Indonesian Journal of Electrical Engineering and Computer Science
Vol. 37, No. 2, February 2025, pp. 937~947
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v37.i2.pp937-947

Adversarially robust federated deep learning models for intrusion detection in IoT

El Mahfoud Ennaji, Salah El Hajla, Yassine Maleh, Soufyane Mounir
LaSTI Laboratory, National School of Applied Sciences Khouribga, Sultan Moulay Slimane University, Beni Mellal, Morocco

Article Info

Article history:

Received May 4, 2024
Revised Sep 15, 2024
Accepted Sep 29, 2024

Keywords:

Adversarial attacks
Adversarial training
Deep learning
Federated learning
Intrusion detection
IoT

ABSTRACT

Ensuring the robustness, security, and privacy of machine learning is a pivotal objective, crucial for unlocking the complete potential of the internet of things (IoT). Deep neural networks have proven to be vulnerable to adversarial perturbations imperceptible to humans. These perturbations can give rise to adversarial attacks, leading to erroneous predictions by deep neural networks, particularly in intrusion detection within the IoT environment. This paper introduces a federated adversarial learning framework designed to protect both data privacy and deep neural network models. This framework consists of federated learning for data privacy and adversarial training on IoT devices to enhance model robustness. The experiments show that adversarial training at the Fog node devices significantly improves the robustness of a federated learning model against adversarial attacks when compared to normal training. Furthermore, the proposed adversarial deep federated learning model is validated using the Edge-IIoTset dataset, achieving an accuracy rate of 91.23% in the detection of attacks.

Corresponding Author:

El Mahfoud Ennaji
LaSTI Laboratory, National School of Applied Sciences Khouribga, Sultan Moulay Slimane University
Beni Mellal 23000, Morocco
Email: elmahfoud.ennaji@usms.ac.ma

1. INTRODUCTION

The proliferation of IoT devices has transformed our world by interlinking a vast array of smart devices across various domains such as smart homes, transportation, industrial automation, and healthcare. However, this interconnectedness introduces numerous vulnerabilities, making the IoT landscape susceptible to a wide range of security threats [1]-[4] including network intrusions, which can be identified and mitigated using network intrusion detection systems (NIDS) [5]. Traditional NIDS solutions struggle with the sheer volume, heterogeneity, and dynamic nature of IoT data, necessitating advanced techniques rooted in machine learning (ML) [6] and deep learning (DL) to enhance IoT security [7]. While DL can identify intricate patterns in vast datasets, its application to IoT security faces challenges like data privacy concerns, resource limitations, and the decentralized nature of IoT networks. To address data privacy challenges, Federated Learning (FL) offers a promising solution.

FL is a distributed approach to machine learning, wherein a central server gathers model updates computed locally by multiple decentralized devices on their private data and consolidates these updates to train a globally learned model [8], [9]. This approach offers several advantages: clients avoid sharing their data, while still benefiting from a model trained on the collective dataset of all participating clients, and distributing the computational workload across all contributing devices [8]. Despite these benefits, models trained through FL are susceptible to adversarial attacks [10], similar to centrally trained models.

Adversarial attacks refer to inputs given to machine learning models that have been subtly altered, leading to misclassification with strong confidence. Techniques to generate such adversarial samples include the fast gradient sign method (FGSM), Torres *et al.* [11], which perturbs input data by adjusting it in the direction that maximizes the model's loss, controlled by a small value (epsilon). The basic iterative method (BIM), Kurakin *et al.* [12], enhances FGSM by iteratively adjusting pixel values in small steps within the original image bounds, making the perturbations less noticeable. Projected gradient descent (PGD) [13] is an advanced iterative attack that improves upon FGSM by incorporating more iterations and randomness, resulting in more effective and subtle adversarial perturbations. PGD shares the underlying mechanism with BIM but enhances its effectiveness with additional iterations and careful pixel adjustments.

To counter adversarial attacks, adversarial training (AT) provides a promising defense. Adversarial training is a technique used to enhance the robustness of models against adversarial attacks, where the model is deliberately exposed to perturbed or adversarial samples during training. These samples undergo label modification to align with their accurate values before input into the neural network. This process enables the trained model to acquire the capability to withstand adversarial samples. This technique was initially introduced by Rumelhart [14]. A machine learning model is deemed susceptible to adversarial samples if it reacts to minute alterations in a plausible input, leading to a change in the predicted label. Consequently, a deep neural model is considered robust against adversarial examples if it remains insensitive to small changes in potential inputs during the prediction phase [15]. In essence, adversarial robustness aims to ensure that minor alterations in the input of the deep learning model do not lead to significant variations in the predicted output.

Several studies have been conducted on federated training for IoT attack datasets to classify anomalies and threats. Shah *et al.* [16] investigated the integration of adversarial training in FL, particularly with non-IID data and a fixed communication budget. They found that adversarial training led to a decline in both natural and adversarial accuracies compared to centralized training, attributed to local adversarial training epochs. To address this, they proposed the FedDynAT algorithm, which improved accuracy and reduced model drift. Rashid *et al.* [17] studied the impact of adversarial machine learning (AML) attacks on IoT intrusion detection systems (IDSs), using the DS2OS dataset. They trained models like decision tree, multilayer perceptron, and random forest and evaluated AML attacks such as FGSM, JSMA, and C&W. The results showed significant performance declines, especially with JSMA and C&W, with MLP models demonstrating relatively higher robustness. Adversarial training using C&W attack samples was effective in maintaining classifier performance, with retrained models achieving accuracies close to the baseline. Ni *et al.* [18] introduced rFedFW, a Byzantine-robust FL framework for IoT, employing dual filtering and adaptive weight adjustments to mitigate malicious gradients. Ibitoye *et al.* [19] proposed DiPSeN, a differentially private self-normalizing neural network, to enhance adversarial robustness in FL without compromising privacy. Ferrag *et al.* [20] presented the Edge-IIoTset dataset, assessing machine learning techniques in both centralized and federated scenarios, achieving 93.04% accuracy for IID data and 77.04% for non-IID data in federated learning with five rounds.

However, while these contributions have been crucial, they have not extensively addressed adversarial attacks in the context of IoT intrusion detection and federated learning. These Studies often encountered significant limitations, such as a notable decline in both natural and adversarial accuracies when incorporating adversarial training in federated learning frameworks. Additionally, adversarial training [21], [22] methods, while enhancing robustness, often compromised privacy guarantees and required a substantial portion of adversarial samples for effective training. Complex integration methods, such as dual filtering mechanisms and adaptive weight adjustments, added further implementation challenges, potentially impacting scalability and computational efficiency. Performance variations in non-IID data and the specificity of certain datasets also highlighted the need for broader validation and more extensive evaluations to ensure robustness and applicability across diverse IoT environments.

This research paper proposes a federated learning approach for IoT-based intrusion detection that incorporates adversarial training to enhance the robustness of the FL model against adversarial attacks. Our method facilitates the adversarial training of deep learning models directly on IoT fog computing devices, thereby eliminating the need to transmit device data to a central server. Using the IIoTEdge dataset, we compare the robustness of baseline federated learning with federated learning that employs adversarial training against adversarial evasion attacks. Additionally, we focus on addressing the vulnerability of federated learning models to adversarial attacks. Our approach aims to enhance the robustness of these models by incorporating adversarial training.

The main contributions of this research are outlined:

- We proposed a federated learning model for intrusion detection in IoT environment: we enable deep learning model training directly on IoT fog computing devices to protect the fog data privacy and make the network more efficient.

- We implement the adversarial federated training on IoT fog computing devices to make our model more robust against adversarial attacks.
- We test the robustness of the baseline federated learning and adversarial federated learning model against adversarial evasion attacks, particularly the projected gradient descent attack over a range of perturbation sizes.

The rest of this paper is structured as follows: section 2 presents our proposed system model, outlining the architecture and components of the framework, and adopted methodology. we will demonstrate the practical implementation of our federated learning and adversarial training approach, explaining how it addresses privacy concerns and enhances model robustness. Section 3 is dedicated to showcasing the results of our comprehensive experiments and a detailed discussion of our findings. we will provide empirical evidence of the effectiveness of our approach, highlighting its relevance and the improvements achieved in terms of robustness against adversarial attacks. Finally, section 4 we offer a concluding summary of our contributions and delineate avenues for future research.

2. METHOD

This section describes the federated learning framework, the overall architecture of the FL model, the model architecture, the data description, the preprocessing techniques, the evaluation metrics used to assess the federated learning model, and the experimental setup.

2.1. Federated learning framework

The proposed framework consists of federated learning for data privacy and adversarial training for the clients to ensure model robustness. We consider a deep neural network (DNN) as the model for each client in federated learning to analyze network traffic and detect attacks on the IoT device. We integrated an adversarial training technique to make our model robust against adversarial attacks, and conducted comprehensive experiments to adjust the model using various parameters in order to identify the most suitable values to utilize.

The schematic representation of our proposed intrusion detection system (IDS) based on deep neural network federated learning (DNN-FL) is depicted in Figure 1. Initially, the central server disseminates the global DNN-FL model to participating clients. Each client then trains its local model using its local data on the fog node device. Subsequently, the fog node device transmits its local model parameters and corresponding weights to the central server. Following this, the central server aggregates local models from all clients by applying the FedAvg algorithm [8]. This aggregation step, executed by the central server, averages the parameters of the client models, effectively combining them into an updated global model. Finally, the aggregated global model is returned to the Fog Node devices for the subsequent training round. The algorithm of federated learning is shown in Algorithm 1.

To assess our model's performance comprehensively, we will conduct evaluations with and without adversarial training in federated learning. This comparative study aims to analyze the impact of adversarial training on our model's effectiveness.

Normal training scenario: the normal training process typically involves fog node devices training their models locally on their respective local datasets, before sending model updates to the server for aggregation, preprocessing techniques are applied to the local dataset. These preprocessing steps involve techniques such as data normalization, feature scaling, transformation, and balancing to enhance data quality and privacy. Once preprocessed, the fog nodes train on it and send their model updates to the server, which aggregates these updates using federated averaging (FedAvg) algorithm to create a global model.

Adversarial training scenario: The purpose of adversarial training [11] is to improve the robustness of the classifier by including adversarial samples in the training set. These adversarial samples are then combined with normal samples. In our case, we take 50% for adversarial simple and 50% for normal simple and the local model is trained on the combined dataset to enhance robustness against adversarial attacks. The adversarial samples are generated using PGD [23] attacks with perturbation strengths (ε=0.3). The choice of perturbation strength ε=0.3 is based on findings from Madry *et al.* [13], who suggests that this level of perturbation is strong enough to create challenging adversarial examples that can significantly enhance the robustness of the model without rendering the input unrecognizable. It strikes a balance between making the model robust and maintaining the fidelity of the input data. The rationale for including PGD attacks during training is that PGD is considered a strong adversarial example generator that can produce perturbations that are likely to fool the model. By training with these examples, the model can learn to resist them.

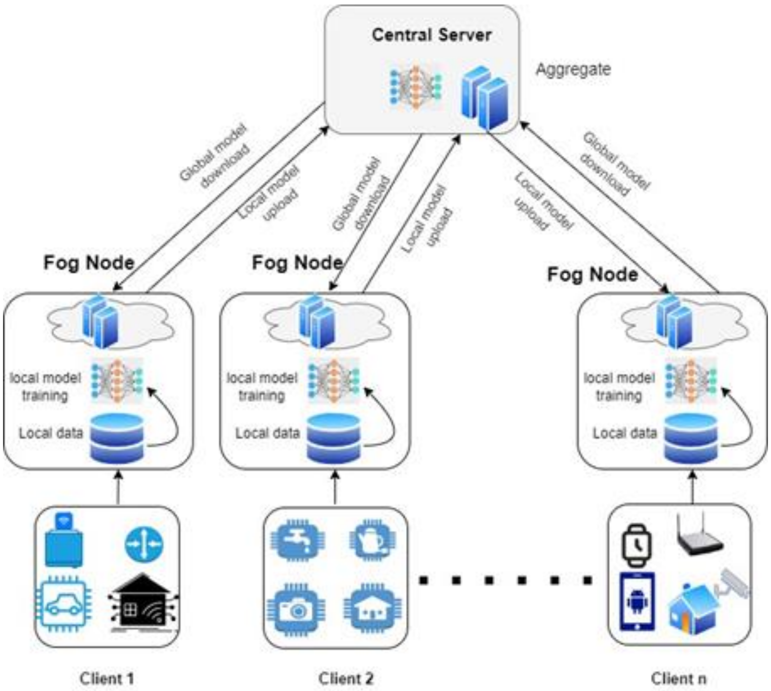


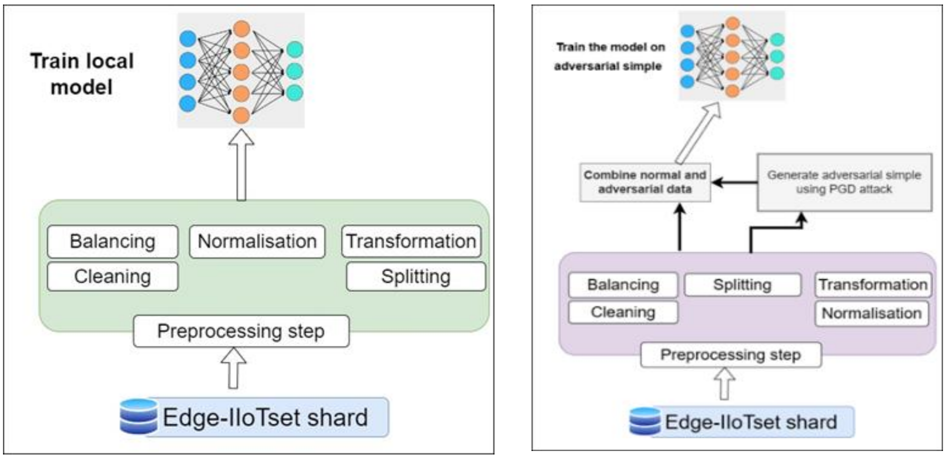
Figure 1. Scheme of our framework

Algorithm 1. DNN-FL based IDS

```
Require: General DNN-FL Model, Set of Clients {C1, C2, …, Cn}, Fog Node
Device {F1, F2, …, Fn}, Averaging Algorithm for Model Aggregation
Ensure: Global DNN-FL Model
Server Distribution:
GlobalMOdel<-InitializeGeneralDNNFLModel()
Distribute (GlobalModel, Client)
Federated Learning Iterations:
for each iteration do
    Local Computation on Fog Nodes:
    for
        LocalModel<-RunLocalDNNFL(Fi, GlobalModel)
        SendModelToServer(Fi, LocalModel)
    end for
    Model Aggregation at Server:
    AggregatedModel<- AggregatedModels(Clients)
    Global Model Update:
    GlobalModel<-UpdateGlobalModel(AggregatedModel)
    Model Distribution to Fog Nodes:
    Distribute (GlobalModel, Fog Nodes)
end for
return GlobalModel
```

We implement adversarial training using IBM's publicly available adversarial robustness toolbox (ART) framework [6]. Figure 2 illustrates the federated learning for normal training and adversarial training schema in the fog node device.

Our implementation of federated learning utilizes the Python-based flower framework [24]. The simulation environment consists of five clients and one central cloud server. The dataset is distributed independently and identically (IID) across the clients, divided into 5 unique subsets to ensure each client receives a distinct portion of the data. For local training at the client level, each client runs for 25 epochs. This epoch count was selected based on observations of model convergence, ensuring that the training process is both efficient and effective. At the central server, we have configured 2 aggregation rounds. This configuration allows us to evaluate the effectiveness of adversarial training in FL as a defense against adversarial evasion attacks.



(a) (b)

Figure 2. Fog node architecture (a) with normal training and (b) with adversarial training

2.2. Model architecture

The proposed DNN-FL model has two hidden layers with 90 nodes, utilizing L2 regularization, ReLU activation function, and Softmax for classification. The training process employs the cross-entropy loss function, the Adam optimizer set at a learning rate of 0.01, and involves 25 local epochs with a batch size of 800 for each participant. Local epochs are set at 25 to ensure sufficient learning while avoiding excessive computation, considering that it multiplies with the number of participants and training rounds, which are limited to 2 to mitigate communication overhead. The overall training is conducted over two rounds. The choice of hyperparameters is geared towards optimizing the learning capabilities of our model across multiple participants, and taking into account the resource constraints of IoT devices. Table 1 shows the different parameters used to implemente the federated deep neural classifier.

Table 1. Federated learning setup

Parameters	Value
Number of participants	‘5’
Hidden layer	‘2’
Hidden node	‘90’
Regulation	‘L2’
Activation function	‘Relu’
Classification function	‘softmax’
Loss function	‘Cross-entropy’
Optimizer	‘Adam’
learning rate	‘0.01’
Local epochs	‘25’
Batch size	‘800’
Training rounds	‘2’

2.3. Dataset description

Introduced by Beutel *et al.* [24], the Edge-IoTset presents a robust dataset specifically designed for IoT environments, enhancing cybersecurity measures across various applications. The dataset supports advanced machine learning techniques for intrusion detection and can be utilized in both centralized and federated learning settings. Developed within a dedicated IoT/IoT laboratory, the dataset incorporates a diverse array of IoT devices and sensors, including but not limited to, cost-effective digital temperature and humidity sensors, water level and pH meters, ultrasonic sensors, heart rate sensors, soil moisture sensors, and flame sensors, covering over ten distinct types.

The dataset addresses multiple types of cybersecurity threats classified into five major categories: denial of service, distributed denial of service [25], intelligence gathering, man in the middle (MITM) [26], injection, and malware attacks [27]. This comprehensive dataset is instrumental for refining the efficacy of intrusion detection systems within IoT frameworks, providing detailed insights into attack vectors and vulnerabilities. Details on the dataset's statistics and classifications are provided in Table 2.

Table 2. Edge-IIoTset statistics and taxonomies

IoT traffic	Class	Records	Total
Normal	Normal	11223940	11223940
Attack	Backdoor	24862	9728708
	HTTP DDoS	229022	
	ICMP DDoS	2914354	
	TCP DDoS	2020120	
	UDP DDoS	3201626	
	Fingerprinting	1001	
	Man in the Middle	1229	
	Password	1053385	
	Port Scanning	22564	
	Ransomware	10925	
	SQL injection	51203	
	Uploading	37634	
	Vulnerability Scanner	145869	
	Cross-Site Scripting	15915	
Total			20952648

2.4. Data preprocessing

The initial steps involve examining and preparing the Edge-IioTset dataset through a series of preprocessing operations. Step 1: involves dividing the dataset among five different clients to simulate a federated learning scenario. All the subsequent steps are applied on the client side. Step 2: focuses on cleaning the data by removing NaN values and duplicates. Step 3: involves deleting unnecessary data elements such as ports, IP addresses, timestamps, and payload information. In step 4, we transform categorical variables into binary or dummy variables. Step 5: employs a splitting method to randomly divide the dataset into training and testing subsets. In step 6, categorical variables are encoded to numerical. Step 7: applies features normalization [28], ensuring each has zero mean and unit variance. Finally, step 8: integrates the synthetic minority over-sampling technique (SMOTE) [29] to address class imbalance issues.

2.5. Evaluation metrics

This section describes the evaluation metrics used to assess the federated deep learning model. The key metrics include:

- Accuracy represents the proportion of correct predictions and is computed as follows:

$$Accuracy=\frac{\text{Number of correct predictions}}{\text{Total number of predictions}}$$

- Precision determines the ratio of correctly predicted positive instances relative to the total predicted positives:

$$\text{Precision}=\frac{\text{TP}}{\text{TP+FP}}$$

- Recall computes the ratio of correctly predicted positive outcomes to all outcomes in a specific class:

$$\text{Recall}=\frac{\text{TP}}{\text{TP+FN}}$$

- F1 Score is a combined metric of recall and precision, ranging from 0 to 1, defined as:

$$\text{F1 score}=\frac{2\times(\text{recall}\times\text{precision})}{\text{recall+precision}}$$

- Confusion matrix records the number of accurate and inaccurate predictions generated by a classification model.

2.6. Experimental setup

The experiments were carried out using a system powered by an 11th-generation Intel Core i7 processor, an NVIDIA Quadro M2000M GPU, and 32 GB of RAM. Flower is a flexible and user-friendly framework for federated learning, supporting various machine learning libraries like PyTorch and TensorFlow. It facilitates collaborative model training among edge devices while preserving local data.

3. RESULTS AND DISCUSSION

This section delves into the comprehensive analysis of our findings. Through rigorous evaluation and comparison, we shed light on the effectiveness of deep neural adversarial training in federated learning setting. Our discussion encompasses insights gleaned from the results, highlighting key observations and implications.

3.1. Performance of deep neural federated learning model

The evaluation of the federated model in various scenarios, normal training, and adversarial training for an IoT intrusion detection dataset is detailed in Tables 3 and 4. Table 3 represents normal training, the model exhibited commendable performance with an accuracy of 89.89%, indicating the percentage of correctly classified instances. The precision, measuring the accuracy of positive predictions, stood at 95.19%, while recall, gauging the model's ability to identify all relevant instances, reached 89.89%. The F1 score, balancing precision and recall, registered at 89.20%.

Table 4 represents the results of adversarial training, the federated model demonstrated further improvement. The accuracy increased to 91.23%, showcasing the model's enhanced overall correctness. Precision, reflecting the accuracy of positive predictions, achieved a value of 91.32%, while recall, measuring the model's capacity to identify relevant instances, reached 91.23%, and an F1 score reached 89.79%. %, these results indicate that adversarial training contributed positively to the model's robustness and performance across various metrics.

From these findings, we can notice that adversarial training enhances the model's ability to correctly classify instances and make accurate positive predictions. The increase in accuracy and precision suggests that the model becomes more reliable in identifying true positives and reducing false positives. Additionally, the improved recall indicates that the model is better at detecting all relevant instances, which is crucial for effective intrusion detection in an IoT environment.

Table 3. Performance of model in normal training

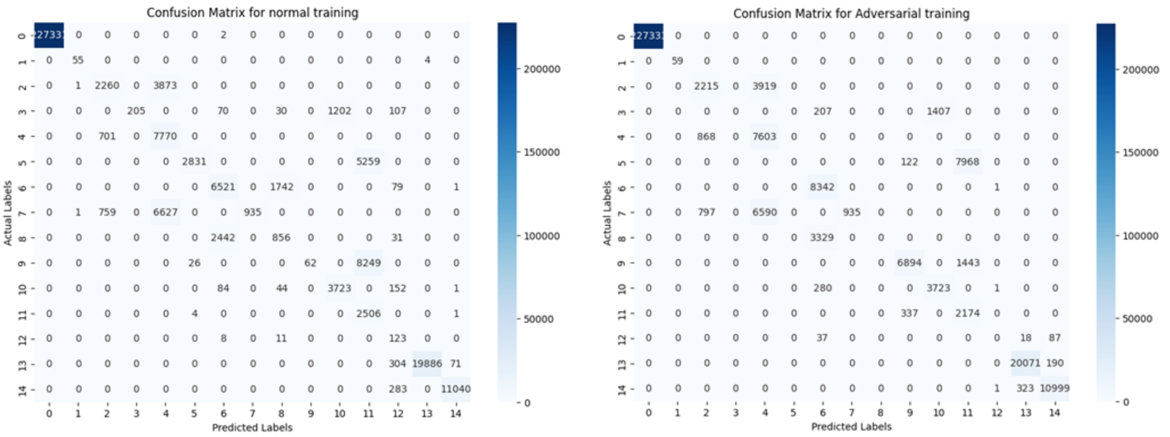
Accuracy	Precision	Recall	F1
89.89 %	95.19%	89.89%	89.20%

Table 4. Performance of model with adversarial training

Accuracy	Precision	Recall	F1
91.23%	91.32%	91.23%	89.79%

Figure 4 represents the confusion matrices of federated learning classification model, with normal training and with adversarial training. FL model in normal training performs exceptionally well, with high true positive rates for most classes, the model demonstrates a strong diagonal concentration, particularly for classes 0 and 14, indicating accurate predictions. In contrast, the adversarial training matrix demonstrated further improvement for some classes, and a decrease for other classes, likely due to the trade-off introduced by AT. Despite these variations, the true positives remain consistently high.

These results demonstrate that while normal training provides a solid baseline with high accuracy for most classes, adversarial training introduces a balance between robustness and accuracy. The improvements in some classes suggest that the model is better equipped to handle adversarial attacks, making it more resilient in challenging scenarios. However, the trade-offs seen in other classes underscore the complexity of optimizing model performance across all categories simultaneously.



(a) (b)

Figure 4. Confusion matrix of the model with (a) normal training and (b) adversarial training

3.2. The performance of the original and the robust FL model against evasion adversarial attacks

We evaluate our classifiers against adversarial attacks generated by the PGD method [30]. The PGD attacker is a common adversarial attack that aims to mislead classifiers by applying small perturbations to the input data. The `eps` parameter controls the size of these perturbations. The evaluation is conducted over a range of increasing `eps` values to assess how the classifier accuracies deteriorate as the perturbation size increases. Figure 5 shows that the robust classifier's accuracy degrades much slower than the original classifier's as the perturbation size increases, indicating that the robust classifier is more resistant to adversarial attacks. The original classifier's accuracy drops sharply even with small perturbations and approaches zero as the perturbation size increases, highlighting its vulnerability to such attacks. The robust classifier, while also experiencing a decrease in accuracy, maintains a higher level of accuracy across all tested perturbation sizes, affirming its effectiveness in mitigating the impact of adversarial noise. These results demonstrate that adversarial training significantly enhances the classifier's resilience against adversarial perturbations. The original classifier's rapid decline in accuracy underscores the necessity for robust training methods to protect against adversarial attacks. The robust classifier's ability to maintain higher accuracy despite increasing perturbation sizes demonstrates its enhanced reliability and effectiveness against adversarial attacks.

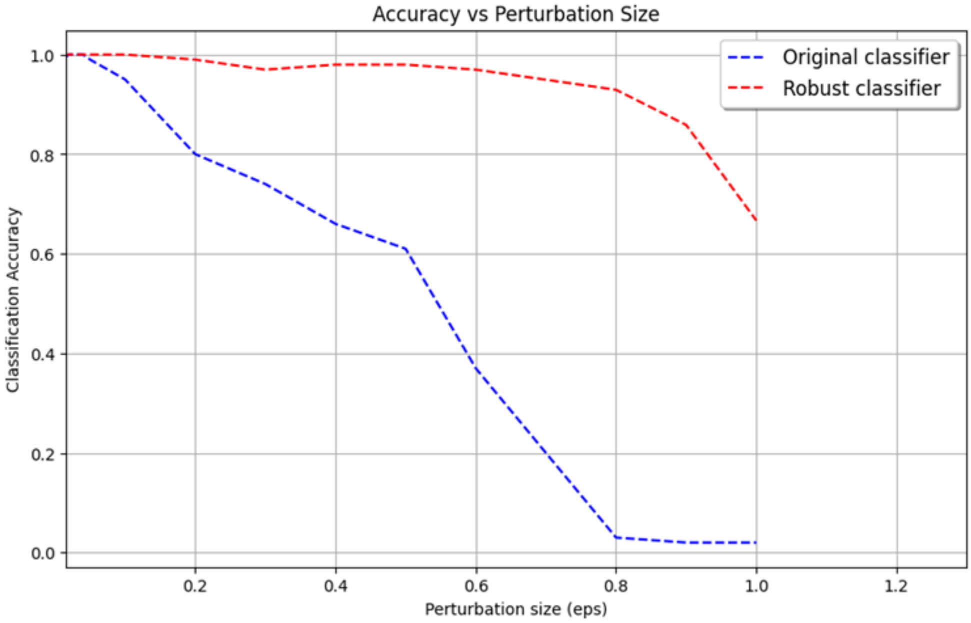


Figure 5. Performance of the original and the robust classifier over a range of `eps`

3.3. Comparison with literature review

We compare our work to the literature review. Table 5 illustrates the comparison between our work and previous studies. This comparison reveals a gap in existing research, as no prior study has assessed the efficacy of adversarial training utilizing PGD attacks in federated learning for intrusion detection in IoT. Our study goes further by comparing the robustness of the baseline federated learning and federated learning model with adversarial training against adversarial evasion attacks, in particular the PGD attacks with a range of perturbation sizes.

Table 5. Literature review of federated IDS in IoT

Paper author	Federated learning	IoT dataset	IDS	Adversarial attack	Defense approach
Shah *et al.* [16]	Yes	No	No	No	Yes
Rashid *et al.* [17]	No	Yes	Yes	Yes	Yes
Ni *et al.* [18]	Yes	No	No	No	No
Ibitoye *et al.* [19]	Yes	No	Yes	No	Yes
Ferrag *et al.* [20]	Yes	Yes	Yes	No	No
Our work	Yes	Yes	Yes	Yes	Yes

Shah *et al.* [16] implemented adversarial training in federated learning but did not assess IoT datasets or the impact of PGD attacks, observing a decline in accuracies compared to centralized training. Rashid *et al.* [17] examined the effects of AML attacks on IoT intrusion detection systems, reporting performance drops of 36% to 78% due to attacks like JSMA and C&W, but their study was limited to centralized models and did not include federated learning. Ni *et al.* [18] introduced a Byzantine-robust federated learning framework without addressing adversarial attacks, while Ibitoye *et al.* [19] developed DiPSeN to enhance adversarial robustness in federated learning without targeting IoT datasets or PGD attacks, focusing instead on privacy preservation. Ferrag *et al.* [20] achieved 93.04% accuracy for IID data and 77.04% for Non-IID data using federated learning in IoT contexts but did not incorporate adversarial attacks or defense strategies. In contrast, our study addresses the gap by applying federated learning to IoT intrusion detection and evaluating PGD attacks with varying perturbation sizes, demonstrating that adversarial training significantly enhances the resilience and robustness of federated learning models against such attacks, confirming its efficacy in securing IoT systems.

However, it is important to address the limitations inherent in our study. The computational overhead associated with adversarial training and the potential impact on model convergence and communication efficiency in federated learning setups are areas that warrant further investigation. Additionally, our results showed that adversarial training improved overall performance but did not uniformly enhance performance across all classes. The confusion matrix indicated a decline in performance for some classes. In summary, our framework demonstrates the efficacy of adversarial training in mitigating adversarial attacks within federated learning environments, particularly for intrusion detection in IoT systems.

4. CONCLUSION

This study assessed the impact of adversarial training on the robustness of federated learning models for intrusion detection, using the Edge-IIoTset dataset. By comparing models under normal training conditions and adversarial training, and evaluating their performance against adversarial attacks, particularly PGD attacks with varying perturbation sizes. We demonstrated that adversarial training significantly enhances the model's resilience to such attacks. Our findings underscore the importance of incorporating adversarial training in federated learning frameworks to improve security and reliability in IoT environments. For future research, we will explore additional strategies to safeguard federated learning models from potential poisoning attacks and other emerging threats. We will investigate diverse defense mechanisms and their integration into federated learning systems to improve the robustness and efficacy of models in real-world applications.

REFERENCES

[1] X. Wu, Z. Jing, and X. Wang, “The security of IoT from the perspective of the observability of complex networks,” *Heliyon*, vol. 10, no. 5, p. e27104, Mar. 2024, doi: 10.1016/j.heliyon.2024.e27104.

[2] T. Sasi, A. H. Lashkari, R. Lu, P. Xiong, and S. Iqbal, “A comprehensive survey on IoT attacks: Taxonomy, detection mechanisms and challenges,” *Journal of Information and Intelligence*, vol. 2, no. 6, pp. 455–513, Nov. 2023, doi: 10.1016/j.jiixd.2023.12.001.

[3] A. Buja, M. Apostolova, and A. Luma, “A model proposal for enhancing cyber security in industrial IoT environments,” *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 36, no. 1, pp. 231–241, Oct. 2024, doi: 10.11591/ijeecs.v36.i1.pp231-241.

[4] S. El Hajla, E. M. Ennaji, Y. Maleh, and S. Mounir, “Enhancing IoT network defense: advanced intrusion detection via ensemble learning techniques,” *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 35, no. 3, pp. 2010–2020, Sep. 2024, doi: 10.11591/ijeecs.v35.i3.pp2010-2020.

[5] S. El Hajla, E. Mahfoud, Y. Maleh, and S. Mounir, “Attack and anomaly detection in IoT Networks using machine learning approaches,” in *Proceedings - 10th International Conference on Wireless Networks and Mobile Communications, WINCOM 2023*, IEEE, Oct. 2023, pp. 1–7. doi: 10.1109/WINCOM59760.2023.10322991.

[6] Y. Maleh, A. Sahid, and M. Belaissaoui, “Optimized machine learning techniques for IoT 6LoWPAN cyber attacks detection,” in *Advances in Intelligent Systems and Computing*, vol. 1383 AISC, 2021, pp. 669–677. doi: 10.1007/978-3-030-73689-7_64.

[7] M. Karanfilovska, T. Kochovska, Z. Todorov, A. Cholakoska, G. Jakimovski, and D. Efnusheva, “Analysis and modelling of a ML-based NIDS for IoT networks,” *Procedia Computer Science*, vol. 204, pp. 187–195, 2022, doi: 10.1016/j.procs.2022.08.023.

[8] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, “Communication-efficient learning of deep networks from decentralized data,” *Journal of the American Statistical Association*, vol. 60, no. 309, 2016, [Online]. Available: http://arxiv.org/abs/1602.05629%0Ahttps://paperswithcode.com/paper/communication-efficient-learning-of-deep

[9] F. M. Ribeiro Junior and C. A. Kamienski, “Federated learning for performance behavior detection in a fog-IoT system,” *Internet of Things (Netherlands)*, vol. 25, p. 101078, Apr. 2024, doi: 10.1016/j.iot.2024.101078.

[10] N. Carlini and D. Wagner, “Towards evaluating the robustness of neural networks,” in *2017 IEEE Symposium on Security and Privacy (SP)*, IEEE, May 2017, pp. 39–57. doi: 10.1109/SP.2017.49.

[11] C. A. Torres, A. A. Orozco, and M. A. Alvarez, “Explaining and harnessing adversarial examples,” in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, 2013, pp. 4330–4333.

[12] A. Kurakin, I. J. Goodfellow, and S. Bengio, “Adversarial examples in the physical world,” in *5th International Conference on Learning Representations, ICLR 2017 - Workshop Track Proceedings*, 2017.

[13] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, “Towards deep learning models resistant to adversarial attacks,” in *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*, 2018.

[14] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, “Learning representations by back-propagating errors,” *Nature*, vol. 323, no. 6088, pp. 533–536, Oct. 1986, doi: 10.1038/323533a0.

[15] A. Raghunathan, J. Steinhardt, and P. Liang, “Certified defenses against adversarial examples,” in *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*, 2018.

[16] D. Shah, P. Dube, S. Chakraborty, and A. Verma, “Adversarial training in communication constrained federated learning.” 2021. [Online]. Available: http://arxiv.org/abs/2103.01319

[17] M. M. Rashid *et al.*, “Adversarial training for deep learning-based cyberattack detection in IoT-based smart city applications,” *Computers & Security*, vol. 120, p. 102783, Sep. 2022, doi: 10.1016/j.cose.2022.102783.

[18] L. Ni, X. Gong, J. Li, Y. Tang, Z. Luan, and J. Zhang, “rFedFW: Secure and trustable aggregation scheme for Byzantine-robust federated learning in Internet of Things,” *Information Sciences*, vol. 653, p. 119784, Jan. 2024, doi: 10.1016/j.ins.2023.119784.

[19] O. Ibitoye, M. O. Shafiq, and A. Matrawy, “Differentially private self-normalizing neural networks for adversarial robustness in federated learning,” *Computers and Security*, vol. 116, p. 102631, May 2022, doi: 10.1016/j.cose.2022.102631.

[20] M. A. Ferrag, O. Friha, D. Hamouda, L. Maglaras, and H. Janicke, “Edge-IIoTset: a new comprehensive realistic cyber security dataset of iot and iiot applications for centralized and federated learning,” *IEEE Access*, vol. 10, pp. 40281–40306, 2022, doi: 10.1109/ACCESS.2022.3165809.

[21] Y. Cao, J. Zhang, Y. Zhao, P. Su, and H. Huang, “SRFL: a secure & amp; robust federated learning framework for iot with trusted execution environments,” *Expert Systems with Applications*, vol. 239, p. 122410, Apr. 2024, doi: 10.1016/j.eswa.2023.122410.

[22] Z. Alebouyeh and A. J. Bidgoly, “Benchmarking robustness and privacy-preserving methods in federated learning,” *Future Generation Computer Systems*, vol. 155, pp. 18–38, Jun. 2024, doi: 10.1016/j.future.2024.01.009.

[23] L. Wang *et al.*, “Progressive defense against adversarial attacks for deep learning as a service in internet of things.” 2020. [Online]. Available: https://arxiv.org/abs/2010.11143

[24] D. J. Beutel *et al.*, “Flower: A Friendly Federated Learning Research Framework.” 2020. [Online]. Available: http://arxiv.org/abs/2007.14390

[25] X. H. Nguyen and K. H. Le, “Robust detection of unknown DoS/DDoS attacks in IoT networks using a hybrid learning model,” *Internet of Things (Netherlands)*, vol. 23, p. 100851, Oct. 2023, doi: 10.1016/j.iot.2023.100851.

[26] M. Conti, N. Dragoni, and V. Lesyk, “A survey of man in the middle attacks,” *IEEE Communications Surveys and Tutorials*, vol. 18, no. 3, pp. 2027–2051, 2016, doi: 10.1109/COMST.2016.2548426.

[27] Y. Zhang, F. Feng, Z. Liao, Z. Li, and S. Yao, “Universal backdoor attack on deep neural networks for malware detection,” *Applied Soft Computing*, vol. 143, p. 110389, Aug. 2023, doi: 10.1016/j.asoc.2023.110389.

[28] F. Pedregosa *et al.*, “Scikit-learn: machine learning in Python,” *Journal of Machine Learning Research*, vol. 12. pp. 2825–2830, 2011. [Online]. Available: http://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html%5Cnhttp://arxiv.org/abs/1201.0490

[29] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: synthetic minority over-sampling technique,” *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, Jun. 2002, doi: 10.1613/jair.953.

[30] A. Athalye, N. Carlini, and D. Wagner, “Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples,” *35th International Conference on Machine Learning, ICML 2018*, vol. 1. pp. 436–448, 2018.

BIOGRAPHIES OF AUTHORS

El Mahfoud Ennaji is Ph.D. student at the University of Sultan Moulay Slimane Beni Mellal, Laboratory of Science and Engineering Technologies at ENSA Khouribga. He is an engineer in network, security and system from National Institute of Post and Telecommunication, Rabat (2016). His research interests are cybersecurity, IoT, and machine learning. He can be contacted at email: elmahfoud.ennaji@usms.ac.ma.

Salah El Hajla is currently a Ph.D. student at the Laboratory of Sciences and Techniques for Engineering at Sultan Moulay Slimane University, Morocco. With a master's degree in distributed computing systems and big data from the Faculty of Science ibn Zohr, Agadir (2019). His research focuses on developing innovative solutions for IoT network security and applying advanced machine learning techniques to enhance cyber defenses. He can be contacted at email: salah.elhajla@usms.ac.ma.

Prof. Dr. Yassine Maleh is an associate professor of cybersecurity and IT governance at Sultan Moulay Slimane University, Morocco, since 2019. He is a double Ph.D. in computer sciences and IT management. He is the founding chair of IEEE Consultant Network Morocco and founding president of the African Research Center of Information Technology & Cybersecurity. He is a senior member of IEEE. He has published over than 140 papers (international journals, book chapters and conferences/workshops), 27 edited books, and 5 authored books. He is the editor-in-chief of the International Journal of Information Security and Privacy. He can be contacted at email: y.maleh@usms.ma.

Prof. Dr. Soufyane Mounir is an associate professor at the National School of Applied Sciences of Sultan Moulay Slimane University, Beni Mellal, Morocco, since 2014. He got his Ph.D. in electronics and telecommunication, from University Hassan 1^{s}t, Morocco. His research is multidisciplinary that focuses on telecommunications, VoIP, signal processing, embedded systems and cyber security. He is an active member of LaSTI Laboratory, ENSA Khouribga. He can be contacted at email: s.mounir@usms.ma.