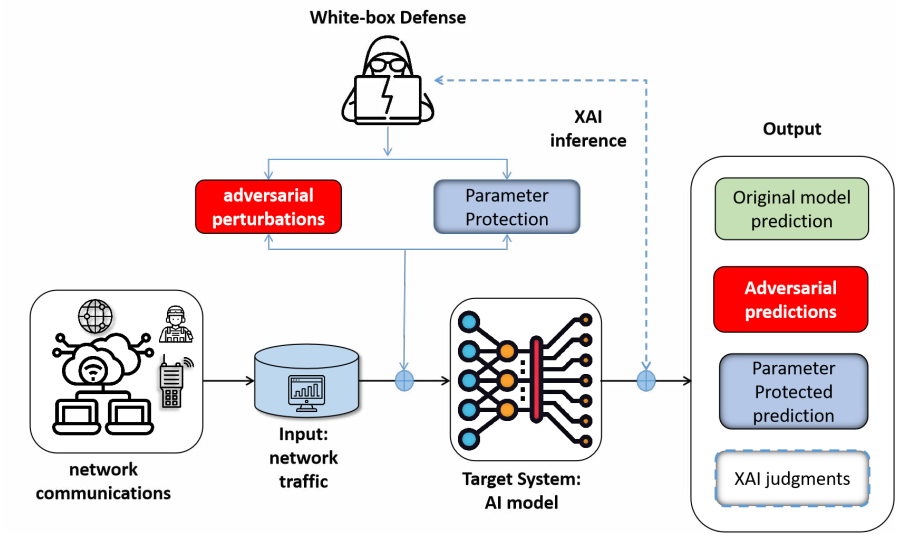

Fig. 1. Zero-trust Adversarial mitigation framework with XAI judgment

implicit trust in and outside computing resources. Under the trust analytics and orchestration capabilities of the US defense ZTA, [8] machine learning analytics, real-time network traffic monitoring, and orchestration capabilities are employed to enforce the security of data and enterprises against cyber threats. Also, evaluating the confidence levels of ML models, devices, users, and resources ensures support for mission requirements. Our proposed defense approach is designed with continuous security vetting of ML models to ensure they meet trustworthiness against cyber threats.

B. IDS Classification Model

This work adopts a feed-forward neural network architecture, which consists of multiple dense layers with a rectified linear unit (ReLU) activation functions (σ) and dropout regulation which prevents overfitting. Following the input data dimensionality, the dense layer comprises 30 neurons, followed by a dropout layer that sets 50% of the input units to 0 (to avoid overfitting). The next dense layer has 15 neurons, followed by the dropout layer with an equal configuration. The Adam optimizer is employed for training, which minimizes the loss function during training while optimizing the weights and biases of the model.

Equation 1 represents the model:

$$\hat{y} = \sigma(w\frac{T}{l}.X_i + b_l). \tag{1}$$

Where a data sample X_i is passed to the l^{th} layer given as $X_i \in R^{1\times d}$. d signifies the number of features for the model prediction.

C. Threat Model

The proposed attack in this paper assumes a white-box threat model, in order to understand and manipulate the model accordingly as in the same case where the attacker gains full knowledge of the model architecture, training data, hyperparameters, and neural network layers. Although white-box models are less challenging than black-box models, the model's performance, when altered with perturbed samples and manipulation, is highly severe and evaluated. The perturbation strength crafted using the fast gradient sign method calculates the gradient of the loss function for the input. Then it applies the *epsilon* scaling to generate adversarial samples as shown in equation 2:

$$adv_x = x + \epsilon.sign(\Delta_x J(\Theta, x, y)), \tag{2}$$

where x is the original sample, ϵ is a very small number, Δ is the gradient function, J is the loss function, Θ is the model weights, and y is the true label. As the *epsilon* ϵ, increases, the model is more likely to get *fooled*.

Algorithm 1 Parameter Protection Technique with Gradient Masking

Input: Neural network, *SDGOptimizerWithMask, maskstrength*
Output: Masked gradients
// Gradient Masking
Set the masking noise strength (*maskstrength*)
for each value of *maskstrength* in (0.01, 0.05, 0.1, 0.2, 0.5) **do**
  Apply *SDGOptimizerWithMask* with the given *maskstrength*
  // Evaluation
  Test the neural network with perturbed inputs
  // Security Assessment
  Evaluate the effectiveness of gradient masking in hiding internal weights
end for

D. Parameter Protection

The parameter protection technique employed in this experiment is a deterrent cyber defense mechanism that aims at hiding the internal weights of the neural networks from the outside world in the form of gradient masking using the (*SDGOptimizerWithMask*). External intruders/ hackers cannot obtain a significant gradient because the internal weights of the neural networks are less exposed. The strength of the masking noise controlled by the *(maskstrength)*, is tested against different ranges of values (0.01, 0.05, 0.1, 0.2, 0.5) to evaluate a moderate level of security (shielding) for gradients and at different perturbation levels. Algorithm 1 summarizes the parameter protection technique using gradient masking.

E. Model Explainability

The SHAP explainer provides the marginal value of contributions made by a feature or subset of features within a model prediction. While the LIME explainer generates local surrogate models to approximate the decision-making process of a complex model, providing interpretable explanations for individual predictions by highlighting important features. Equation 3 below evaluates the original prediction model using the following metrics as shown:

- Decision Impact Ratio (DIR): DIR refers to the rate of change in decisions due to the omissions of critical features in the interpretation method.

$$DIR = \sum_{i}^{N} \frac{1D(x_i) \neq D(x_i - c_i)}{N}. \tag{3}$$

  Where x_i denotes the i^{th} original sample, and c_i denotes the critical area marked by the model for the i^{th} sample.
- Confidence impact ratio (CIR): CIR signifies the percentage decline in confidence due to the omissions of a critical feature in the interpretation method as in equation 4:

$$CIR = \sum_{i}^{N} \frac{max(C(x_i) - C(x_i - c_i), 0)}{N}. \tag{4}$$

An evaluation of the explainability methods helps to obtain a subjective assessment of the security expert's trust and assessments of the model decisions.

IV. EXPERIMENTATION AND EVALUATION

This paper carried out systematic experiments to evaluate the original model, adversarial prediction, parameter protection predictions, and the XAI model judgments. The neural network employed in this experiment is for binary classification of anomalous from benign network attack types, and showing how white-box adversarial examples can fool ML model decisions.

A. Dataset Selection, Preprocessing, and Simulation Setup

The Network Intrusion Detection dataset is publicly available on Kaggle [14], consists of a wide variety of intrusions in a military network environment, and is used in this paper. The network environment is simulated following a typical US Air Force LAN and flooded with multiple attacks. 41 network packet flow features were obtained, with two classes: normal and anomalous (normal 13449, anomaly 11743). The dataset is then split into training and testing sets using the train-test-split Keras and Scikit-learn modules, with data samples split 70% for training and 30% for testing while keeping a random state for reproducibility. The choice of this dataset for this experiment is due to its relevance and related representation of the nature of cyber attacks reflecting military-crafted cyber-intrusions.

Within the data preprocessing stage, feature selection is an essential technique that aids model performance. The Pearson coefficient correlation (PCC) algorithm calculates the congruence between network traffic features, thereby removing the uncorrelated features within the set variance threshold of K-highest scores (0 to 1). 0 denotes no correlation and 1 is a positive linear correlation. The correlation threshold of this experiment is set to 0.8, using the Scikit learn Min-Max Scaling function, leaving only 30 relevant and correlated features. The simulation for this study was in a Python environment with the Tensorflow 2.9.0 library on a Windows 10 OS with the configuration of Intel(R) Core(TM) i3-7100 CPU @ 3.90GHz, 8GB RAM, GPU Tesla K80.

B. Model Performance and Evaluation Metrics

The performance of the ML models is evaluated adequately to the degree of correctness and rationality of the intrinsic behavior of the model. The evaluation metrics within the experiment include the F1-Score, accuracy, precision, and recall, test loss value, and optimal timely prediction. The F1-Score is a better metric since it provides a balance between precision and recall. F1-Score is denoted as $F1 - Score = \frac{2 \times precision \times recall}{precision + recall}$. While accuracy is the simple mean of model correctness obtained from the difference in predictions from the labeled ground truth data.

The results shown in Table I summarizes the prediction performance of the different model output to the test dataset in predicting anomalous from benign military network attack types. As shown in Table I, the *original prediction* attains an accuracy of 94% with a minimal test loss of 0.124.

The model's *adversarial prediction with Epsilon strength* yields a decreasing test loss value as the perturbation strength increases. With an excessive perturbation strength of 500, (for robustness evaluation) the model accuracy forcefully decreases to 50% with a loss of 1095, thus posing a security risk to any network.

The parameter protection model using an increasing range of *maskstrength* reveals an increased accuracy, reduced loss value down the table, and added security feature that protects the model's accurate gradients from bad actors. The *maskstrenght* of 0.2 yields the best parameter protection results with fair accuracy, F1-Score, and recall of 94%. In a real-world scenario, parameter protection would provide an added layer of security with a balance of optimal model performance.

Concerning the timely prediction of attacks, the time required for the model ordinary prediction is 0.31s without any

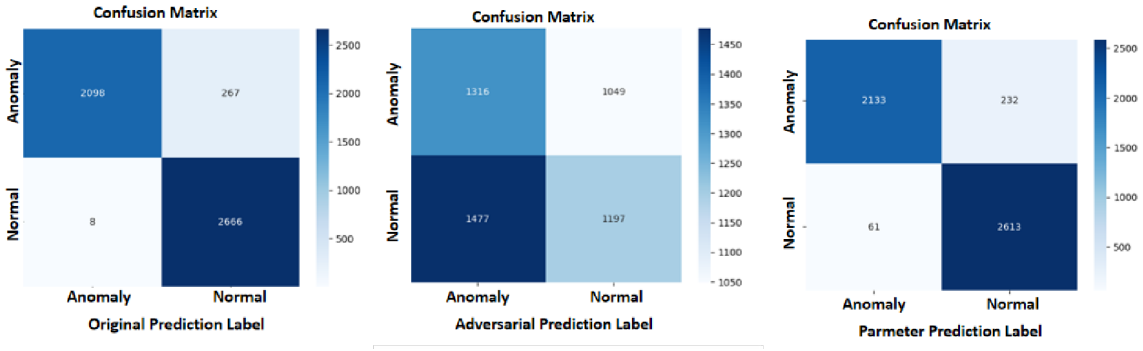

Fig. 2. Confusion Matrix showing the original vs adversarial (epsilon = 500) and parameter protection model (maskstrength = 0.2) results

perturbations. During the perturbations, the average prediction time is 0.298s. On the other hand, the trade-off for security and model performance consumes more time within the parameter protection technique, yielding an average time of 0.34s.

The confusion matrix in Fig. 2 shows the overall rate of true positive (TP), true negative (TN), false positive (FP), and false-negative (FN) for the binary classifications of the three models (original prediction, adversarial prediction, and parameter protection) experimented. Furthermore, the original model yields an increased performance without any perturbations or shielding. The FNs and FPs are the lowest in this model. (FP=267, FN=8). The adversarial confusion matrix shows a high rate of FPs and FNs as the epsilon strength increases (FP=1049, FN=1477). Lastly, the parameter-protected confusion matrix achieves fairness of (FP=232, FN=61) results considering the noise from the gradient mask, adding the essential security feature shielding the gradient of the neural networks from bad actors.

C. Model Explainability Performance

Firstly, we leverage the noise from the adversarial and parameter protection neural networks, to prevent anomalous explanation of the DNN model. The SHAP and LIME visual explainer could not handle the additional complexities for interpretation. The noise is an obscuring tool and significantly impacts the interpretability of the SHAP and LIME explainer within our experiments. This issue remains an open area of concern in the field of XAI.

However, the original model explainability with an F1-Score of 94% of explanations is provided in Table II. As explained earlier in the DIR and CIR equations 3 and 4, which can equally provide military cybersecurity teams with a proper judgment of any model decisions before adoption.

In the quantitative assessment of SHAP and LIME, the SHAP achieves a better decision and confidence impact than the LIME explainer and wins as a better performer within our experiment.

The MEAN Vote improves model inference and the Cybersecurity Experts' Trust by aggregating the mean DIR of the explainability methods with the CIR of the same explainability methods (SHAP and LIME). If the mean DIR value is greater than or equal to the mean CIR value, the mean vote will be 1. Otherwise, it will be 0. The mean vote can improve decision-making for military security experts based on this consensus and provide real-time judgments and remediation to attacks, as required by the Zero-trust defense architecture.

TABLE I
SUMMARY OF MODEL PERFORMANCE AND DETECTION

Model	**Accuracy (%)**	**F1-Score (%)**	**Precision (%)**	**Loss**	**time (s)**
Original Prediction	*94*	*94*	*95*	*0.124*	*0.31*
Adversarial Perturbation with ϵ					
0.01	94	94	94	0.128	0.33
0.05	94	94	94	0.128	0.29
0.1	94	94	94	0.129	0.29
0.5	94	94	94	0.133	0.29
500	*51*	*51*	*51*	*1095*	*0.29*
PP with ***maskstrength***					
0.01	90	90	90	0.2642	0.34
0.05	92	92	92	0.2156	0.33
0.1	92	92	92	0.2171	0.33
0.2	*94*	*94*	*94*	*0.1983*	*0.37*
0.5	93	93	94	0.203	0.33

TABLE II
EXPLAINABILITY CHECKLIST BASED ON THE ORIGINAL PREDICTION WITH THE F1 SCORE OF 94%

Explainability Method	Decision Impact Ratio (DIR)	Confidence Impact Ratio (CIR)
SHAP	0.8649	0.1239
LIME	0.7680	0.2776
Cybersecurity Experts' Trust MEAN Vote	0.8165	0.20

The SHAP visual plot is shown in Fig. 3, where the various network traffic features are ranked based on their average impact on the model, thus providing various insights like model debugging, feature selection, feature engineering, more precise explanations, and the particular feature importance for model prediction. For instance, if the *dsthostsrvcount* randomized to have no prediction power in the model, this action will force a low accuracy and overall model performance of the binary class prediction. This plot gives a better understanding of the model's black box nature, relying on and transparently attributing training features for better understanding and tactical

decisions within military networks.

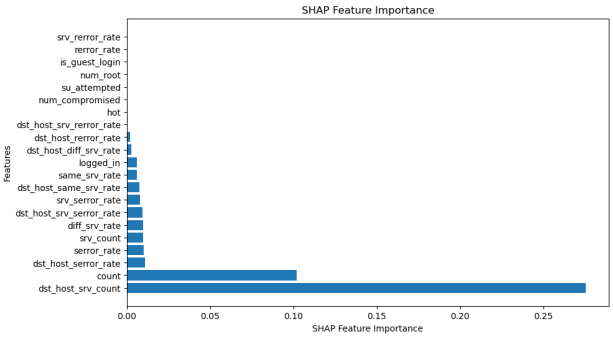

Fig. 3. SHAP feature importance on model output, with test data

Fig. 4 highlights the LIME model probability prediction. It interprets and measures the probability of the model prediction within a specific attack to ascertain if anomalous or benign using the LIME explainability method. As shown in Fig. 4, the LIME explainer is 92% confident that the individual network traffic is normal (1) and 8% certain that of not an attack (0).

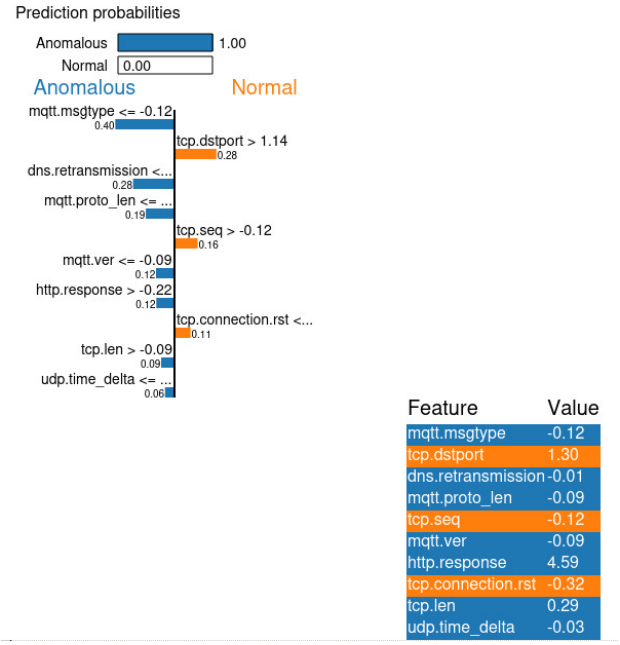

Fig. 4. Sample data LIME model probability prediction

V. CONCLUSION

Adopting an adversarial model within a zero-trust security framework, re-thinks how to respond to advanced persistent threats within cyber warfare scenarios to mitigate constant severe intrusions and attacks. The zero-trust approach helps to enhance the system's overall security posture and resilience. The parameter protection technique ensures a more secure model against gradient exposure and manipulation by either white-box or black-box attacks. Our experimental results have shown that the PP method improves model security while offering a good classification result of anomalous from benign attacks. Additionally, the XAI methods explored in this paper provide trust and reliable results toward effective cyber threat mitigation.

ACKNOWLEDGMENT

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the Innovative Human Resource Development for Local Intellectualization support program (IITP-2023-2020-0-01612) supervised by the IITP (Institute for Information & communications Technology Planning & Evaluation) and Priority Research Centers Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science, and Technology (2018R1A6A1A03024003).

REFERENCES

[1] D. Serpanos and T. Komninos, "The Cyberwarfare in Ukraine," *Computer*, vol. 55, no. 7, pp. 88–91, 2022.

[2] E. C. Nkoro, C. I. Nwakanma, J.-M. Lee, and D.-S. Kim, "Industrial network attack vulnerability detection and analysis using shodan eye scanning technology," in *2022 13th International Conference on Information and Communication Technology Convergence (ICTC)*, 2022, pp. 886–889.

[3] P. Phiayura and S. Teerakanok, "A comprehensive framework for migrating to zero trust architecture," *IEEE Access*, vol. 11, pp. 19 487–19 511, 2023.

[4] D. L. Marino, C. S. Wickramasinghe, and M. Manic, "An adversarial approach for explainable ai in intrusion detection systems," in *IECON 2018 - 44th Annual Conference of the IEEE Industrial Electronics Society*, 2018, pp. 3237–3243.

[5] M. M. Hassan, M. R. Hassan, S. Huda, and V. H. C. de Albuquerque, "A robust deep-learning-enabled trust-boundary protection for adversarial industrial iot environment," *IEEE Internet of Things Journal*, vol. 8, no. 12, pp. 9611–9621, 2021.

[6] V. U. Ihekoronye, S. O. Ajakwe, D. Kim, and J. M. Lee, "Hierarchical intrusion detection system for secured military drone network: A perspicacious approach," in *MILCOM 2022 - 2022 IEEE Military Communications Conference (MILCOM)*, 2022, pp. 336–341.

[7] R. Freter, "Department of Defence (DoD) Zero Trust Reference Architecture, Version 2.0 ," in *Defense Information Systems Agency (DISA) and National Security Agency (NSA)*, July 2022. [Online]. Available: https://dodcio.defense.gov/Portals/0/Documents/Library/(U)ZT_RA_v2.0(U)_Sep22.pdf

[8] M. Usama, A. Qayyum, J. Qadir, and A. Al-Fuqaha, "Black-box adversarial machine learning attack on network traffic classification," in *2019 15th International Wireless Communications & Mobile Computing Conference (IWCMC)*, 2019, pp. 84–89.

[9] A. Kerman, O. Borchert, S. Rose, and A. Tan, "Implementing a zero trust architecture," *National Institute of Standards and Technology (NIST)*, 2020.

[10] M. Abdelaty, S. Scott-Hayward, R. Doriguzzi-Corin, and D. Siracusa, "Gadot: Gan-based adversarial training for robust ddos attack detection," in *2021 IEEE Conference on Communications and Network Security (CNS)*, 2021, pp. 119–127.

[11] J. Wang, J. Pan, I. AlQerm, and Y. Liu, "Def-ids: An ensemble defense mechanism against adversarial attacks for deep learning-based network intrusion detection," in *2021 International Conference on Computer Communications and Networks (ICCCN)*, 2021, pp. 1–9.

[12] N. Capuano, G. Fenza, V. Loia, and C. Stanzione, "Explainable Artificial Intelligence in CyberSecurity: A Survey," *IEEE Access*, vol. 10, pp. 93 575–93 600, 2022.

[13] I. Vaccari, A. Carlevaro, S. Narteni, E. Cambiaso, and M. Mongelli, "explainable and reliable against adversarial machine learning in data analytics," *IEEE Access*, vol. 10, pp. 83 949–83 970, 2022.

[14] Kaggle. Network intrusion detection dataset. [Online]. Available: https://www.kaggle.com/datasets/sampadab17/network-intrusion-detection