Enhancing IoT Security: A Meta-Learning Approach to Adversarial Robustness

Zil e Huma
Edinburgh Napier University
Edinburgh, United Kingdom
zile.huma@napier.ac.uk

Sana Ullah Jan
Edinburgh Napier University
Edinburgh, United Kingdom
S.Jan@napier.ac.uk

Jawad Ahmad
Prince Mohammad Bin Fahd University
Alkhobar, Saudi Arabia
JAhmad@pmu.edu.sa

William J. Buchanan
Edinburgh Napier University
Edinburgh, United Kingdom
B.Buchanan@napier.ac.uk

Nikolaos Pitropakis
Department of Information Technology,
The American College of Greece, Athens, Greece
npitropakis@acg.edu

***Abstract*—The increasing connectivity of Internet of Things (IoT) devices and networks has significantly raised security concerns. Intrusion detection systems (IDSs) serve as a first-line defense mechanism to detect and identify various cyber threats. However, traditional IDSs frameworks come with their own challenges, such as high computational costs, limited generalization to evolving variants of cyberattacks, increasing complexity, and vulnerability to adversarial attacks. This paper proposes a meta-learning-based IDS framework using Model-Agnostic Meta-Learning (MAML) to particularly combat adversarial attacks in IoT networks. The designed architecture optimizes model initialization by performing inner-loop updates using adversarially perturbed data. It aggregates gradients from these adversarially adapted models in the outer loop to achieve a resilient initialization that generalizes well against adversarial attacks. The proposed approach is rigorously evaluated by training and testing the model on three major adversarial attacks: Fast Gradient Sign Method (FGSM), Projected Gradient Descent (PGD), and DeepFool. The experimental outcomes indicate the promising performance of the proposed architecture with an average attack detection accuracy of 95.97%. The compact model size of 0.06MB makes it suitable for deployment on resource-constrained IoT devices and networks. Furthermore, the lower inferencing time ensures the timely detection of intrusion, which is significant for real-time IDSs.**

***Index Terms*—Adversarial Machine Learning, Cybersecurity, Internet of Things, Intrusion Detection, Meta-learning**

I. Introduction

The IoT has revolutionized various industries by enabling seamless connectivity between smart devices, leading to advancements in healthcare, industrial automation, smart cities, and many other applications [1], [2]. The modern IoT system frequently utilizes advanced Machine Learning (ML) and Deep Learning (DL) techniques for optimal decision-making and automating industrial processes [3]. However, the traditional ML/DL-based approaches are highly vulnerable to adversarial attacks. In these attacks, the sophisticated perturbations can manipulate model predictions and pose serious security threats, leading to data breaches, system failures, and unauthorized access to critical infrastructures [4]. The recourse-constrained and heterogeneous nature of IoT systems demands robust learning models that can withstand adversarial manipulations while maintaining high performance.

Traditional techniques, such as Standard Adversarial Training (SAT), Defensive Distillation (DD), and Gradient Masking (GM), have been extensively explored in existing literature to enhance DL models' robustness against adversarial attacks. These methods are effective in certain situations, but they also come with significant limitations. SAT augments training data with adversarial examples and improves the system's robustness. However, the constraint of overfitting to specific perturbations often results in poor generalization to unseen attack types and degrades clean-data accuracy [5]. DD trains a model on softened predictions from a previously trained model, which can obscure gradient information but is still susceptible to stronger adaptive attacks [6]. GM attempts to hide gradient information from attackers but fails against adaptive attack strategies that approximate or circumvent the masked gradients [7]. Additionally, methods like certified defenses using Interval-Bound Propagation (IBP) provide theoretical robustness guarantees with high computational costs, [8], making them impractical for real-time IoT systems. These limitations highlight the need for a more adaptable and generalizable adversarial defense mechanism that can rapidly adjust to new and evolving threats without requiring exhaustive retraining. Meta-learning offers a promising solution to overcome the shortcomings of traditional adversarial defense mechanisms by enabling models to quickly adapt to new attack scenarios rather than relying on fixed training distributions [9], [10]. Unlike standard adversarial training, which requires extensive retraining to counter evolving threats, meta-learning, particularly MAML [11], learns a prior that facilitates rapid adaptation to novel adversarial attacks with minimal fine-tuning. By integrating adversarial training into a meta-learning framework, our approach enhances the robustness and generalization of IoT systems. This strategy ensures a lightweight, adaptive,

979-8-3503-9281-4/25/$31.00 ©2025 IEEE

and efficient solution to combat adversarial attacks in IoT. The key contributions of the paper are summarized in the following.

1) This paper proposes an adversarial defense framework integrating MAML with adversarial training, enabling IoT models to rapidly adapt to evolving adversarial threats. The proposed meta-learning approach enhances generalization and robustness with minimal fine-tuning.
2) The proposed framework is evaluated against multiple adversarial attack strategies, including FGSM, PGD, and DeepFool. The proposed approach significantly improves the model's ability to defend against adversarial attacks by incorporating diverse attack types during training.
3) To validate the effectiveness of the proposed framework we utilized a real-time IoT security dataset IDSIoT2024 [12], which ensures the applicability of our approach in practical IoT environments. Experimental results demonstrate that the proposed approach successfully balances adversarial robustness and clean-data accuracy, making it a scalable and efficient solution for real-world IoT security challenges.

The remainder of the article is organized as follows. Section II presents an overview of some of the latest contributions related to adversarial ML-based solutions. Section III elaborates on the key architecture of the proposed approach. Section IV presents a detailed discussion of experimental outcomes. Section V concludes the research.

II. Related Work

This section summarizes some significant contributions related to AML-based IDS. Yuan et al. [13] proposed a novel IDS to enhance robustness against adversarial attacks by combining DL and conventional ML models. The proposed approach incorporated an adversarial example detector based on Local Intrinsic Dimensionality (LID) and leveraged the low attack transferability between DL and ML models. Roshan et al. [14] explored AML threats to NIDS and introduced three different security frameworks. The performance of designed models was investigated in real-time network environments. In another study. Ali et al. [15] integrated Generative Adversarial Networks (GANs) and knowledge distillation for better efficiency and resilience. GANs were incorporated to generate diverse training data, improving adaptability against various adversarial attacks.

Badjie et al. [16] improved defensive distillation by incorporating a Denoising AutoEncoder (DAE) to improve robustness against data poisoning attacks. In the developed model, a reconstruction and filtering pipeline enhanced the training data's resilience. Chen et al. [17] evaluated the robustness of Federated Learning (FL) based NIDS and proposed FedDef, an optimization-based defense. This proposed scheme improved privacy by maximizing input distance while preserving utility by minimizing gradient distance.

Wang et al. [18] demonstrated that incorporating robustness-promoting regularization specifically during the meta-update stage of MAML is essential for enhancing adversarial robustness in few-shot learning. Yin et al. [19] proposed Adversarial Meta-Learner (ADML), a meta-learning algorithm that enhances adversarial robustness by leveraging both clean and adversarial samples during model initialization. In another study, Ji et al. [11] introduced a multi-improved MAML approach that enhances few-shot malware classification by integrating novel data augmentation techniques and customized neural architectures with tailored learning rate schedules. Alrayes et al. [20] developed an adaptive intrusion detection framework for IoT environments using MAML and few-shot learning to rapidly adapt to diverse attack scenarios. Their approach demonstrated high performance on the UNSW-NB15 and NSL-KDD99 datasets, showing strong potential for resilient, real-time IoT security.

While existing AML-based IDS approaches improve robustness through hybrid models, GANs, defensive distillation, and meta-learning variants, they often lack rapid adaptability to evolving IoT-specific threats, integration of diverse adversarial attacks during training, and validation on realistic IoT datasets. The proposed framework bridges these gaps by combining MAML with adversarial training, enabling fast adaptation, enhancing resilience against multiple attack types, and demonstrating practical effectiveness on the real-time IDSIoT2024 dataset while preserving clean-data accuracy.

III. The Proposed Intrusion Detection System

This section presents the threat modeling, and design of the proposed architecture.

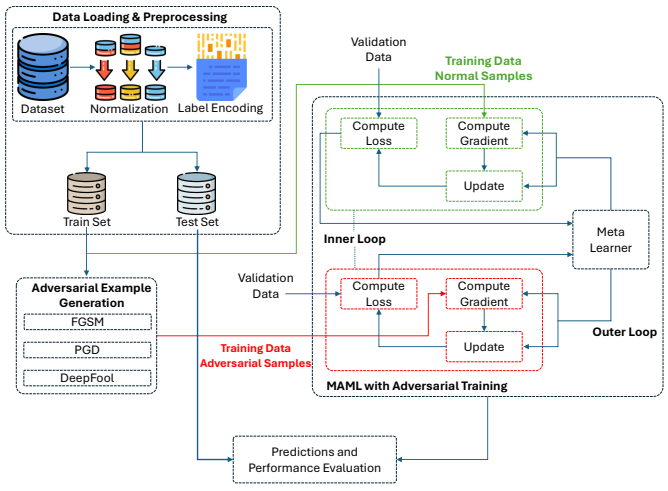


Fig. 1: Workflow of the proposed architecture.

A. Threat Modeling

In IDS, adversarial attacks attempt to mislead the model by introducing small, carefully crafted perturbations to

input samples. We consider three major types of adversarial attacks: FGSM, PGD, and DeepFool.

1) Fast Gradient Sign Method: FGSM perturbs the input sample x in the direction of the gradient of the loss function to maximize the model's classification error as (1):

$$x' = x + \epsilon \cdot \text{sign}\left(\nabla_x L\left(y, f_\theta(x)\right)\right) \tag{1}$$

where, ϵ is a small perturbation magnitude, $L\left(y, f_\theta(x)\right)$ indicate the classification loss function, and $\nabla_x L\left(y, f_\theta(x)\right)$ represents the gradient of the loss with respect to the input.

This method generates adversarial examples in a single step, making it computationally efficient but relatively weak against defenses.

2) Projected Gradient Descent: PGD is an iterative variant of FGSM that repeatedly applies small perturbations while ensuring they remain within a given bound:

$$x^{t+1} = \Pi_{B(x,\epsilon)}\left(x^t + \alpha \cdot \text{sign}\left(\nabla_x L\left(y, f_\theta\left(x^t\right)\right)\right)\right) \tag{2}$$

where, x^t is the perturbed input at iteration t, α represent the step size, $B(x,\epsilon)$ defines an ℓ_∞-bounded neighborhood around x, and $\Pi_{B(x,\epsilon)}$ ensures the perturbed sample stays within the bound.

PGD is one of the strongest adversarial attacks, as it approximates the worst-case perturbation within the allowed limit.

3) DeepFool Attack: DeepFool generates adversarial examples by iteratively linearizing the decision boundary and finding the minimal perturbation r needed to cross it:

$$x' = x + r, \quad \text{where} \quad r = -\frac{f_\theta(x)}{\left\|\nabla_x f_\theta(x)\right\|} \tag{3}$$

DeepFool creates subtle perturbations, making adversarial examples harder to detect while achieving a high misclassification rate.

B. Workflow of the proposed IDS

To mitigate the aforementioned attacks, we propose a MAML-based IDS with adversarial training. The working process of the proposed architecture is presented in Fig. 1.

The following presents a detailed description of the key modules of the proposed architecture.

1) Input Data Processing and Preprocessing: The dataset consists of N labeled samples represented as (4):

$$\mathscr{D} = \left\{\left(x_i, y_i\right)\right\}_{i=1}^{N}, \quad x_i \in \mathbb{R}^d, \quad y_i \in \{0, 1, \ldots, C-1\} \tag{4}$$

where each sample x_i is a d-dimensional feature vector and y_i is a categorical label from C possible attack categories.

To ensure numerical stability and improve model convergence, each feature is standardized using zeromean unit-variance normalization as shown in (5):

$$\tilde{x}_{i,j} = \frac{x_{i,j} - \mu_j}{\sigma_j}, \quad \forall i \in \{1, \ldots, N\}, \quad j \in \{1, \ldots, d\} \tag{5}$$

where μ_j and σ_j are the mean and standard deviation of feature j across the dataset.

The categorical labels are encoded into numerical form using a function $\phi : y \to \mathbb{Z}^+$, mapping each attack category to a unique integer as (6):

$$\tilde{y}_i = \phi\left(y_i\right) \tag{6}$$

where $\tilde{y}_i$ represents the encoded label. The dataset is partitioned into training and testing sets:

$$\mathscr{D}_{\text{train}} \cup \mathscr{D}_{\text{test}} = \mathscr{D}, \quad |\mathscr{D}_{\text{train}}| = 0.7N, \quad |\mathscr{D}_{\text{test}}| = 0.3N \tag{7}$$

where $\mathscr{D}_{\text{train}}$ and $\mathscr{D}_{\text{test}}$ contain disjoint samples.

2) Neural Network Model and Loss Function: The classification model is parameterized by θ and defines a function $f_\theta : \mathbb{R}^d \to \mathbb{R}^C$, mapping an input x to a probability distribution over C classes:

$$p(y \mid x; \theta) = \text{softmax}(z), \quad z = f_\theta(x) \tag{8}$$

where:

$$p(y = c \mid x; \theta) = \frac{\exp\left(z_c\right)}{\sum_{j=1}^{C} \exp\left(z_j\right)} \tag{9}$$

The model is trained to minimize the categorical cross-entropy loss, defined as (10):

$$L\left(y, f_\theta(x)\right) = -\sum_{c=1}^{C} 1(y = c) \log p(y = c \mid x; \theta) \tag{10}$$

where $1(y = c)$ is an indicator function that equals 1 if $y = c$, and 0 otherwise.

3) Adversarial Perturbation Generation: To improve robustness, adversarial perturbations are introduced by modifying input samples in the direction of the gradient of the loss function. The adversarial perturbation δ is computed using gradient ascent on the loss function:

$$\delta = \epsilon \cdot \text{sign}\left(\nabla_x L\left(y, f_\theta(x)\right)\right) \tag{11}$$

where ϵ is a hyperparameter controlling the perturbation magnitude.

The adversarial example x' is then generated by adding δ to the original input:

$$x' = x + \delta \tag{12}$$

For each batch, adversarial training involves augmenting the dataset with both clean and adversarial examples as (13):

$$\mathscr{D}'_{\text{train}} = \left\{\left(x_i, y_i\right) \cup \left(x'_i, y_i\right)\right\}, \quad \forall \left(x_i, y_i\right) \in \mathscr{D}_{\text{train}} \tag{13}$$

Thus, during training, the model is optimized using both clean and adversarial samples, making it resistant to adversarial perturbations.

4) MAML with Adversarial Training: MAML aims to learn an optimal initial set of model parameters θ such that the model can quickly adapt to new attack patterns. The meta-learning process consists of:

Step 1: Task-Specific Adaptation (Inner Loop): A task T_i is sampled from a distribution over tasks $p(T)$, where $T_i = (\mathscr{D}_i, \mathscr{D}'_i)$, consisting of clean samples $\mathscr{D}_i$ and adversarial samples $\mathscr{D}'_i$.

For each task T_i, the model is trained using gradient-based adaptation as (14):

$$\theta'_i = \theta - \alpha \nabla_\theta L\left(\mathscr{D}_i \cup \mathscr{D}'_i; \theta\right) \tag{14}$$

where:

- α is the inner learning rate.
- $L\left(\mathscr{D}_i \cup \mathscr{D}'_i; \theta\right)$ is the loss computed on both clean and adversarial examples.

Expanding the loss function as (15):

$$L\left(\mathscr{D}_i \cup \mathscr{D}'_i; \theta\right) = \sum_{(x,y)\in\mathscr{D}_i\cup\mathscr{D}'_i} -\sum_{c=1}^{C} 1(y=c)\log p(y=c \mid x;\theta) \tag{15}$$

where $\mathscr{D}'_i$ contains adversarial examples x'.

Step 2: Meta-Optimization Across Tasks (Outer Loop): Once task-specific adaptation has been performed across multiple tasks, the meta-update step optimizes the global parameters θ by aggregating gradients:

$$\theta = \theta - \beta \sum_{i=1}^{N_T} \nabla_\theta L\left(\mathscr{D}'_i, Y_i; \theta'_i\right) \tag{16}$$

where N_T is the number of sampled tasks, β is the outer learning rate, and θ'_i is the task-specific adapted parameter.

Using Taylor series approximation, the gradient of the loss function with respect to θ as (17):

$$\nabla_\theta L\left(\mathscr{D}'_i, Y_i; \theta'_i\right) \approx \nabla_\theta L\left(\mathscr{D}'_i, Y_i; \theta\right) - \alpha H_\theta \nabla_\theta L\left(\mathscr{D}_i \cup \mathscr{D}'_i; \theta\right) \tag{17}$$

where H_θ is the Hessian matrix of the loss function. The final meta-update equation is shown in (18):

$$\theta = \theta - \beta \sum_{i=1}^{N_T} \left(\nabla_\theta L\left(\mathscr{D}'_i, Y_i; \theta\right) - \alpha H_\theta \nabla_\theta L\left(\mathscr{D}_i \cup \mathscr{D}'_i; \theta\right)\right) \tag{18}$$

This ensures that the model learns a robust initialization θ that generalizes well across different adversarial settings.

IV. Experimental Outcomes

The performance of the proposed architecture is analyzed using a real-time dataset IDSIoT2024 [12]. Initially, the model is tested under clean conditions (no adversarial attack), followed by implementing three adversaries: FGSM, PGD, and DeepFool. Furthermore, MAML-based adversarial training is applied to improve the system's performance against these attacks. Key hyperparameters are manually tuned through systematic trial-and-error to achieve optimal performance. Additionally, inference time and model size are evaluated to analyze computational performance and resource efficiency.

A. Performance Evaluation on Clean Data

The system achieved accuracy, precision, recall, and an F1 score of 97.87%, 97.84%, 97.87%, and 97.59%, respectively, in an ideal scenario without adversarial interference. These metrics indicate that the model is highly effective at correctly identifying normal and malicious instances, demonstrating its capability to learn and generalize patterns. This high performance establishes a strong baseline, showcasing the model's effectiveness as represented in the normalized confusion matrix in Fig. 2.

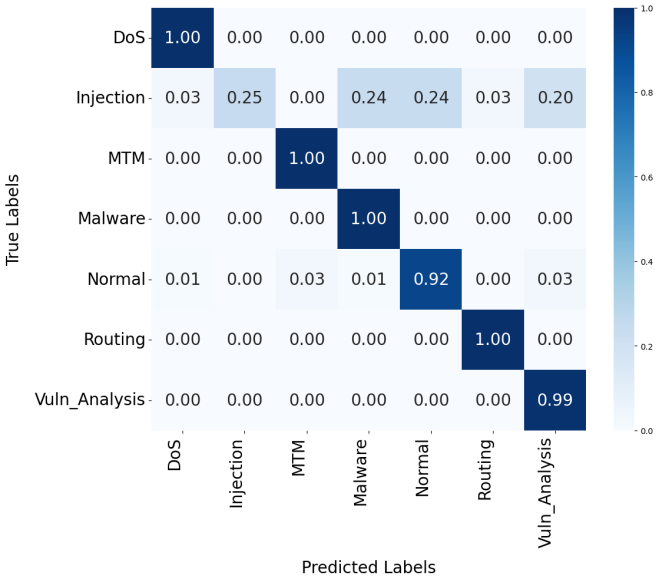


Fig. 2: The normalized confusion matrix for clean data evaluation.

B. Performance Evaluation (Before and After Adversarial Training)

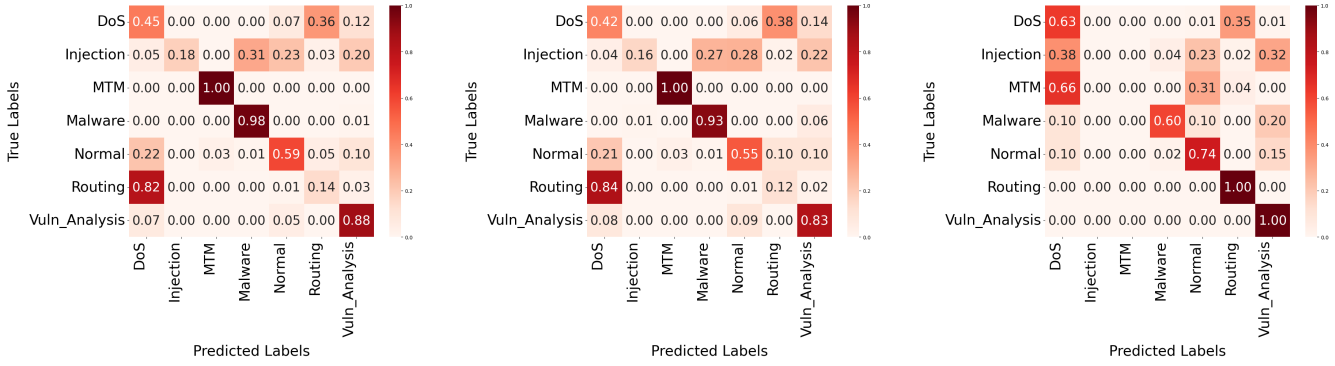
The experimental outcomes for the attack detection under all scenarios are summarized in Table I. The model's vulnerability to adversarial attacks became evident when exposed to three distinct adversarial threats. Before adversarial training, the model's performance significantly declined under all three attack scenarios, as shown in the normalized confusion matrices in Fig. 3.

For the FGSM attack, the accuracy and an F1-score dropped to approximately 60%. The PGD attack resulted in even lower metrics, with an accuracy of 56.73%. Although slightly less effective, the DeepFool attack still substantially impacted the model, reducing its accuracy to 70.90%.

These results demonstrate the model's susceptibility to adversarial manipulations, where even small perturbations can significantly mislead the classifier. The PGD attack is the most potent among the three, leading to the lowest

TABLE I: Performance comparison of proposed architecture before adversarial training (BAT) and after adversarial training (AAT)

Metric	FGSM		PGD		DeepFool	
	BAT	AAT	BAT	AAT	BAT	AAT
Accuracy	0.6004	0.9629	0.5673	0.9472	0.7090	0.9692
Precision	0.6279	0.9651	0.6029	0.9550	0.7227	0.9726
Recall	0.6004	0.9629	0.5673	0.9472	0.7090	0.9692
F1 Score	0.6046	0.9636	0.5750	0.9495	0.6919	0.9692



(a) FGSM case (b) PGD case (c) DeepFool case

Fig. 3: Multiclass performance evaluation before adversarial training.

accuracy and F1 scores. This indicates its effectiveness in creating adversarial examples that closely resemble legitimate data while deceiving the model. In contrast, the DeepFool attack is comparatively less effective, suggesting a slightly better resistance to this perturbation, although the performance was still substantially lower than on clean data.

After incorporating adversarial training into the MAML framework, the model exhibited a remarkable improvement in robustness across all attack types. Fig. 4 illustrates the system performance after adversarial training.

Against the FGSM attack, the model's accuracy and recall are significantly improved. Similarly, under the PGD attack, the model's accuracy increased to 94.72% and an F1 score of 94.95%. This enhancement reflects the model's significant resilience, demonstrating that adversarial training effectively mitigates the impact of even the most potent PGD attack. For the DeepFool attack, the model maintained high robustness, achieving an accuracy of 96.92%. The improvement in performance across all metrics underscores the model's enhanced capacity to generalize and defend against diverse attack vectors.

The substantial performance gains after adversarial training demonstrate the effectiveness of integrating MAML with adversarial strategies. The meta-learning capability of MAML enables the model to learn adaptable parameters that generalize well even in the presence of adversarial perturbations. This adaptability allows the model to quickly adjust to unseen attacks, significantly enhancing its robustness compared to traditional training approaches. The near-clean-data performance achieved after adversarial training underscores the model's ability to maintain high accuracy, precision, recall, and F1 scores, making it a highly reliable solution for IDS.

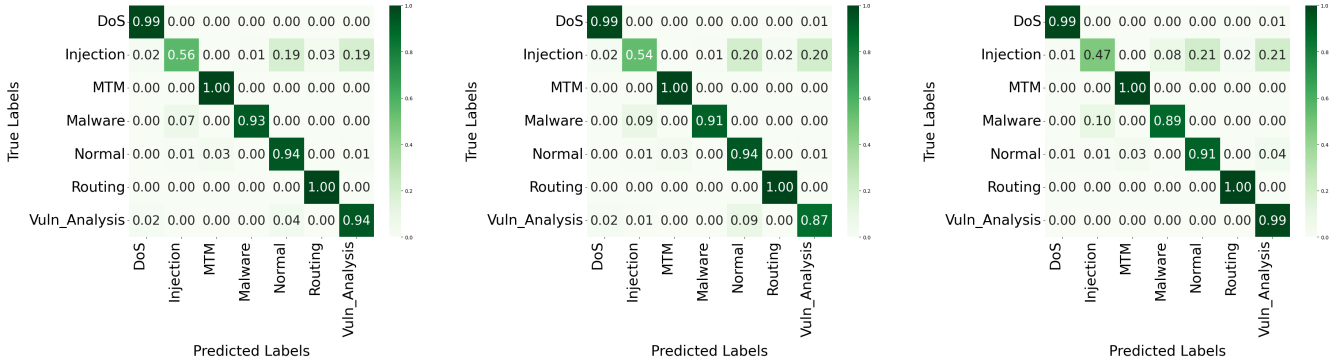
C. Computational Performance and Resource Efficiency

The proposed architecture demonstrated competitive inference times across different attack scenarios. The average inferencing time for the three adversaries is approximately 50 microseconds, highlighting the framework's rapid response capability. This low latency ensures real-time detection and mitigation of adversarial threats, making it suitable for practical cybersecurity applications. The low inferencing time showcases the proposed architecture's ability to enhance system resilience without compromising computational performance. Therefore, it offers a scalable and effective solution for securing IDSs.

The proposed architecture is also highly resource-efficient, with a model size of only 0.06MB. This compact size demonstrates its potential for deployment in resource-constrained environments such as IoT devices, edge computing systems, and other embedded systems. The lightweight nature of the model, coupled with its high accuracy and robustness, makes it an ideal choice for practical cybersecurity solutions.

V. CONCLUSION

The article proposed lightweight and efficient MAML-based adversarial training to combat adversaries in IDSs. The proposed architecture is tested under different cases, such as clean data and adversarial perturbations. Initially,



(a) FGSM case (b) PGD case (c) DeepFool case

Fig. 4: Multiclass performance evaluation after adversarial training.

the model achieved high accuracy on clean data, but its performance significantly degraded under adversarial attacks. To address this, we integrated adversarial training using MAML, which substantially improved robustness across multiple attack scenarios. The model also exhibited efficient computational performance under three adversaries. Moreover, its lightweight architecture ensured resource efficiency, making it suitable for IoT applications. For future endeavors, we aim to extend our work by implementing on embedded GPU devices for real-time IoT applications. Additionally, we plan to explore automated hyperparameter optimization, evaluate on diverse IoT datasets across domains, and conduct real hardware experiments to further validate generalizability and practical deployment.

Acknowledgment

The authors would like to express their sincere gratitude to Edinburgh Napier University for its generous support and sponsorship of this project.

References

[1] J. Ahmad, S. Latif, I. U. Khan, M. S. Alshehri, M. S. Khan, N. Alasbali, and W. Jiang, "An interpretable deep learning framework for intrusion detection in industrial internet of things," *Internet of Things*, p. 101681, 2025.

[2] O. Aouedi, T.-H. Vu, A. Sacco, D. C. Nguyen, K. Piamrat, G. Marchetto, and Q.-V. Pham, "A survey on intelligent internet of things: Applications, security, privacy, and future directions," *IEEE communications surveys & tutorials*, 2024.

[3] M. A. Khan, M. A. Khan, S. U. Jan, J. Ahmad, S. S. Jamal, A. A. Shah, N. Pitropakis, and W. J. Buchanan, "A deep learning-based intrusion detection system for mqtt enabled iot," *Sensors*, vol. 21, no. 21, p. 7016, 2021.

[4] H. Baniecki and P. Biecek, "Adversarial attacks and defenses in explainable artificial intelligence: A survey," *Information Fusion*, p. 102303, 2024.

[5] L. Rice, E. Wong, and Z. Kolter, "Overfitting in adversarially robust deep learning," in *International conference on machine learning*. PMLR, 2020, pp. 8093–8104.

[6] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *2017 ieee symposium on security and privacy (sp)*. Ieee, 2017, pp. 39–57.

[7] A. Athalye, N. Carlini, and D. Wagner, "Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples," in *International conference on machine learning*. PMLR, 2018, pp. 274–283.

[8] M. Mirman, M. Baader, and M. Vechev, "The fundamental limits of interval arithmetic for neural networks," *arXiv preprint arXiv:2112.05235*, 2021.

[9] T. Zoppi, M. Gharib, M. Atif, and A. Bondavalli, "Meta-learning to improve unsupervised intrusion detection in cyber-physical systems," *ACM Transactions on Cyber-Physical Systems (TCPS)*, vol. 5, no. 4, pp. 1–27, 2021.

[10] T. Hospedales, A. Antoniou, P. Micaelli, and A. Storkey, "Meta-learning in neural networks: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 9, pp. 5149–5169, 2021.

[11] Y. Ji, K. Zou, and B. Zou, "Mi-maml: classifying few-shot advanced malware using multi-improved model-agnostic meta-learning," *Cybersecurity*, vol. 7, no. 1, p. 72, 2024.

[12] M. Koppula and L. J. L.M.I., "A real time dataset "idsiot2024"," 2024. [Online]. Available: https://dx.doi.org/10.21227/gfaz-t124

[13] X. Yuan, S. Han, W. Huang, H. Ye, X. Kong, and F. Zhang, "A simple framework to enhance the adversarial robustness of deep learning-based intrusion detection system," *Computers & Security*, vol. 137, p. 103644, 2024.

[14] K. Roshan, A. Zafar, and S. B. U. Haque, "Untargeted white-box adversarial attack with heuristic defence methods in real-time deep learning based network intrusion detection system," *Computer Communications*, vol. 218, pp. 97–113, 2024.

[15] T. Ali, A. Eleyan, T. Bejaoui, and M. Al-Khalidi, "Lightweight intrusion detection system with gan-based knowledge distillation," in *2024 International Conference on Smart Applications, Communications and Networking (SmartNets)*. IEEE, 2024, pp. 1–7.

[16] B. Badjie, J. Cecílio, and A. Casimiro, "Denoising autoencoder-based defensive distillation as an adversarial robustness algorithm against data poisoning attacks," *ACM SIGAda Ada Letters*, vol. 43, no. 2, pp. 30–35, 2024.

[17] J. Chen, Y. Zhao, Q. Li, X. Feng, and K. Xu, "Feddef: defense against gradient leakage in federated learning-based network intrusion detection systems," *IEEE Transactions on Information Forensics and Security*, vol. 18, pp. 4561–4576, 2023.

[18] R. Wang, K. Xu, S. Liu, P.-Y. Chen, T.-W. Weng, C. Gan, and M. Wang, "On fast adversarial robustness adaptation in model-agnostic meta-learning," *arXiv preprint arXiv:2102.10454*, 2021.

[19] C. Yin, J. Tang, Z. Xu, and Y. Wang, "Adversarial meta-learning," *arXiv preprint arXiv:1806.03316*, 2018.

[20] F. S. Alrayes, S. U. Amin, and N. Hakami, "An adaptive framework for intrusion detection in iot security using maml (model-agnostic meta-learning)," *Sensors (Basel, Switzerland)*, vol. 25, no. 8, p. 2487, 2025.