AITA
ARTIFICIAL INTELLIGENCE
THEORY AND APPLICATIONS

■ RESEARCH ARTICLE

Mitigating Adversarial Attacks on ECG Classification in Federated Learning via Adversarial Training

Eyüpcan Çelik[a†], **Mehmet Kemal Güllü**[a]

[a] Department of Electrical and Electronics Engineering, İzmir Bakırçay University, İzmir, Türkiye
[†] 6017019@bakircay.edu.tr, corresponding author

RECEIVED DECEMBER 6, 2024
ACCEPTED JANUARY 29, 2025

CITATION Çelik, E., & Güllü, M. K. (2025). Mitigating adversarial attacks on ECG classification in federated learning via adversarial training. *Artificial Intelligence Theory and Applications, 5(1)*, 18-28

Abstract

Federated Learning (FL) has become an important research area in recent years, particularly when dealing with sensitive data such as healthcare information. Since healthcare data contains critical and personal information, FL provides a major advantage by enabling training on local devices without requiring data to be collected on a central server. In the analysis of healthcare data, such as electrocardiography (ECG), FL enables local processing of data while preserving privacy. However, despite its privacy benefits, FL can be vulnerable to attacks. Malicious inputs aim to degrade model accuracy, known as adversarial attacks (AA), can pose a major threat. Adversarial Training (AT) offers a defence mechanism by increasing model's robustness against such attacks. Federated Adversarial Training (FAT) extends AT into the FL environment, combining privacy advantages with enhanced resistance to adversarial inputs. In this work, we propose the use of FAT to improve both privacy and security when classifying ECG signals, ensuring robustness against AAs. This approach involves applying AT at the client level by augmenting clean ECG data with adversarial examples generated using the Projected Gradient Descent (PGD) method. A Convolutional Neural Network (CNN) architecture was employed for local training. Experiments are conducted on the MIT-BIH Arrhythmia Database (MIT-DB). For comparison, we also trained an FL model without incorporating FAT. Both models were tested on the original test data as well as on adversarially attacked versions generated using PGD, Fast Gradient Sign Method (FGSM), Carlini & Wagner (CW), and Basic Iterative Method (BIM). The results show that the FL system with FAT significantly outperforms the system without FAT in resisting AAs, with a slight compromise in performance on the original test data, thus highlighting the effectiveness of FAT in enhancing model robustness against AAs for ECG classification tasks. Code is available at *https://github.com/Skyress1/ECG-FAT-Code*.

Keywords: federated adversarial training, federated learning, adversarial attacks, ECG

1. Introduction

Machine learning enables the development of more accurate and intelligent systems by processing large datasets and provides revolutionary advances in various fields such as health [1], finance [2], and the Internet of Things [3]. However, in traditional machine learning methods, the need to collect data on a central server leads to privacy and security issues. Especially in cases where personal and sensitive data is used, these

Artificial Intelligence Theory and Applications, ISSN: 2757-9778. ISBN: 978-605-69730-2-4 © 2025 İzmir Bakırçay University

issues can expose user privacy. In response to these challenges, FL [4] is a machine learning paradigm that allows data to be processed on local devices without creating a centralized dataset. FL allows each device to train models on its local data instead of collecting and processing data from different sources. Thus, it offers significant advantages in terms of data privacy and security, especially in areas where sensitive data is used. In recent years, FL has been widely used in studies on health data to reduce privacy concerns and improve the accuracy of machine learning models.

Health data is highly sensitive, especially as data sets containing personal and biometric information. Among such data, electrocardiography (ECG) signals provide critical information about heart health by monitoring heart rhythms. While ECG data is a widely used tool for diagnosing heart conditions, the collection and processing of this data can also pose the risk of privacy violations. The high sensitivity of health data makes it imperative to ensure data privacy and security.

Protecting privacy in health data is crucial to prevent unauthorized use and sharing of individuals' personal information. While traditional centralized data processing methods require data to be collected and stored in a single location, FL reduces this risk and allows data to be processed on local devices. Thus, the protection of privacy in health data can be secured with a FL approach. Especially when it comes to highly sensitive biometric data such as ECG, the data privacy advantage provided by FL plays a critical role.

The impact of FL on health data is dramatic. FL not only ensures data privacy but also enables collaboration between different data sources. Thus, data from different hospitals or clinics can be used to train the same model without being aggregated in a centralized system. This enables improved diagnostic models for diverse patient populations.

Attacks in FL are a vulnerability that needs to be addressed in addition to the advantages offered by this technology. Especially since FL systems have a structure where model updates are made locally by each participant, malicious participants can manipulate this process. These attacks, known as AA, can degrade the accuracy and performance of the model by introducing misleading data into the model. Such attacks pose a serious threat as they can have irreversible consequences in critical areas such as health data.

One of the methods used to prevent AA is the AT approach. AT aims to make the model more resistant to attacks by using adversarial data during model training. AT provides a more robust learning process by not only improving the accuracy of the model but also its reliability. AT is used in traditional centralized learning systems. Its equivalent in FL systems is FAT.

FAT is a technique that combines the approaches of FL and AT. The FAT technique was proposed by Zizzo et al [5]. This method aims to develop models that are more resilient to AAs while maintaining the privacy advantages of FL. In terms of protecting the privacy and security of health data, FAT is considered as an important step towards developing more secure and effective machine learning models in the future.

There have been many studies on FL and ECG in the literature. Tang et al. [6] proposed a personalized FL method for ECG classification task. Manocha et al. [7] proposed a new algorithm using deep learning to classify ECG arrhythmias in a federated environment. In their proposed algorithm, they integrated a Support Vector Machine classifier with a Bi-directional Long Short-Term Memory based Auto-Encoder network. Alreshidi et al. [8] presented Fed-CL, an advanced method that combines Long Short-

Term Memory networks and Convolutional Neural Networks to accurately predict AFib utilizing FL. Çelik and Güllü [9] conducted a comparison study on server-side aggregation algorithms on Independently and Identically Distributed (IID) and Non-IID data distributions for ECG classification task.

There have also been many studies on FAT in the literature. Bondok et al. [10] used FL and AT to address privacy and security concerns in smart grids. Catak and Kuzlu [11] used FL to train a segmentation model for spectrum sensing in the presence of radar and wireless communication systems. They also used AT to combine model flexibility and local model updates into a robust global model. Luo et al. [12] proposed a new Ensemble Federated Adversarial Training (EFAT) method that enables AT to perform better in non-IID environments by extending the training data with different distortions.

In this study, we propose the use of FAT for ECG classification task to be robust against AAs while maintaining privacy and security. For this purpose, in each of the clients, the PGD [13] discarded versions of the clean data were added to the training set and the clients were made to perform AT. The original test data, PGD attacked version of the test data, FGSM [14] attacked version, CW [15] attacked version and BIM [16] attacked version of the test data were tested respectively. The results obtained are compared.

2. Materials and Methods

2.1. MIT-BIH Arrhythmia Database

The MIT-BIH Arrhythmia Database [17] (MIT-DB) was used in this study. The MIT-DB contains 48 ECG recordings of 30 minutes each. These 48 ECG recordings belong to 47 patients. The sampling frequency of all recordings is 360 Hz. The labels in the MIT-DB were edited according to the Association for the Advancement of Medical Instrumentation (AAMI) standard. The label editing is shown in Table 1. There are 5 classes in total from the AAMI standard. These are N (normal beats), S (supraventricular ectopic beats), V (ventricular ectopic beats), F (fusion beats), and Q (unclassifiable beats).

Table 1. AAMI Standards and MIT-BIH Annotation

AAMI	MIT-BIH
Normal Beat (N)	N, L, R, j, e
Supraventricular Ectopic Beat (S)	a, S, A, J
Ventricular Ectopic Beat (V)	E, V
Fusion Beat (F)	F
Unknown Beat (Q)	/, Q, f

2.2. Data Preparation and Normalization

Each ECG signal in MIT-DB was divided into 180-length windows. 180-length windows were created by taking 90 indices before and 90 indices after the beats in the ECG signals. The values in each window were normalized using Min-Max scaler to be in the range [0, 1]. The min-max scaler is given in equation (1).

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{1}$$

2.3. Deep Learning Architecture

In this study, Convolutional Neural Network (CNN) is used as the deep learning architecture. CNN architecture has 4 Convolutional layers, 4 MaxPool layers, 1 Flatten layer and 3 Linear layers. Convolutional layers consist of 16, 32, 64 and 128 filters respectively. They all have a kernel_size of 3 and a padding of 1. There is also a ReLU activation function at the output of each convolutional layer. In the MaxPool layers, kernel_size is set to 2. Linear layers consist of 256, 64 and 5 units respectively. The 256- and 64-unit linear layers have a ReLU activation function at the output. The 5-unit linear layer is the output layer of the model. The number of parameters of the architecture used is 410021. The architecture used in the study is given in Figure 1.

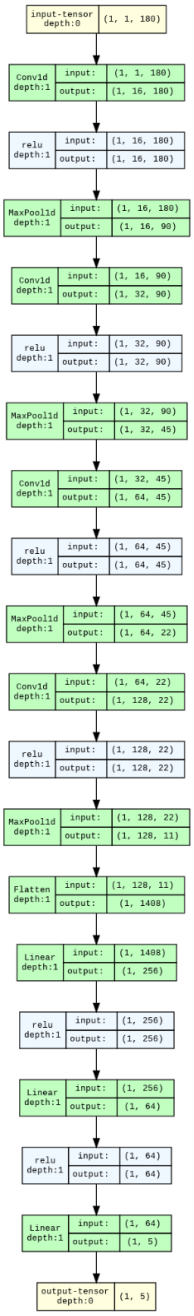

Figure 1. CNN Architecture

2.4. Algorithm

This section explains the proposed FAT system's defence mechanism against AAs in the context of ECG signal classification. The methodology is outlined in Algorithm 1, which details the local training process for a client. Initially, the client receives the global parameters. A PGD attack is then applied to the input data x_j and x_j^{adv} is obtained. Input data x_j and adversarial examples x_j^{adv} are combined to obtain $\hat{x}_j$. Original label y_j and duplicate label y_j are concatenated to obtain $\hat{y}_j$. New parameters are obtained using $\hat{x}_j$ containing both adversarial and clean samples and their labels $\hat{y}_j$. Finally, the ClientUpdate procedure is terminated by sending the new parameters to the server.

Algorithm 1. FAT for ECG signals

Input:

Client i, global parameters $\hat{\theta}$, local dataset D_i, local epoch number *E*, batch size *b*, adversarial perturbation function PGD, learning rate η

procedure ClientUpdate
  $\theta_i \leftarrow \hat{\theta}$
  for local epoch = 1, ⋯ , *E* **do**
    for mini-batch $\{x_j, y_j\}_{j=1}^{b} \sim D_i$ **do**
      $x_j^{adv} \leftarrow PGD(x_j, y_j)$
      $\hat{x}_j \leftarrow concatenate(x_j, x_j^{adv})$
      $\hat{y}_j \leftarrow concatenate(y_j, y_j)$
        $\theta_i \leftarrow \theta_i - \eta \nabla_{\theta_i} \ell_{CE}(\hat{x}_j, \hat{y}_j; \theta_i)$
    end for
  end for
  return θ_i
end procedure

After the ClientUpdate procedure is completed on all clients selected for training, the parameters sent by the clients are collected on the server. Using these parameters, the new global parameters are determined using equation (2).

$$\hat{\theta} \leftarrow \sum_{i \in S} \frac{n_i}{m} \theta_i \tag{2}$$

Where S is the list of clients selected in the current round, n_i is the data count of the ith client, m is the sum of the data counts of the clients selected in the round, θ_i is the local parameters sent by the ith client, $\hat{\theta}$ is the global parameters. This equation belongs to the FederatedAveraging (FedAvg) [4], which is used as the server-side aggregation method in this study.

3. Experimental Results

In this study, the experiments were conducted using the MIT-BIH Arrhythmia Database (MIT-BIH DB). The 48 ECG signals in the MIT-BIH DB were first divided into windows with length of 180 samples and normalized with the Min-Max scaler. A total of 109468

windows were obtained. Of these, 80% were used as training data and 20% were reserved as testing. The training data was distributed across 10 clients. Therefore, the training data was divided into 10 parts for 10 clients. In each round, 5 randomly selected clients participated in the training. The training continued for 10 rounds in total. In each round, the selected clients were trained locally for 5 epochs. FedAvg was used as the server-side aggregation algorithm. The study utilized two training approaches. First, the Non-FAT Model was trained using only clean data on clients. Second, the FAT Model incorporated AT by augmenting the training data with adversarial examples generated using a PGD attack. Testing was conducted at the end of each training round, using the test data in five variations: original (Clean), PGD attacked, FGSM attacked, CW attacked, and BIM attacked. For the PGD attack, epsilon was set to 8/255, alpha to 1/255 and the number of steps to 20. For the FGSM attack, the epsilon value was set to 8/255. For the CW attack parameters were set to c = 1, kappa = 0, 50 steps, and a learning rate of 0.01. Finally, the BIM attack used an epsilon of 8/255, alpha of 2/255, and 10 steps.

The CNN architecture was implemented using the Pytorch [18] library in Python. The FL environment was set up with the Flower [19] library, while the torchattacks [20] library was utilized for generating AAs (PGD, FGSM, CW, BIM). The training was performed on a system equipped with an AMD Ryzen 5 5600H processor, 16 GB of RAM and an NVDIA Geforce RTX 3050 graphics card. For the CNN architecture, the Adam optimizer was employed, and Cross Entropy Loss was used as the loss function. The results obtained are presented in tables and graphs. Table 2, 3, 4 and 5 show the Accuracy, Precision, Recall and F1 Score metrics for both FL system without FAT (Non-FAT Model) and FL system with FAT (FAT Model) across original test data (Clean) and adversarial datasets (PGD, FGSM, CW and BIM) in the 10^{th} round of training. Figure 2, 3, 4, 5 and 6 presents the variations in Accuracy, Precision, Recall and F1 Scores over 10 rounds on original test data and adversarial datasets. Note that round 0 represents the baseline results obtained using randomly initialized parameters.

Table 2. Accuracy results of 10th round for clean and adversarial data

Model	Clean (%)	PGD (%)	FGSM (%)	CW (%)	BIM (%)
Non-FAT Model	98.44	35.14	73.87	7.47	35.36
FAT Model	97.60	94.82	95.44	46.65	94.81

Table 3. Precision results of 10th round for clean and adversarial data

Model	Clean (%)	PGD (%)	FGSM (%)	CW (%)	BIM (%)
Non-FAT Model	93.95	30.12	47.05	7.27	30.15
FAT Model	90.55	81.72	83.76	26.80	81.55

Table 4. Recall results of 10th round for clean and adversarial data

Model	Clean (%)	PGD (%)	FGSM (%)	CW (%)	BIM (%)
Non-FAT Model	92.80	21.27	47.29	2.42	21.05
FAT Model	88.79	80.13	81.85	26.78	80.06

Table 5. F1 Score results of 10th round for clean and adversarial data

Model	Clean (%)	PGD (%)	FGSM (%)	CW (%)	BIM (%)
Non-FAT Model	92.94	22.61	45.46	3.20	22.55
FAT Model	89.06	79.75	81.70	24.02	79.63

Table 2, Table 3, Table 4 and Table 5 show that the Non-FAT Model outperformed the FAT Model in Accuracy, Precision, Recall and F1 Score metrics in the original (clean) test data. However, the FAT Model also achieved very high performance. On PGD,

FGSM, CW and BIM attacked test data, the FAT Model outperformed the Non-FAT Model in Accuracy, Precision, Recall and F1 Score metrics.

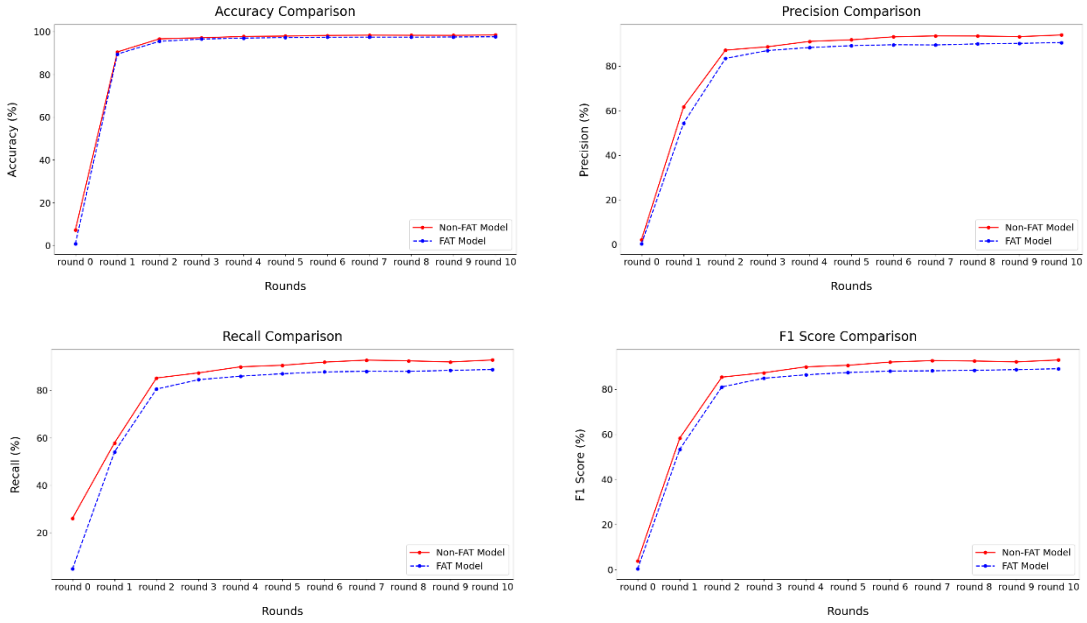

Figure 2. Clean Data Metric Results (Accuracy, Precision; Recall, F1 Score)

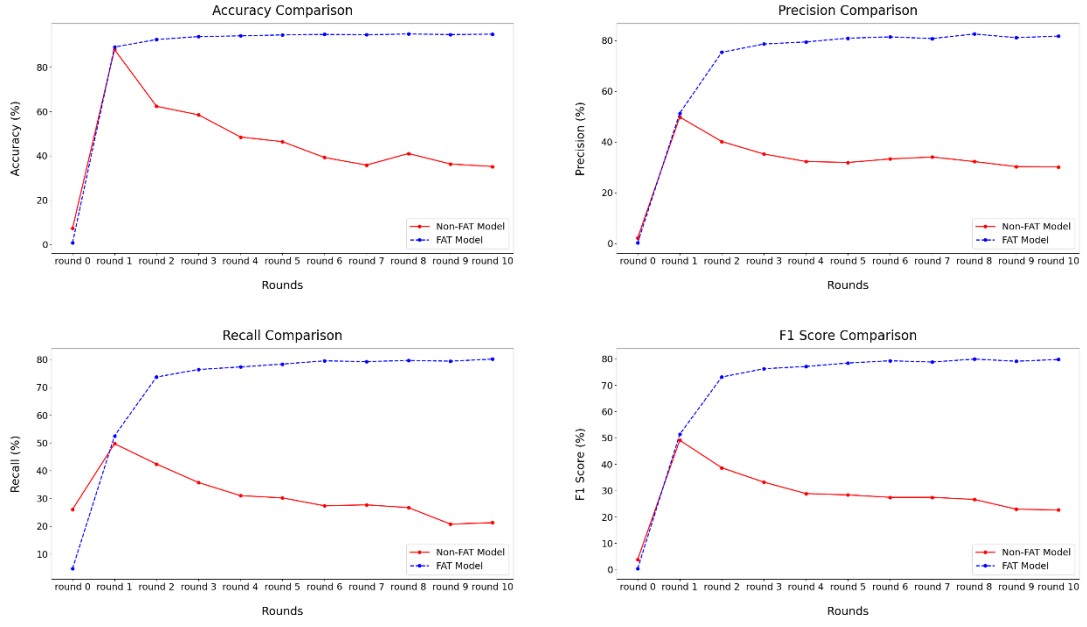

Figure 3. PGD Attacked Data Metric Results (Accuracy, Precision; Recall, F1 Score)

Figure 2 shows that the Non-FAT Model outperforms the FAT-Model across all rounds in all four metrics on the original test data. However, the performance gap between the two models is minimal. The FAT Model also achieves consistently high performance across all rounds and metrics, demonstrating its robustness even with slight compromises in comparison to the Non-FAT model.

Figure 3 shows that in round 1, the FAT and Non-FAT models exhibit comparable performance across all metrics on PGD-attacked test data, but the FAT model slightly outperforming the Non-FAT model. From round 2 onward, the FAT model performs even better, while the Non-FAT model performs very poorly against PGD-attacked data. The Non-FAT model was not able to achieve high performance on PGD-attacked data due to the lack of AT during local training and the absence of PGD-attacked instances in the dataset.

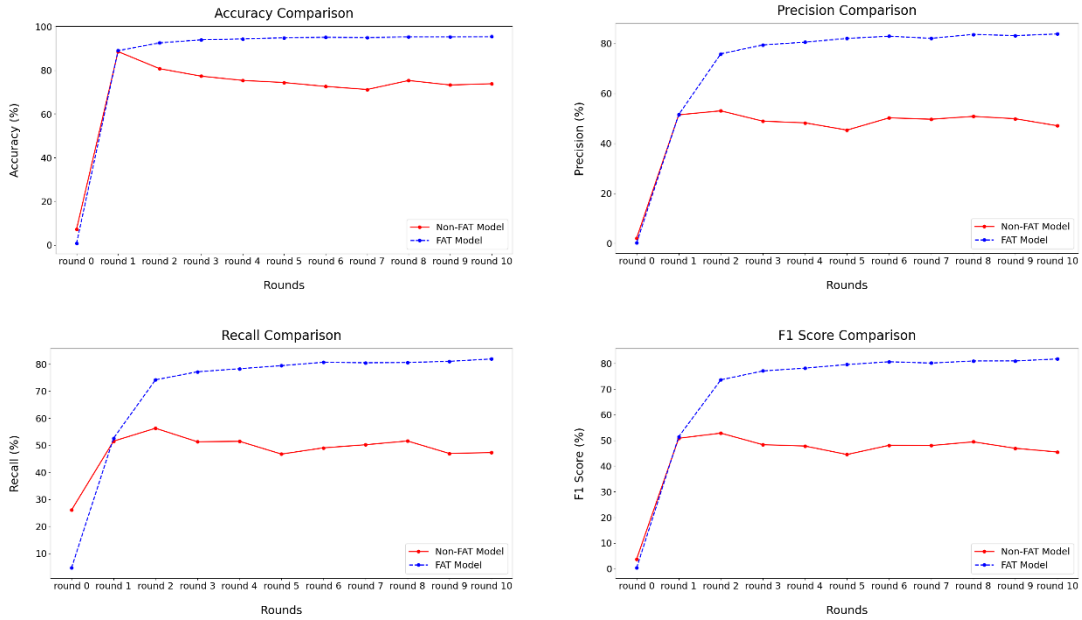

Figure 4. FGSM Attacked Data Metric Results (Accuracy, Precision; Recall, F1 Score)

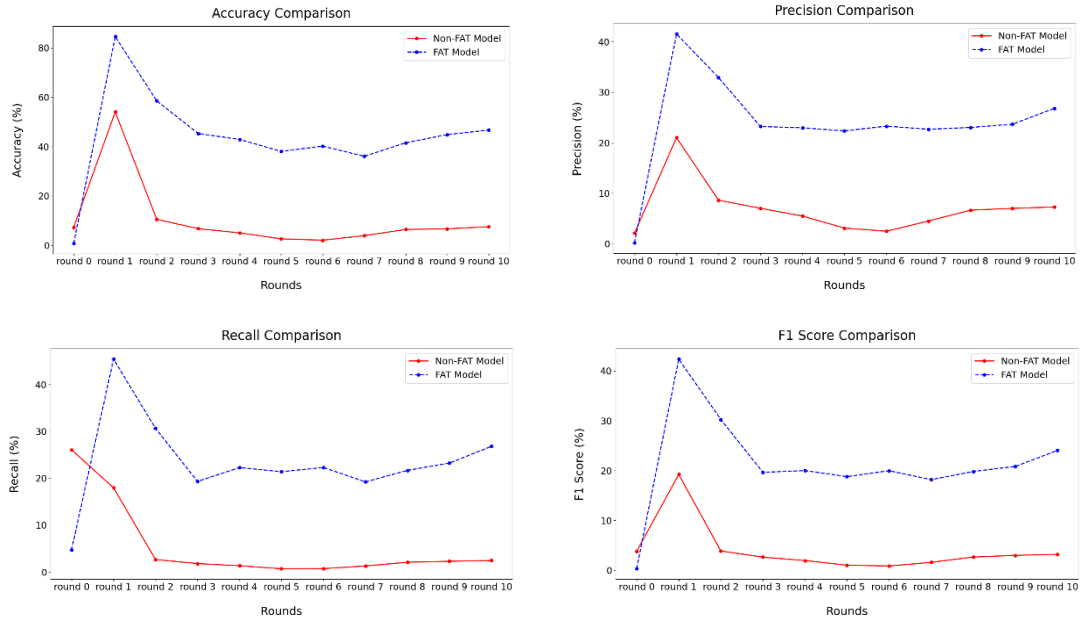

Figure 5. CW Attacked Data Metric Results (Accuracy, Precision; Recall, F1 Score)

Figure 4 illustrates that in round 1, the FAT and Non-FAT models exhibit similar performance across all four metrics on FGSM-attacked test data. However, from round

2 onward, the FAT model performs even better, while the Non-FAT model demonstrates poor performance against FGSM-attacked data. The Non-FAT model was not able to achieve high performance on FGSM-attacked data, since no AT was performed during the local training of its clients or the lack of FGSM-attacked instances in the dataset.

Figure 5 shows that the FAT model outperformed the Non-FAT model across all rounds (excluding round 0, which reflects the initial weights and is not evaluated) in all four metrics on the CW-attacked test data. The Non-FAT model achieved considerably lower results. Both models showed their highest performance in Round 1 across all metrics. However, in the following rounds, they achieved lower performance than this round. Although the FAT model outperformed the Non-FAT model, it's performance remained below the levels achieved against other types of AAs.

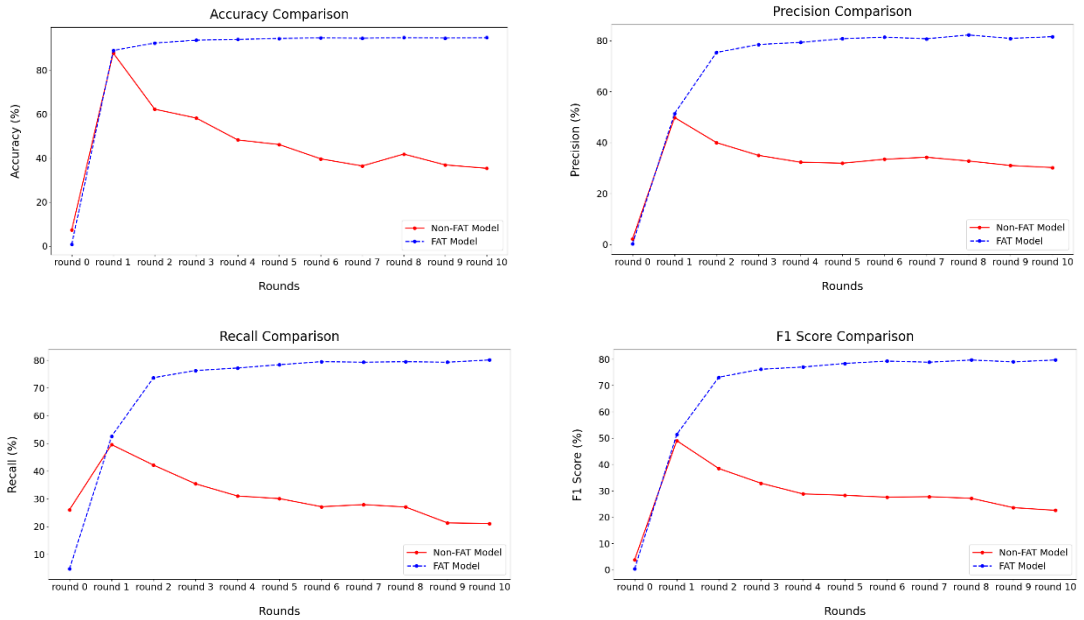

Figure 6. BIM Attacked Data Metric Results (Accuracy, Precision; Recall, F1 Score)

Figure 6 shows that in round 1, the FAT and Non-FAT Models exhibit similar performance across all four metrics on the BIM-attacked test data, with the FAT model performing slightly better. From round 2 onward, the FAT model consistently outperforms the Non-FAT model, which demonstrates very poor performance on BIM-attacked data. This underperformance of the Non-FAT model is likely due to the absence of Adversarial Training (AT) during local client training and the lack of BIM-attacked instances in the dataset.

Overall, the Non-FAT model demonstrated strong classification performance on the original data. However, it struggled with low classification performance when tested against PGD, FGSM, CW and BIM adversarial attacks. In contrast, the FAT model achieved high performance against all data except for the CW attacked data. It maintained its high performance across all datasets, except for the CW attacked data. Although its performance on CW attacks is superior to the non-FAT model, it fails to achieve the same success on CW attacks as it achieves on other adversarial data. Notably, the FAT model maintained robust performance on adversarial data with only minimal compromise in accuracy on the original data. Incorporating adversarially attacked versions of the training data during the training process, through AT, proves to be effective in enhancing the model's resilience against adversarially attacked test data.

4. Conclusion

In this study, we propose the use of FAT for ECG classification to enhance robustness against AAs while preserving privacy and security. For this purpose, in addition to clean data, adversarial examples generated using the PGD method are also used during local training on clients. The proposed framework is tested against the original test data, and adversarially attacked versions created using PGD, FGSM, CW, and BIM. Its performance was compared with that of an FL system without FAT.

The results showed that the FL system without FAT achieved high performance in Accuracy, Precision, Recall and F1 Scores on the original test data. However, it's performance dropped significantly across all four metrics for PGD, FGSM, CW and BIM attacked data. It achieved a very low performance especially against CW attacked data. In the proposed structure, i.e. the FL system using FAT, high performance is achieved in all four metrics for the original test data, PGD, FGSM and BIM attacked data. The performance on CW-attacked data, while improved compared to the Non-FAT system, was lower than for other types of adversarial attacks. When comparing the two systems, the FL system without FAT is more successful on the original test data. For the PGD, FGSM, CW and BIM attacked data, the FL system with FAT is more successful. However, the FL system with FAT is also successful on the original test data. The FL system with FAT achieves very high performance against adversarial attacked data with a little performance compromise from the original test data.

Through the FAT, ECG signal classification can achieve both enhanced privacy and security using FL, while simultaneously providing a robust defence against potential AAs. This study lays a foundation for future research exploring similar techniques with diverse types of health data.

Acknowledgement

This research has been presented in IV. International Congress on Artificial Intelligence in Health.

References

[1] Habehh, H., & Gohel, S. (2021). Machine learning in healthcare. Current Genomics, 22(4), 291–300. https://doi.org/10.2174/1389202922666210705124359

[2] Nazareth, N., & Reddy, Y. V. R. (2023). Financial applications of machine learning: A literature review. Expert Systems With Applications, 219, 119640. https://doi.org/10.1016/j.eswa.2023.119640

[3] Cui, L., Yang, S., Chen, F., Ming, Z., Lu, N., & Qin, J. (2018). A survey on application of machine learning for Internet of Things. International Journal of Machine Learning and Cybernetics, 9(8), 1399–1417. https://doi.org/10.1007/s13042-018-0834-5

[4] McMahan, H. B., Moore, E., Ramage, D., Hampson, S., & Arcas, B. a. Y. (2017). Communication-Efficient Learning of Deep Networks from Decentralized Data. International Conference on Artificial Intelligence and Statistics, 1273–1282. http://proceedings.mlr.press/v54/mcmahan17a/mcmahan17a.pdf

[5] Zizzo, G., Rawat, A., Sinn, M., & Buesser, B. (2020). FAT: Federated Adversarial Training. arXiv. https://arxiv.org/abs/2012.01791

[6] Tang, R., Luo, J., Qian, J., & Jin, J. (2021). Personalized federated learning for ECG classification based on feature alignment. Security and Communication Networks, 2021, 1–9. https://doi.org/10.1155/2021/6217601

[7] Manocha, A., Sood, S. K., & Bhatia, M. (2024). Federated learning-inspired smart ECG classification: an explainable artificial intelligence approach. Multimedia Tools and Applications. https://doi.org/10.1007/s11042-024-20084-3

[8] Alreshidi, F. S., Alsaffar, M., Chengoden, R., & Alshammari, N. K. (2024). Fed-CL- an atrial fibrillation prediction system using ECG signals employing federated learning mechanism. Scientific Reports, 14(1). https://doi.org/10.1038/s41598-024-71366-7

[9] Çelik, E., & Güllü, M. K. (2023). Comparison of federated learning strategies on ECG classification. 2023 Innovations in Intelligent Systems and Applications Conference (ASYU), 1-4. https://doi.org/10.1109/asyu58738.2023.10296796

[10] Bondok, A. H., Mahmoud, M., Badr, M. M., Fouda, M. M., Abdallah, M., & Alsabaan, M. (2023). Novel evasion attacks against adversarial training defense for smart Grid federated learning. IEEE Access, 11, 112953–112972. https://doi.org/10.1109/access.2023.3323617

[11] Catak, F. O., & Kuzlu, M. (2024). A federated adversarial learning approach for robust spectrum sensing. 2024 13th Mediterranean Conference on Embedded Computing (MECO), 1-4. https://doi.org/10.1109/meco62516.2024.10577941

[12] Luo, S., Zhu, D., Li, Z., & Wu, C. (2021). Ensemble Federated Adversarial Training with Non-IID data. arXiv. https://arxiv.org/abs/2110.14814

[13] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2017). Towards deep learning models resistant to adversarial attacks. arXiv. https://arxiv.org/abs/1706.06083

[14] Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. arXiv. https://arxiv.org/abs/1412.6572

[15] Carlini, N., & Wagner, D. (2017). Towards evaluating the robustness of neural networks. 2017 IEEE Symposium on Security and Privacy (SP), 39-57. https://doi.org/10.1109/sp.2017.49

[16] Kurakin, A., Goodfellow, I. J., & Bengio, S. (2018). Adversarial examples in the physical world. In Chapman and Hall/CRC eBooks (pp. 99–112). https://doi.org/10.1201/9781351251389-8

[17] Moody, G., & Mark, R. (2001). The impact of the MIT-BIH Arrhythmia Database. IEEE Engineering in Medicine and Biology Magazine, 20(3), 45–50. https://doi.org/10.1109/51.932724

[18] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., . . . Chintala, S. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. arXiv. https://arxiv.org/abs/1912.01703

[19] Beutel, D. J., Topal, T., Mathur, A., Qiu, X., Fernandez-Marques, J., Gao, Y., Sani, L., Li, K. H., Parcollet, T., Buarque, D. G. P. P., & Lane, N. D. (2020). Flower: a friendly federated Learning research framework. arXiv. https://arxiv.org/abs/2007.14390

[20] Kim, H. (2020). Torchattacks: a PyTorch repository for adversarial attacks. arXiv. https://arxiv.org/abs/2010.01950