Federated Learning with Adversarial Optimisation for Secure and Efficient 5G Edge Computing Networks

Saniya Zafar * , Jonathan White , Phil Legg

Posted Date: 4 August 2025

doi: 10.20944/preprints202508.0130.v1

Keywords: adversarial attacks; cybersecurity; federated learning; privacy attacks

 doi:10.20944/preprints202508.0130.v1

Article

Federated Learning with Adversarial Optimisation for Secure and Efficient 5G Edge Computing Networks

Saniya Zafar *, Jonathan White and Phil Legg

Computer Science Research Centre, University of the West of England, Bristol, UK
* Correspondence: saniya.zafar@uwe.ac.uk

Abstract

With the evolution of 5G edge computing networks, privacy-aware applications are gaining significant attention due to their decentralised processing capabilities. However, these networks face substantial challenges to ensure privacy and security, specifically in a Federated Learning (FL) setup, where adversarial attacks can potentially influence the model integrity. Conventional privacy-preserving FL mechanisms are often susceptible to such attacks, leading to degraded model performance and severe security vulnerabilities. To address this issue, this article proposes FL with adversarial optimisation framework to improve adversarial robustness in 5G edge computing networks while ensuring privacy preservation. The proposed framework considers two models; a classifier model and an adversary model, where the classifier model is integrated with adversary model, trained jointly considering Fast Gradient Sign Method (FGSM) for generation of adversarial perturbations. This adversarial optimisation enhances classifier's resilience to attacks, thereby improving both privacy preservation and model accuracy. Experimental analysis reveals that the proposed model achieves up to 99.44% accuracy on adversarial test data, while improving robustness and sustaining high precision and recall across varying client scenarios. The experimental results further ensure the effectiveness of the proposed model in terms of communication efficiency and computational efficiency while reducing inference time and FLOPs making it ideal for secure 5G edge computing applications.

Keywords: adversarial attacks; cybersecurity; federated learning; privacy attacks

1. Introduction

The emergence of 5G wireless networks has significantly transformed computing capabilities by enabling ultra-low latency, high data throughput, and massive device connectivity [1]. One of the significant advancements empowered by 5G is edge computing, which enables computations to be performed at edge users instead of exclusively depending on centralised computing. 5G edge computing is capable of dealing with low-latency applications including smart healthcare, autonomous vehicles, and Industrial Internet-of-Things (IIoT) because it offers reduced bandwidth consumption and lower data transmission delays. However, the increase in the number of distributed devices opens up increased possibilities of privacy issues and security threats, specifically in Federated Learning (FL)-based environments that consider decentralised training across various edge nodes [2].

FL enables edge devices to individually train their local models by maintaining their data privacy while collaboratively training a global model. Despite its privacy-preserving benefits, FL is susceptible to adversarial threats that can compromise both data privacy and model integrity [3]. More precisely, attackers can influence FL models by introducing malicious updates via model poisoning or manipulate training data through backdoor attacks. Moreover, they may infer private information using membership inference attacks and gradient leakage. Such adversarial activities are particularly concerning in 5G edge computing networks due to the heterogeneous nature of edge devices in terms of computational power and security capabilities, which makes them attractive targets for attackers [4].

 doi:10.20944/preprints202508.0130.v1

Conventional privacy-preserving FL algorithms including differential privacy, homomorphic encryption, and secure multi-party computation offer certain levels of privacy. However, they encounter multiple limitations while handling adversarial attacks in the FL environment. The aforementioned traditional mechanisms focus primarily on protecting data privacy, but often become unsuccessful in the detection and mitigation of malicious model updates from compromised edge devices, making them inefficient in combating model poisoning attacks and backdoor attacks [3,5]. Moreover, traditional techniques such as secure aggregation and homomorphic encryption introduce a considerable amount of communication overhead, which restricts their scalability and practicality for 5*G* resource-limited edge devices. Furthermore, FL methods based on differential privacy inject noise to protect data privacy, which hinders the performance of the model and does not efficiently impede membership inference attacks while leaving FL models susceptible to inference and gradient-based attacks [5,6].

To address the aforementioned challenges, adversarial optimisation is an emerging solution to improve the robustness of FL models in 5*G* edge computing. Adversarial optimisation involves the training of FL models to combat adversarial attacks through various techniques including adversarial training, intrusion detection-based filtering of malicious updates, and robust aggregation methods [7]. Moreover, adversarial optimisation allows the development of effective defence mechanisms while ensuring the optimal balance between security, model performance, and efficiency [8,9]. This article provides a novel FL with adversarial optimisation framework for 5*G* edge computing networks while enhancing FL privacy, security, and robustness against adversarial attacks. The key contributions of this work are summarised as follows:

1. This paper proposes a novel FL framework with adversarial optimisation to strengthen the security of 5*G* edge computing networks. By incorporating a classifier model and an adversary model, the proposed algorithm simultaneously improves the robustness of the FL model against adversarial attacks. This ensures more secure and private FL training across edge devices.
2. To train the proposed framework, adversary model considers the Fast Gradient Sign Method (FGSM) for generation of stronger perturbations based on the classifier model's responses in an iterative manner. This guarantees the improvement of model resilience in privacy sensitive 5*G* edge computing networks.
3. For performance evaluation, extensive level simulations have been performed considering the 5*G*- Network Intrusion Detection Dataset (NIDD) [10], which validates the adaptability of the proposed algorithm to 5*G* edge computing networks including 5*G*-enabled IoT. Comprehensive experimental analysis reveals valuable insights of the proposed FL with adversarial optimisation algorithm in terms of accuracy, scalability, computational efficiency, and time efficiency.

The rest of the article is organised as follows. Section 2 presents an overview of the most related contributions from the existing literature. Section 3 provides the details of the problem formulation and the mathematical model. Section 4 elaborates the design of the proposed methodology. Section 5 provides a detailed discussion of the experimental results, and Section 6 concludes the article.

2. Related Work

Several researchers have explored FL for 5*G* edge computing networks to enable privacy-preserving data analysis. For instance, the authors of [11] examined secure aggregation protocol for FL to prevent data leakage during model updates. In [12], the authors introduced a foundational FL framework as federated averaging to efficiently aggregate decentralised model updates. Despite these advancements, conventional FL algorithms remain vulnerable to adversarial attacks, including data poisoning attacks and backdoor attacks which pose significant threats to FL models. [13] demonstrated how malicious clients can inject backdoor triggers into global models without detection. In addition, [14] presented a comprehensive analysis of poisoning attacks while illustrating their detrimental impact on FL models. To counteract such malicious activities, various defensive strategies have been proposed. Differential privacy mechanisms, such as those discussed by the authors of [15] add noise to model updates to obscure sensitive information. However, an unreasonable amount of noise can

 doi:10.20944/preprints202508.0130.v1

degrade the model's accuracy. Similarly, [16] demonstrated the use of secure multi-party computation to ensure privacy, but introduces significant computational overhead. The authors of [17] proposed homomorphic encryption methods to enable secure model aggregation but remain computationally intensive for large-scale FL applications. To address the limitations of existing approaches, recent works have examined adversarial learning techniques. In [18], the authors proposed an adversarial training approach to improve the robustness of the model against inference attacks. In contrast, the authors of [19] provide a comprehensive analysis of adversarial vulnerabilities in FL and presented the decision boundary-based federated adversarial training algorithm to enhance both accuracy and robustness in FL models. Furthermore, [20] provides a comprehensive analysis of current backdoor attack strategies and defences in FL, highlighting challenges and potential future directions. Prior work conducted by our research group explores defending against adversarial machine learning attacks [21,22], as well as data privacy and performance challenges within federated learning environments [23,24]. Inspired by these advancements, we propose a novel FL framework that incorporates these two domains, extending adversarial optimisation to leverage both an adversary model and a classifier model in a two-way manner that can dynamically simulate and mitigate adversarial attacks, leading to a more robust classification model.

3. System Model and Problem Formulation

This section presents the system model and problem formulation to enhance the security in 5G edge computing networks while improving robustness against adversarial attacks. The proposed system model formulates the FL with adversarial optimisation as a min-max optimisation problem. The core idea revolves around the simultaneous training of a robust classifier and an adversary model in a decentralised learning environment. Overall, the classifier aims to minimise the classification loss, whereas the adversary attempts to maximise it by generating strong adversarial perturbations. This adversarial setup enhances the resilience of the classifier against malicious attacks.

In the considered system model, let M represent the number of clients participating in the FL system, where each client $m \in \{1, 2, \ldots, M\}$ possesses a local dataset denoted as:

$$\mathcal{D}_m = \{(\mathbf{x}_{mi}, y_{mi})\}_{i=1}^{n_m}, \tag{1}$$

where $\mathbf{x}_{mi} \in \mathbb{R}^d$ represents the input data, $y_{mi} \in \{1, \ldots, C\}$ is the corresponding class label, and n_m denotes the number of data samples at client m. The total number of data samples across all clients is:

$$N = \sum_{m=1}^{M} n_m. \tag{2}$$

The classifier model is parametrised by θ and is represented as:

$$f_\theta : \mathbb{R}^d \to \mathbb{R}^C. \tag{3}$$

The adversary model, designed to generate adversarial perturbations, is parametrised by ϕ and denoted as:

$$g_\phi : \mathbb{R}^d \to \mathbb{R}^d. \tag{4}$$

The adversary employs FGSM to generate adversarial examples that are aimed at misleading the classifier. The adversarial perturbation is expressed as follows:

$$\mathbf{x}^{adv} = \mathbf{x} + \epsilon \cdot \text{sign}(\nabla_{\mathbf{x}} \mathcal{L}(f_\theta(\mathbf{x}), y)), \tag{5}$$

 doi:10.20944/preprints202508.0130.v1

where: $\mathcal{L}(f_\theta(\mathbf{x}), y)$ represents the loss function, typically cross-entropy loss, ϵ is the perturbation budget, restricting the distortion introduced by the adversary. To ensure the perturbations remain realistic, the generated adversarial examples are constrained by the L_∞ norm:

$$\left\|\mathbf{x}^{adv} - \mathbf{x}\right\|_\infty \leq \epsilon. \tag{6}$$

This ensures that the adversarial examples remain within a feasible range while maintaining perceptual similarity to the original data.

In the considered FL framework, the objective is modelled as a min-max optimisation problem. The classifier model f_θ minimises a joint loss function, while the adversary model g_ϕ maximises the adversarial loss to induce model degradation. This adversarial optimisation objective can be expressed as:

$$\begin{aligned}\min_\theta \max_\phi \frac{1}{M}\sum_{m=1}^{M} \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}_m}[\lambda\mathcal{L}(f_\theta(\mathbf{x}), y) \\ + (1-\lambda)\mathcal{L}\big(f_\theta\big(g_\phi(\mathbf{x})\big), y\big)],\end{aligned} \tag{7}$$

$\lambda \in [0,1]$ is a trade-off parameter that balances the clean loss and adversarial loss, $\mathcal{L}(f_\theta(\mathbf{x}), y)$ is the clean classification loss, $\mathcal{L}\big(f_\theta\big(g_\phi(\mathbf{x})\big), y\big)$ is the adversarial classification loss. Each client performs local training by optimising its own classifier objective using stochastic gradient descent. The local classifier objective for each client m is given by:

$$\min_{\theta_m} \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}_m}\big[\lambda\mathcal{L}(f_{\theta_m}(\mathbf{x}), y) + (1-\lambda)\mathcal{L}\big(f_{\theta_m}\big(g_\phi(\mathbf{x})\big), y\big)\big]. \tag{8}$$

The model parameters are updated using gradient-based optimisation:

$$\theta_m^{t+1} = \theta_m^t - \eta\nabla_{\theta_m}\mathcal{L}_m, \tag{9}$$

where: $\mathcal{L}_m$ is the local loss function, and η is the learning rate. After a specified number of local training epochs, the updated model parameters are transmitted to the central server. The server aggregates these models using the federated averaging algorithm to obtain the global model:

$$\theta^{t+1} = \sum_{m=1}^{M} \frac{n_m}{N}\theta_m^t. \tag{10}$$

This aggregated model serves as the updated global model for the next communication round.

Overall, to maintain effective training dynamics, several constraints and regularisation mechanisms are incorporated. The adversarial perturbations generated by adversary model g_ϕ are strictly bounded by the L_∞ norm:

$$\left\|g_\phi(\mathbf{x}) - \mathbf{x}\right\|_\infty \leq \epsilon. \tag{11}$$

Furthermore, gradient clipping is applied to prevent gradient explosion during training, ensuring that the gradients are within a manageable range:

$$\|\nabla_\theta\mathcal{L}\| \leq G_{\max}, \tag{12}$$

where $G_{\max}$ is a predefined gradient threshold. Additionally, a L_2 regularisation term is introduced to mitigate overfitting and encourage generalisation:

$$\mathcal{L}_{reg}(\theta) = \frac{\lambda_r}{2}\|\theta\|^2, \tag{13}$$

 doi:10.20944/preprints202508.0130.v1

where $\lambda_r \geq 0$ controls the strength of regularisation. The final mathematical formulation of the FL with adversarial optimisation can be represented as:

$$\min_{\theta} \max_{\phi} \sum_{m=1}^{M} \frac{n_m}{N} \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}_m} \Big[\lambda \mathcal{L}(f_\theta(\mathbf{x}), y) + (1-\lambda)\mathcal{L}\big(f_\theta(g_\phi(\mathbf{x})), y\big) + \frac{\lambda_r}{2}\|\theta\|^2\Big] \tag{14a}$$

$$\text{subject to:} \quad \big\|g_\phi(\mathbf{x}) - \mathbf{x}\big\|_\infty \leq \epsilon. \tag{14b}$$

The formulated optimisation problem encompasses the adversarial interaction between the classifier model and the adversary model to ensure the robustness of the FL model. The FL system effectively strengthens its resilience to malicious activities by jointly optimising both clean and adversarial losses while constraining adversarial perturbations. The classifier achieves robust generalisation even under adversarial threats through iterative local training and global aggregation.

4. Proposed Federated Learning with Adversarial Optimisation Algorithm

In this section, we provide the overall methodology of the proposed framework. This section provides the dataset description and pre-processing of the dataset followed by training algorithm of the proposed FL with adversarial optimisation model for secure and robust 5G edge computing networks. Figure 1 illustrates the overall framework of the proposed model.

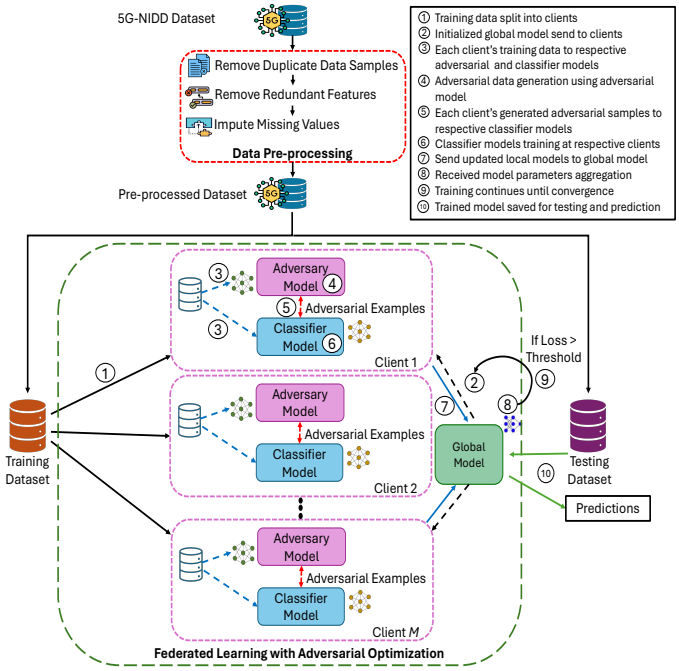


Figure 1. FL with adversarial optimisation workflow.

4.1. Dataset Description and Pre-Processing

In the proposed FL with adversarial optimisation mechanism, we considered the latest and most realistic 5G-NIDD intrusion detection dataset for 5G wireless networks [10]. 5G-NIDD is a fully labelled dataset from 5G testbed at the University of Oulu, Finland. 5G-NIDD features data from base stations, including attack scenarios such as port scans (UDP Scan, SYN Scan, TCP Connect) and Denial-of-Service (Slowrate, ICMP, SYN, UDP, HTTP Floods). Overall, the dataset is comprised of 9 classes with 8 attack classes and a benign class. The class distribution of 5G-NIDD dataset is presented in Figure 2. Moreover, the dataset contains 52 columns. Pre-processing the dataset plays a fundamental role in training the machine and deep learning models. For this, we preprocessed the

 doi:10.20944/preprints202508.0130.v1

dataset by removing one duplicated row i.e., data sample of the dataset. Subsequently, the redundant feature columns with 80% to 99% missing values are eliminated which do not contribute significantly to training the network, resulting in a new dataset with 36 columns including 35 feature columns and a label column. After that, the remaining features with missing values within the acceptable range are identified. The identified feature columns are sTos, sDSb, sTtl, and sHops, each having 214 missing values, which is only 0.0176% of the total feature values, are imputed based on the respective feature distribution.

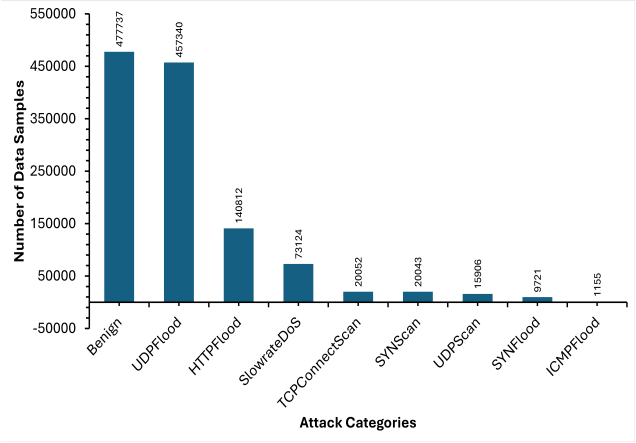


Figure 2. Overview of the dataset distribution.

4.2. Training of Federated Learning with Adversarial Optimisation Model

The training algorithm for FL with adversarial optimisation is provided in Algorithm 1. The training process begins with the server initialising the parameters of both the classifier model and the adversary model. These parameters are then distributed to all participating clients. The clients receive the initial classifier parameters, denoted as θ^t, and the adversary parameters, ϕ^t. At each global communication round t, the server synchronises both models across all clients by broadcasting the updated parameters.

Once the models are initialised, each client proceeds with its local training using its respective dataset. For every local training epoch, the client divides its dataset into mini-batches and, for each batch, generates adversarial examples using FGSM. This is achieved by perturbing the original input data in the direction of the gradient of the classifier's loss with respect to the input. The perturbation is controlled by a predefined parameter ϵ, which determines the strength of the adversarial attack. The adversarial input $\mathbf{x}^{adv}$ is computed as follows:

$$\mathbf{x}^{adv} = \mathbf{x} + \epsilon \cdot \text{sign}\left(\nabla_{\mathbf{x}} \mathcal{L}\left(f_{\theta_m^t}(\mathbf{x}), y\right)\right), \tag{15}$$

where $\mathcal{L}$ is the loss function, $f_{\theta_m^t}$ represents the classifier, and y denotes the true label. Next, the client calculates both the clean loss and the adversarial loss. The clean loss measures the classifier's performance on the original data, while the adversarial loss evaluates its performance on the adversarially perturbed data. The combined loss function, which is a weighted sum of the clean and adversarial losses is then used to update the classifier model. A regularisation term is also included to mitigate overfitting, resulting in the following total loss:

$$\mathcal{L}_{\text{total}} = \lambda \mathcal{L}_{\text{clean}} + (1-\lambda)\mathcal{L}_{\text{adv}} + \frac{\lambda_r}{2}\left\|\theta_m^t\right\|^2, \tag{16}$$

where λ controls the balance between clean and adversarial training, and λ_r is the regularisation coefficient.

 doi:10.20944/preprints202508.0130.v1

The adversary model is updated using gradient ascent to maximise adversarial loss. This ensures the generation of more effective adversarial examples, further challenging the classifier. The update rule for the adversary is given as follows:

$$\phi_m^{t+1} = \phi_m^t + \eta_\phi \nabla_\phi \mathcal{L}_{\text{adv}}, \tag{17}$$

where η_ϕ is the learning rate for the adversary. Conversely, the classifier is updated using gradient descent to minimise the total loss:

$$\theta_m^{t+1} = \theta_m^t - \eta_\theta \nabla_\theta \mathcal{L}_{\text{total}}\,, \tag{18}$$

where η_θ is the classifier's learning rate. After completing the local training for the specified number of epochs, each client sends its updated classifier and adversary parameters back to the server. The server then aggregates the model updates using federated averaging. This involves computing a weighted average of the client models, where the weights are proportional to the number of data samples held by each client. The global classifier and adversary parameters are updated as follows:

$$\theta^{t+1} = \sum_{m=1}^{M} \frac{n_m}{N} \theta_m^t, \tag{19}$$

$$\phi^{t+1} = \sum_{m=1}^{M} \frac{n_m}{N} \phi_m^t, \tag{20}$$

where M is the total number of clients, n_m is the number of samples at client m, and $N = \sum_{m=1}^{M} n_m$ is the total number of samples across all clients.

This process of local training followed by model aggregation is repeated for a specified number of global rounds. At the end of the training, the server returns the final classifier and adversary model parameters. This approach not only ensures the classifier's robustness against adversarial attacks but also maintains data privacy by keeping data decentralised between clients.

4.3. Testing of the Trained Federated Learning with Adversarial Optimisation Model

The trained FL with adversarial optimisation model is evaluated for clean test data as well as adversarial test data to observe the overall model performance in terms of robustness against adversarial attacks for 5G edge computing networks.

 doi:10.20944/preprints202508.0130.v1

Algorithm 1: Federated Learning with Adversarial Optimisation

Input: T: global rounds, M: number of clients, E: local epochs, B: batch size
Input: η_θ, η_ϕ: learning rates, ϵ: perturbation budget
Result: Final global classifier model f_θ^T and adversary model g_ϕ^T
Initialise global classifier model f_θ and adversary model g_ϕ ;
for $t \leftarrow 1$ **to** T **do**
  Server: Broadcast classifier parameters θ^t and adversary parameters ϕ^t to all clients;
  foreach *client* $m \in \{1, \ldots, M\}$ *in parallel* **do**
    Initialise local classifier $\theta_m^t \leftarrow \theta^t$, and local adversary $\phi_m^t \leftarrow \phi^t$;
    foreach *local epoch* $e \leftarrow 1$ **to** E **do**
      foreach *batch* $\mathcal{B} \subseteq \mathcal{D}_m$ *of size* B **do**
        Adversarial Example Generation (FGSM):

$$\boldsymbol{x}^{adv} = \boldsymbol{x} + \epsilon \cdot \text{sign}\Big(\nabla_{\boldsymbol{x}} \mathcal{L}(f_{\theta_m^t}(\boldsymbol{x}), y)\Big)$$

        Compute Loss:

$$\mathcal{L}_{\text{clean}} = \mathcal{L}(f_{\theta_m^t}(\boldsymbol{x}), y)$$

$$\mathcal{L}_{\text{adv}} = \mathcal{L}(f_{\theta_m^t}(\boldsymbol{x}^{adv}), y)$$

$$\mathcal{L}_{\text{total}} = \lambda \mathcal{L}_{\text{clean}} + (1-\lambda)\mathcal{L}_{\text{adv}} + \frac{\lambda_r}{2}\|\theta_m^t\|^2$$

        Adversary Update:

$$\phi_m^{t+1} = \phi_m^t + \eta_\phi \nabla_\phi \mathcal{L}_{\text{adv}}$$

        Classifier Update:

$$\theta_m^{t+1} = \theta_m^t - \eta_\theta \nabla_\theta \mathcal{L}_{\text{total}}$$

      end
    end
    Send updated θ_m^t, ϕ_m^t to server;
  end
  Server: Perform federated averaging for classifier and adversary models:

$$\theta^{t+1} = \sum_{m=1}^{M} \frac{n_m}{N} \theta_m^t, \quad \phi^{t+1} = \sum_{m=1}^{M} \frac{n_m}{N} \phi_m^t$$

end
Return: Final global classifier parameters θ^T and adversary parameters ϕ^T

5. Performance Evaluation

This section presents the performance evaluation of the proposed FL with adversarial optimisation model while providing a comparison of its results with the standard FL algorithm. All the simulations are carried out using the TensorFlow and Scikit-Learn libraries on the Google Compute Engine, which provides an Nvidia Tesla T4 Graphics Processing Unit (GPU) with high RAM for smooth execution of deep learning algorithms.

5.1. Implementation of Federated Learning with Adversarial Optimisation Model

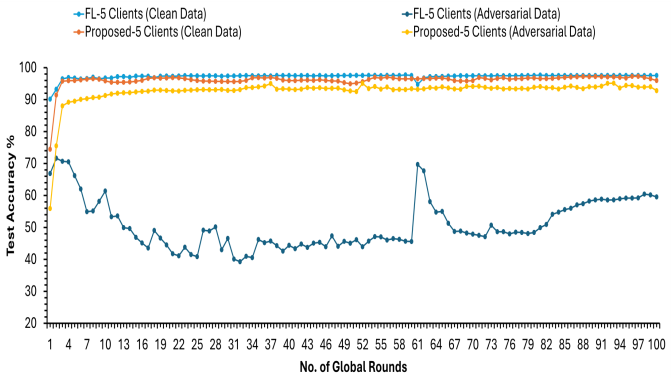
The 5G-NIDD dataset is divided into training dataset and testing dataset as 80 : 20 i.e., 80% training data and 20% testing data. To train the FL with adversarial optimisation model to secure 5G edge computing network, training dataset with its corresponding labels is given as an input to train the proposed model. In the FL setup, the data is Independently and Identically Distributed

 doi:10.20944/preprints202508.0130.v1

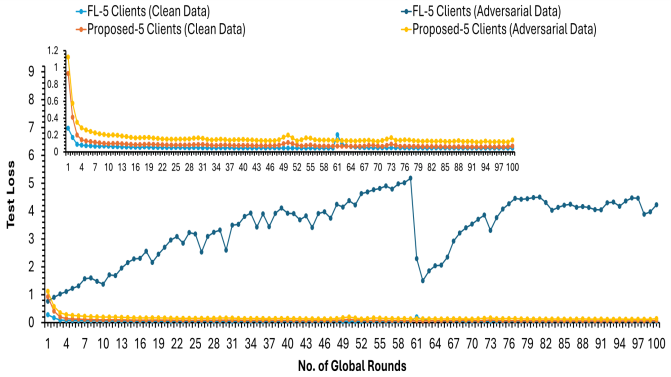
(IID) distributed among the participating clients to ensure balanced local training. Unlike standard FL, the proposed approach incorporates both a classifier model and an adversary model, which are trained in tandem to enhance robustness against adversarial attacks. The classifier is a dense neural network consisting of four fully connected layers with ReLU activation functions. It has an input layer, followed by hidden layers with 128, 64, and 32 neurons, respectively. Dropout layers with a rate of 0.2 are applied after the first two hidden layers to prevent overfitting. The output layer uses a softmax activation function to predict the class probabilities. The adversary model is designed using the FGSM to generate adversarial perturbations. It takes the classifier and the input data as input, computes the gradient of the loss with respect to the input using a gradient tape, and applies perturbations based on the sign of the gradients. The perturbations are scaled using a fixed epsilon value of 0.1 to maximise the classifier's loss, effectively simulating adversarial attacks. The adversary model is trained to iteratively generate more robust adversarial examples by refining its perturbations with each iteration, considering the classifier's response to previous attacks. This training strategy forces the adversary to adapt and optimise its perturbations, effectively challenging the classifier and enhancing the model's resilience against a range of adversarial scenarios. FL is performed with various number of clients scenarios including $5, 10, 20, 30$, and 40 clients scenario, each client in respective case trains its local model for 5 epochs per global round, using a mini-batch size of 128. The global model for each scenario is trained over 100 communication rounds, where local model updates are aggregated using weighted averaging. The learning rate is initialised at 0.001 and follows an exponential decay schedule with a decay factor of 0.95. Both the classifier and adversary models are optimised using the Adam optimiser with a weight decay of 1×10^{-5}. Performance is evaluated using metrics such as accuracy, precision, recall, and F1-score for clean inputs, while adversarial accuracy is measured using adversarial examples generated from the adversary model with perturbation strength 0.1. The proposed adversarial FL approach ensures robust intrusion detection by mitigating the impact of adversarial attacks in FL environments.

The testing dataset and adversarial dataset are used to test the trained FL with adversarial optimisation model. Figure 3 presents the convergence analysis of the standard FL model and the proposed FL with adversarial optimisation model under both the clean test data and adversarial test data. It is evident that the overall performance of the proposed model is not compromised as the global test accuracy remains constantly high. Moreover, the test loss of the proposed approach is significantly lower in comparison with the standard FL. For instance, in Figure 3(**d**), the test loss of the simple FL model with 40 clients reaches 19.9694 on adversarial test data, whereas the proposed framework preserves the lower test loss of 0.0124, indicating enhanced learning capability and greater stability. The effectiveness and generalisation of the model is also evident from its persistent performance over various rounds of training. The performance of the standard FL fluctuates under adversarial attacks over increased rounds of training, while the proposed model maintains stability in terms of accuracy and loss as observed in Figure 3(**a**) - 3(**d**). This illustrates that the proposed framework helps to restrain the accumulation of errors, which commonly affects the standard FL under adversarial attacks.

 doi:10.20944/preprints202508.0130.v1

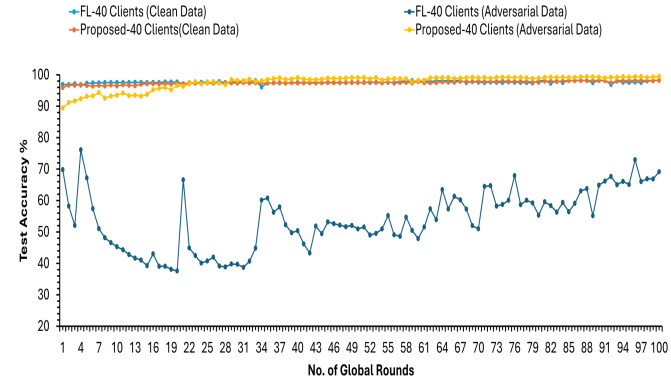


(**a**) Testing accuracy of models with 5 clients.

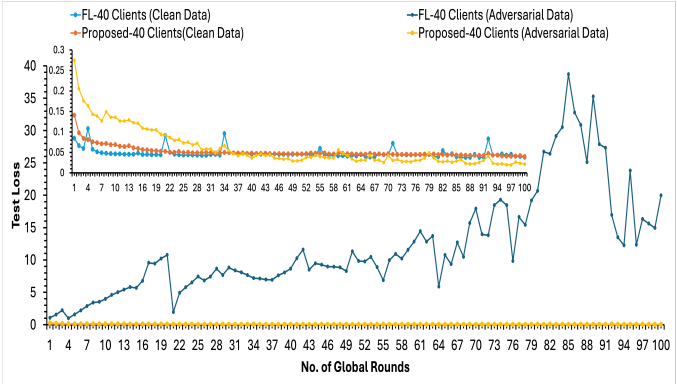


(**b**) Testing loss of models with 5 clients (showing zoomed in view).

.



(**c**) Testing accuracy of models with 40 clients.



(**d**) Testing loss of models with 40 clients (showing zoomed in view).

Figure 3. Convergence analysis of models.

 doi:10.20944/preprints202508.0130.v1

Figure 4 illustrates the robustness and scalability of the proposed model against adversarial attacks. It is observed that the existence of the adversarial data intensively degrades the performance of the simple FL model, while our approach with adversarial optimisation maintains significantly higher accuracy rate. For instance, in models trained with 5 clients, the accuracy rate of simple FL drops to 59.54% on adversarial data, whereas the FL model with adversarial optimisation retains higher accuracy rate of 92.79%. This trend is followed in different client configurations, with the adversarial optimisation-based model attaining the accuracy of 99.44% as compared to standard FL model with accuracy of 69.13% for 40 clients scenario. Overall, the experimental analysis depicts significant rise in model performance on adversarial data in comparison with standard FL model while providing comparable performance on clean data.

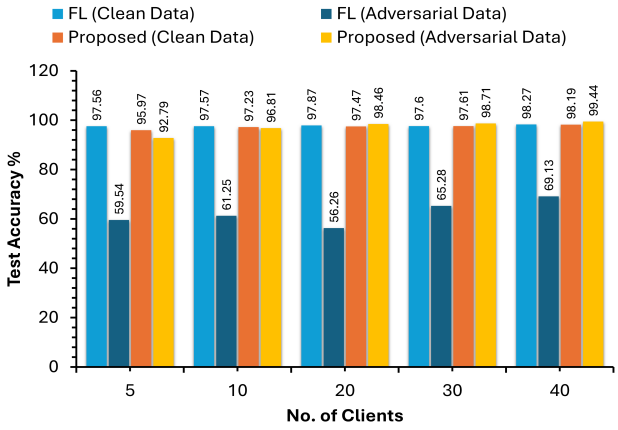


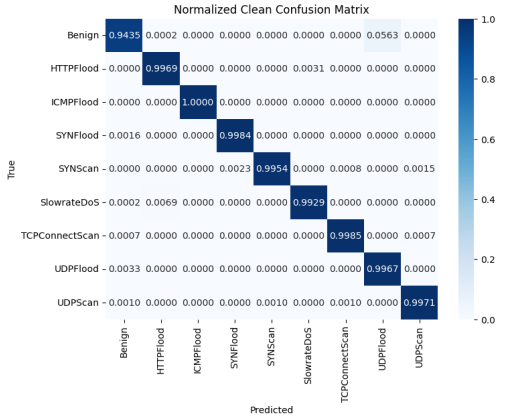
Figure 4. Test accuracy comparison of models with various no. of clients.

The robustness of the proposed FL with adversarial optimisation mechanism is further highlighted by the enhancement in precision, recall, and F1-score, as presented in Table 1. For example, in 40 clients case, it is notable that the simple FL achieves an F1-score of 70% only when compared to the FL with adversarial optimisation which provides an impressive F1-score of 99%. This reveals that the proposed method not only accurately classifies adversarial data but also manages to preserve the balance among precision and recall, resulting in fewer misclassification. The confusion matrices provided in Figure 5 and Figure 6 further signifies the reduction of misclassification on adversarial data in case of FL with adversarial optimisation when compared to simple FL model.

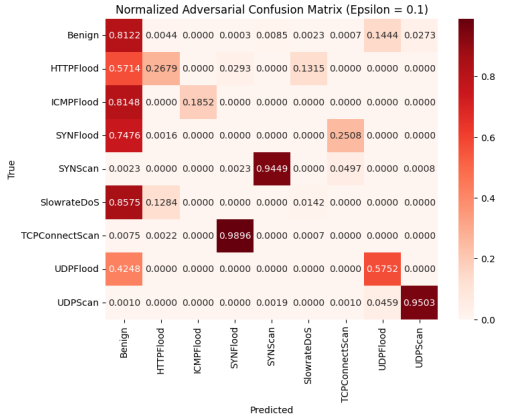
Table 1. Performance comparison of models with various no. of clients.

Clients	Strategy	Clean test data			Adversarial test data		
		Precision	Recall	F1	Precision	Recall	F1
5	**FL**	98%	98%	98%	62%	60%	57%
	Proposed	96%	96%	96%	93%	93%	93%
10	**FL**	98%	98%	98%	67%	61%	63%
	Proposed	97%	97%	97%	97%	97%	97%
20	**FL**	98%	98%	98%	60%	56%	56%
	Proposed	97%	97%	97%	98%	98%	98%
30	**FL**	98%	98%	98%	66%	65%	64%
	Proposed	98%	98%	98%	99%	99%	99%
40	**FL**	98%	98%	98%	72%	69%	70%
	Proposed	98%	98%	98%	99%	99%	99%

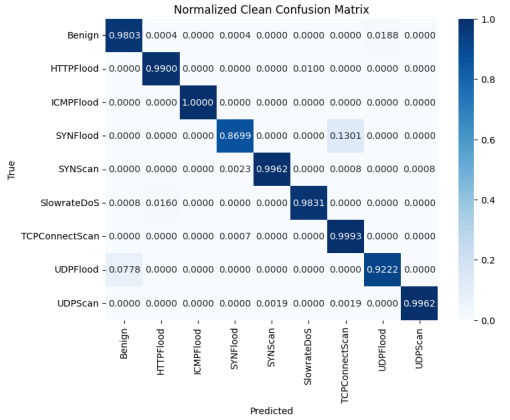
 doi:10.20944/preprints202508.0130.v1



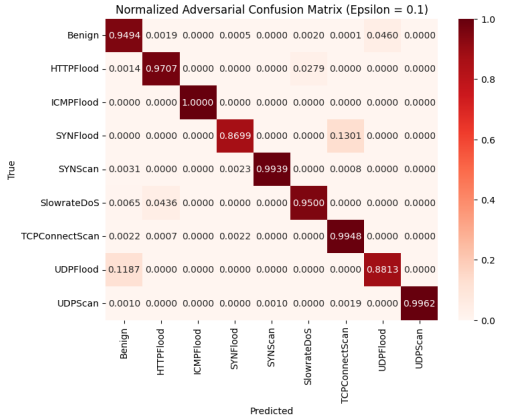
(**a**) Confusion matrix of the FL model with clean test data.



(**b**) Confusion matrix of the standard FL model with adversarial test data.



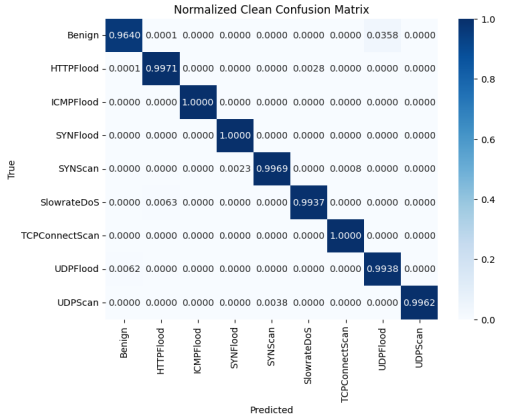
(**c**) Confusion matrix of the proposed model with clean test data.



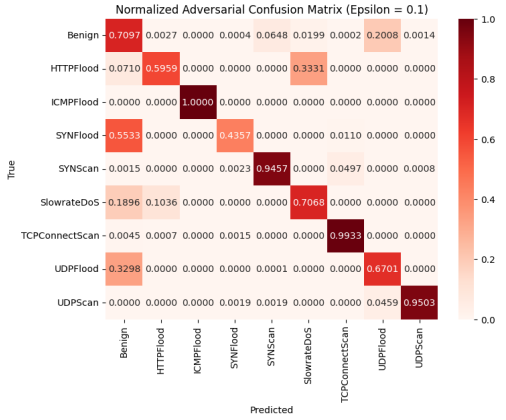
(**d**) Confusion matrix of the proposed model with adversarial test data.

Figure 5. Confusion matrices for models considering 5 clients scenario.

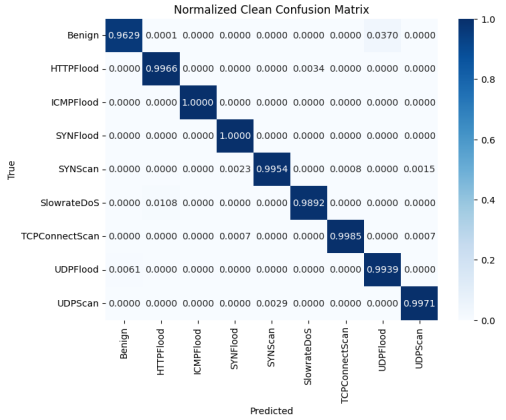
 doi:10.20944/preprints202508.0130.v1



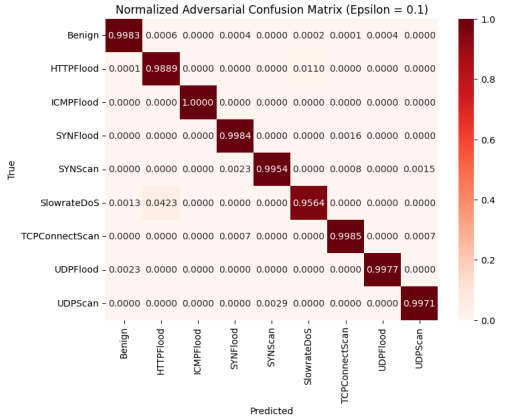
(**a**) Confusion matrix of the FL model with clean test data.



(**b**) Confusion matrix of the standard FL model with adversarial test data.



(**c**) Confusion matrix of the proposed model with clean test data.



(**d**) Confusion matrix of the proposed model with adversarial test data.

Figure 6. Confusion matrices for models considering 40 clients scenario.

To evaluate the classification performance of the proposed algorithm, confusion matrices were generated considering both clean test data and adversarial test data for varying number of clients including 5 clients scenario and 40 clients scenario as presented in Figure 5 and Figure 6 respectively. For the standard FL model with 5 clients, the confusion matrix on clean test data indicates relatively high accuracy as shown in Figure 5(**a**). However, a sharp decline in performance is observed under adversarial conditions with a noticeable increase in false positives and false negatives as displayed in Figure 5(**b**). In contrast, Figure 5(**c**) and Figure 5(**d**) demonstrate that the proposed FL with adversarial optimisation model maintains robust classification performance across both clean test data and adversarial test data highlighting the effectiveness of adversarial optimisation in enhancing resilience to perturbations. When the models are scaled to 40 clients, the standard FL model again performs well on clean test data as shown in Figure 6(**a**) but suffers significantly under adversarial attacks where the misclassification rate dominates the confusion matrix as depicted in Figure 6(**b**). On the contrary, the proposed FL with adversarial optimisation model continues to exhibit strong performance, exhibiting fewer misclassifications and higher consistency in both clean test scenario and adversarial test scenario as illustrated in Figure 6(**c**) and Figure 6(**d**). This emphasizes the robustness and scalability of the proposed model while highlighting its suitability for large-scale deployment in federated edge networks exposed to adversarial threats.

Figure 7 and Table 2 present the sensitivity analysis to assess the performance and reliability of our proposed FL architecture integrated with adversarial optimisation. The detailed sensitivity

 doi:10.20944/preprints202508.0130.v1

analysis involved the performance evaluation of both the standard FL model and the proposed FL with adversarial optimisation model against adversarial test datasets generated using varying perturbation strengths $0.01, 0.05, 0.1, 0.2$, and 0.3. Both the models were trained considering 40 clients scenario, where the proposed model incorporated adversarial training using perturbation strength 0.1 while the standard FL model did not consider any adversarial robustness mechanisms. In Figure 7, it is seen that the proposed model maintains significantly higher accuracy rate and resilience with the increase in perturbation strengths as compared to the achievable accuracy rate of the standard FL model. It is to be noted that the proposed model provides the maximum achievable accuracy of 99.44% when tested on adversarial testing data with perturbation strength 0.1. This is because the proposed model is trained considering the perturbation strength 0.1. Overall, the accuracy of the standard FL model decreases drastically which indicates vulnerability to stronger attacks, whereas the proposed FL with adversarial optimisation model exhibits only a gradual decline while maintaining significantly higher accuracy even at $\epsilon = 0.3$. This stable degradation pattern of achievable accuracy under severe adversarial conditions demonstrates the resilience of the proposed model and underscores the need for adversarial optimisation in privacy preserving FL frameworks. In Table 2, we observed that the performance of both models is comparable at $\epsilon = 0.01$ with F1-score of 93% and F1-score of 94% for the standard FL model and the proposed FL with adversarial optimisation model respectively. However, the performance of standard FL model degraded rapidly, dropping to F1-score of 70% at $\epsilon = 0.1$ and further reduced to 46% at $\epsilon = 0.3$. On the contrary, the proposed FL with adversarial optimisation model consistently outperformed its counterpart while achieving a remarkable F1-score of 99% at $\epsilon = 0.1$ and sustaining robustness even at higher perturbation strengths with F1-score of 80% and 54% at $\epsilon = 0.2$ and $\epsilon = 0.3$ respectively. sensitivity analysis underscores the efficiency of the proposed FL with adversarial optimisation model in improving the resilience of FL models, specifically under stronger and varying adversarial threats. The performance gap between standard FL and proposed FL model highlights the need to integrate adversarial robustness mechanisms in federated environments, particularly in security critical applications like 5G edge computing networks and IoT.

Table 2. Performance comparison of trained models with various perturbation strengths.

Perturbation	Strategy	Precision	Recall	F1
ϵ = 0.01	**FL**	93%	93%	93%
	Proposed	94%	94%	94%
ϵ = 0.05	**FL**	78%	77%	78%
	Proposed	97%	97%	97%
ϵ = 0.1	**FL**	72%	69%	70%
	Proposed	99%	99%	99%
ϵ = 0.2	**FL**	54%	51%	52%
	Proposed	80%	81%	80%
ϵ = 0.3	**FL**	48%	45%	46%
	Proposed	61%	57%	54%

 doi:10.20944/preprints202508.0130.v1

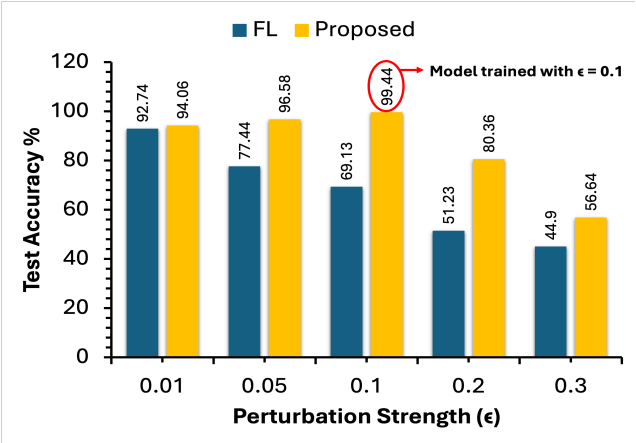


Figure 7. Test accuracy comparison of models with various perturbation strengths.

Integrating adversarial optimisation does not compromise communication efficiency. Regardless of the escalation in the number of clients, communication overhead remains comparatively stable in both the standard FL, and FL with adversarial optimisation approaches. Moreover, the proposed model provides a slight reduction in communication overhead as compared to simple FL as presented in Figure 8. For instance, the communication overhead for simple FL in case of 40 clients is 590.12 MB, while the suggested algorithm shows slight optimisation in this with 578.4 MB communication overhead. These findings suggest that the FL with adversarial optimisation is feasible for large-scaled FL applications without introducing additional communication cost.

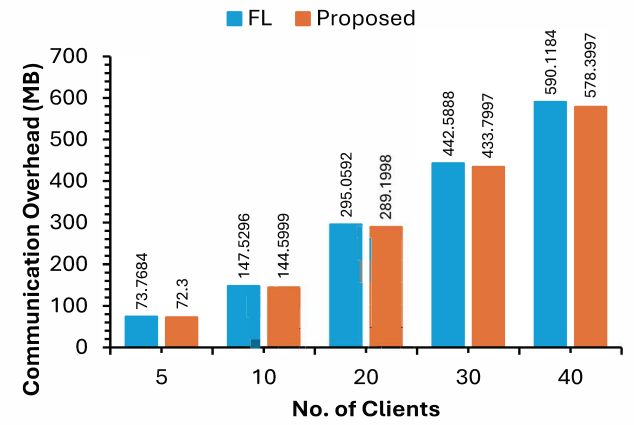


Figure 8. Communication efficiency comparison of models with various no. of clients.

The computational efficiency is also an important factor while dealing with real-time applications. Therefore, computational efficiency and time efficiency are evaluated. The proposed FL with adversarial optimisation model demonstrates an improvement in time efficiency. To be more precise, the proposed mechanism maintains reduced average inference time per data sample of 64.87 msec as compared to standard FL model with average inference time per sample of 69.18 msec. Furthermore, the computational cost of the suggested FL model is also lower as compared to standard FL model in terms of Floating-point Operations Per Second (FLOPS) per inference. Specifically, the simple FL model demands 38,470 FLOPs per inference, whereas the proposed model requires 37,718 FLOPs indicating reduced computational complexity.

Overall, it is to be noted that the proposed FL with adversarial optimisation model strengthens privacy and security while ensuring enhancement in robustness against adversarial attacks, scalability, as well as computational efficiency.

 doi:10.20944/preprints202508.0130.v1

6. Conclusion

This article presents FL with adversarial optimisation framework to enhance security and adversarial robustness in 5G edge computing networks. The proposed mechanism incorporates a classifier model and an adversary model to improve the resilience of FL against malicious activities. Adversary model used FGSM for producing adversarial perturbations to challenge the classifier model and enhance its robustness against adversarial attacks. Comprehensive and real-world 5G network dataset namely, the 5G-NIDD dataset is utilised to train and evaluate the presented model. To evaluate the effectiveness and scalability of the proposed algorithm, we conducted experimental analysis across various client scenarios, including 5, 10, 20, 30, and 40 clients. Experimental results demonstrate that the presented approach significantly improves the achievable adversarial accuracy of 99.44% with 40 clients while preserving competitive clean accuracy. Moreover, the proposed model demonstrates communication efficiency through reduced overhead and computational efficiency in terms of reduced inference time in comparison with conventional FL models. In the future, more sophisticated adversarial methods including Projected Gradient Descent (PGD), Carlini & Wagner (C&W), and DeepFool attacks will be explored to optimise the adversary model for further enhancement of model's robustness against ever evolving adversarial attacks. Additionally, non-IID data distributions will also be considered to evaluate the performance of the proposed model to better simulate the real-world scenario.

Author Contributions: Conceptualization, S.Z., J.W. and P.L.; methodology, S.Z., J.W. and P.L.; software, S.Z.; validation, S.Z., J.W. and P.L.; formal analysis, S.Z., J.W. and P.L.; investigation, S.Z.; resources, S.Z. and P.L.; data curation, S.Z., J.W., P.L.; writing—original draft preparation, S.Z.; writing—review and editing, J.W. and P.L.; visualization, S.Z., J.W., and P.L.; supervision, P.L.; project administration, P.L.; funding acquisition, P.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the College of Arts, Technology and Environment at the University of the West of England.

Institutional Review Board Statement: Not Applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All experimental data can be made available on request from the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

5G-NIDD	5G - Network Intrusion Detection Dataset
FGSM	Fast Gradient Sign Method
FL	Federated Learning
FLOPS	Floating-point Operations Per Second
GPU	Graphics Processing Unit
IID	Independent and Identically Distributed
IIoT	Industrial Internet of Things
PGD	Projected Gradient Descent
C&W	Carlini & Wagner

References

1. Hassan, N.; Yau, K.L.A.; Wu, C. Edge computing in 5G: A review. *IEEE Access* **2019**, *7*, 127276–127289.
2. Lee, J.; Solat, F.; Kim, T.Y.; Poor, H.V. Federated learning-empowered mobile network management for 5G and beyond networks: From access to core. *IEEE Communications Surveys & Tutorials* **2024**.

 doi:10.20944/preprints202508.0130.v1

3. Nowroozi, E.; Haider, I.; Taheri, R.; Conti, M. Federated learning under attack: Exposing vulnerabilities through data poisoning attacks in computer networks. *IEEE Transactions on Network and Service Management* **2025**.
4. Han, G.; Ma, W.; Zhang, Y.; Liu, Y.; Liu, S. BSFL: A blockchain-oriented secure federated learning scheme for 5G. *Journal of Information Security and Applications* **2025**, p. 103983.
5. Feng, Y.; Guo, Y.; Hou, Y.; Wu, Y.; Lao, M.; Yu, T.; Liu, G. A survey of security threats in federated learning. *Complex & Intelligent Systems* **2025**, *11*, 1–26.
6. Rao, B.; Zhang, J.; Wu, D.; Zhu, C.; Sun, X.; Chen, B. Privacy inference attack and defense in centralized and federated learning: A comprehensive survey. *IEEE Transactions on Artificial Intelligence* **2024**.
7. Tahanian, E.; Amouei, M.; Fateh, H.; Rezvani, M. A Game-theoretic Approach for Robust Federated Learning. *International Journal of Engineering, Transactions A: Basics* **2021**, *34*, 832–842.
8. Guo, Y.; Qin, Z.; Tao, X.; Dobre, O.A. Federated Generative-Adversarial-Network-Enabled Channel Estimation. *Intelligent Computing* **2024**, *3*, 0066.
9. Grierson, S.; Thomson, C.; Papadopoulos, P.; Buchanan, B. Min-max training: Adversarially robust learning models for network intrusion detection systems. In Proceedings of the 2021 14th International Conference on Security of Information and Networks (SIN). IEEE, 2021, Vol. 1, pp. 1–8.
10. Samarakoon, S.; Siriwardhana, Y.; Porambage, P.; Liyanage, M.; Chang, S.Y.; Kim, J.; Kim, J.; Ylianttila, M. 5G-NIDD: A Comprehensive Network Intrusion Detection Dataset Generated over 5G Wireless Network, 2022. Dataset, https://doi.org/10.21227/xtep-hv36.
11. Bonawitz, K.; Eichner, H.; Grieskamp, W.; Huba, D.; Ingerman, A.; Ivanov, V.; Kiddon, C.; Konečnỳ, J.; Mazzocchi, S.; McMahan, B.; et al. Towards federated learning at scale: System design. *Proceedings of machine learning and systems* **2019**, *1*, 374–388.
12. McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; y Arcas, B.A. Communication-efficient learning of deep networks from decentralized data. In Proceedings of the Artificial intelligence and statistics. PMLR, 2017, pp. 1273–1282.
13. Bagdasaryan, E.; Veit, A.; Hua, Y.; Estrin, D.; Shmatikov, V. How to backdoor federated learning. In Proceedings of the International conference on artificial intelligence and statistics. PMLR, 2020, pp. 2938–2948.
14. Lyu, L.; Yu, H.; Yang, Q. Threats to federated learning: A survey. *arXiv preprint arXiv:2003.02133* **2020**.
15. Abadi, M.; Chu, A.; Goodfellow, I.; McMahan, H.B.; Mironov, I.; Talwar, K.; Zhang, L. Deep learning with differential privacy. In Proceedings of the Proceedings of the 2016 ACM SIGSAC conference on computer and communications security, 2016, pp. 308–318.
16. Bonawitz, K.; Ivanov, V.; Kreuter, B.; Marcedone, A.; McMahan, H.B.; Patel, S.; Ramage, D.; Segal, A.; Seth, K. Practical secure aggregation for privacy-preserving machine learning. In Proceedings of the proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, 2017, pp. 1175–1191.
17. Aono, Y.; Hayashi, T.; Wang, L.; Moriai, S.; et al. Privacy-preserving deep learning via additively homomorphic encryption. *IEEE transactions on information forensics and security* **2017**, *13*, 1333–1345.
18. Nasr, M.; Shokri, R.; Houmansadr, A. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In Proceedings of the 2019 IEEE symposium on security and privacy (SP). IEEE, 2019, pp. 739–753.
19. Zhang, J.; Li, B.; Chen, C.; Lyu, L.; Wu, S.; Ding, S.; Wu, C. Delving into the adversarial robustness of federated learning. In Proceedings of the Proceedings of the AAAI conference on artificial intelligence, 2023, Vol. 37, pp. 11245–11253.
20. Nguyen, T.D.; Nguyen, T.; Le Nguyen, P.; Pham, H.H.; Doan, K.D.; Wong, K.S. Backdoor attacks and defenses in federated learning: Survey, challenges and future research directions. *Engineering Applications of Artificial Intelligence* **2024**, *127*, 107166.
21. McCarthy, A.; Ghadafi, E.; Andriotis, P.; Legg, P. Defending against adversarial machine learning attacks using hierarchical learning: A case study on network traffic attack classification. *Journal of Information Security and Applications* **2023**, *72*, 103398. https://doi.org/https://doi.org/10.1016/j.jisa.2022.103398.
22. McCarthy, A.; Ghadafi, E.; Andriotis, P.; Legg, P. Functionality-Preserving Adversarial Machine Learning for Robust Classification in Cybersecurity and Intrusion Detection Domains: A Survey. *Journal of Cybersecurity and Privacy* **2022**, *2*, 154–190. https://doi.org/10.3390/jcp2010010.
23. White, J.; Legg, P. Evaluating Data Distribution Strategies in Federated Learning: A Trade-Off Analysis Between Privacy and Performance for IoT Security. In Proceedings of the AI Applications in Cyber Security and Communication Networks; Hewage, C.; Nawaf, L.; Kesswani, N., Eds., Singapore, 2024; pp. 17–37.

 doi:10.20944/preprints202508.0130.v1

24. White, J.; Legg, P., Federated Learning: Data Privacy and Cyber Security in Edge-Based Machine Learning. In *Data Protection in a Post-Pandemic Society: Laws, Regulations, Best Practices and Recent Solutions*; Hewage, C.; Rahulamathavan, Y.; Ratnayake, D., Eds.; Springer International Publishing: Cham, 2023; pp. 169–193. https://doi.org/10.1007/978-3-031-34006-2_6.