A Distribution-Aware Robust Federated Learning Framework for Mobile Edge Networks

Yu Qiao[1], Phuong-Nam Tran[1], and Choong Seon Hong[2]*
[1] *Department of Artificial Intelligence, Kyung Hee University, Yongin-si 17104, Republic of Korea*
[2] *Department of Computer Science and Engineering, Kyung Hee University, Yongin-si 17104, Republic of Korea*
E-mail: {qiaoyu, tpnam0901, cshong}@khu.ac.kr

***Abstract*—Federated learning (FL) is a promising technology for achieving edge intelligence in mobile edge networks while preserving the privacy of local clients. However, a significant challenge in FL is the non-IID data across clients, which can lead to inconsistent updates between local and global models, ultimately hindering convergence. Furthermore, recent research has shown that FL models are susceptible to adversarial attacks, especially in non-IID scenarios, which can significantly impair their performance. This vulnerability poses further challenges to achieving robust and generalizable edge intelligence. In this paper, we first identify that the model's predictions for classes with fewer samples are less confident compared to those with more samples, even in federated adversarial environments. Second, recognizing that adversarial training (AT) is an effective defense mechanism, we propose a distribution-aware-assisted federated adversarial training to balance the model's predictions. Specifically, we suggest assigning higher scores to classes with fewer samples and lower scores to those with more samples during each local AT process, thereby improving the global model's robustness against adversarial attacks. Experimental results on several popular datasets show that our method achieves performance on par with or better than various baseline approaches.**

***Index Terms*—Mobile edge networks, federated learning, edge intelligence, adversarial attack, non-IID.**

I. Introduction

Given the significant advancements in storage and computing capabilities of edge smart devices, many computational tasks can be executed at the edge. This progress has led to the emergence of mobile edge networks as the next-generation computing paradigm [1]–[4]. However, extracting data from edge devices for edge computing involves concerns related to data sensitivity and legal regulations [5], [6]. Therefore, implementing mobile edge computing in a distributed and privacy-preserving manner is becoming increasingly attractive. To address this, federated learning (FL) has emerged as a promising approach for achieving edge intelligence. It allows participants to collaboratively train a shared global model by sharing model parameters while keeping their original data private [7], [8]. However, recent studies [9], [10] indicate that FL, like traditional machine learning, is vulnerable to adversarial attacks. Attackers can create adversarial examples (AEs) with subtle perturbations to deceive the model into incorrect predictions during inference. This highlights the need for a robust FL framework to counter such attacks.

To defend against such attacks, researchers commonly employ adversarial training (AT), a widely regarded method for enhancing model robustness by incorporating adversarial examples (AEs) into the training process [11], [12]. Building on the effectiveness of AT in centralized machine learning, recent studies [9], [13] propose a promising approach to enhance the global model's adversarial robustness by applying AT locally. This approach is termed federated adversarial training (FAT) [9]. However, adversarial training inevitably results in reduced prediction accuracy on clean samples compared to standard training without incorporating adversarial processes [13]. Moreover, recent studies have revealed that the non-IID data challenge in FAT poses relatively greater challenges for federated models compared to vanilla FL. This issue arises primarily because non-IID data across clients causes inconsistencies between local and global update directions, a problem further exacerbated in adversarial environments, ultimately hindering the convergence of federated training [8], [14], [15]. Therefore, it is crucial to design robust federated models that can defend against adversarial attacks even under non-IID data challenges.

In this paper, we introduce Federated Balanced Adversarial Learning (FBAL), a method aimed at enhancing the global model's adversarial robustness to adversarial attacks in non-IID settings. We focus on label non-IID challenges, where sample sizes across clients are unbalanced. This imbalance causes local models to become biased towards classes with larger sample sizes, making it more challenging to improve robustness against adversarial attacks [9]. To this end, by comparing the mean logits output from IID and non-IID data settings, we observe an intriguing trend: the model trained on a balanced dataset exhibits consistent prediction confidence across all classes. In contrast, the model trained on a non-IID dataset shows inconsistent prediction confidence, with higher confidence for classes with larger sample sizes and lower confidence for classes with smaller sample sizes. Second, inspired by the success of AT in countering adversarial attacks, we introduce combining AT with a balanced softmax loss while mitigating the inconsistencies observed, thus enhancing the global model's robustness against adversarial attacks.

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. RS-2023-00207816), (No. RS-2024-00352423), and supported by Institute of Information communications Technology Planning Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2019-0-01287, Evolvable Deep Learning Model Generation Platform for Edge Computing) *Dr. CS Hong is the corresponding author.

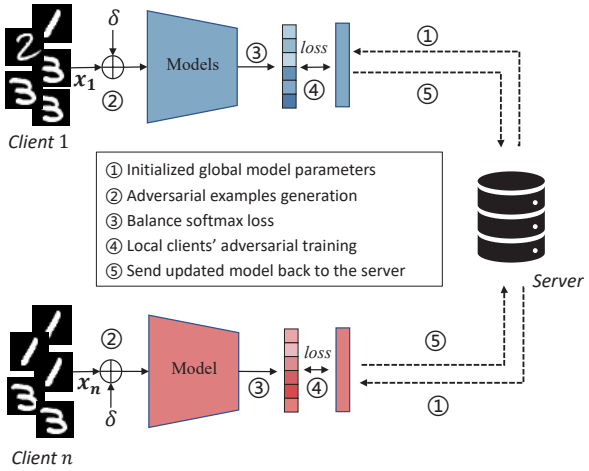


Fig. 1. Illustration of the proposed FBAL framework. We focus on steps 3 and 4 in this paper.

Specifically, by leveraging the prior distribution knowledge from each client, we suggest assigning higher scores to classes with fewer samples and lower scores to those with more samples. We hypothesize that integrating these two aspects can position FBAL as an effective approach in the face of label non-IID and adversarial attack challenges. We summarize our key findings and proposal as follows:

- We propose a robust federated framework by integrating AT with a balanced softmax loss, termed FBAL. This approach is expected to enhance the global model's adversarial robustness under non-IID challenges.
- We introduce a distribution-aware approach under the FAT framework for robust FL against both non-IID and adversarial attack challenges. We achieve this by balancing the softmax output based on the prior distribution information for each client and employing the AT strategy for each client.
- Experimental results over popular benchmark datasets: MNIST [16] and FMNIST [17], demonstrate that our approach achieves a competitive performance improvement over several baseline methods.

II. Background

A. Vanilla Federated Learning

We consider a mobile edge network, which includes N edge clients and a central server. Each edge client has its own private image dataset $\mathcal{D}_i$. The samples and label data in this image dataset are denoted as $\boldsymbol{x}_i$ and y_i, respectively. The objective of each local client is to update the shared global model from the server based on its own dataset. Its objective function can be expressed as follows [18]:

$$\mathcal{L}_i(\omega_i) = -\frac{1}{|\mathcal{D}_i|}\sum_{i\in\mathcal{D}_i}\sum_{j=1}^{C}\mathbb{1}_{y=j}\log\frac{e^{z_j}}{\sum_{j=1}^{C}e^{z_j}}, \tag{1}$$

where z represents the unnormalized prediction value output by the model, $\mathbb{1}(\cdot)$ represents the indicator function, $[C]$ represents the label space containing C classes, and the model parameters of each client are represented as ω_i. Following the standard federated training paradigm, the global objective function is defined as minimizing the sum of local losses of all distributed clients [19]:

$$\min_{\omega}\mathcal{L} = \sum_{i\in[N]}\frac{|\mathcal{D}_i|}{\sum_{i\in[N]}|\mathcal{D}_i|}\mathcal{L}_i, \tag{2}$$

where $|\mathcal{D}_i|$ denotes the size of local dataset for each client.

B. Adversarial Federated Learning

Adversarial federated learning aims to enhance the model's robustness against adversarial attacks by incorporating AEs into the federated training process, enabling the model to defend against such attacks during inference. This approach, known as AT, involves incorporating AEs into each local training process. Specifically, we first use the PGD algorithm to generate AEs, and then these generated AEs are used as inputs in each local training process. Finally, each local model can be optimized by modifying the standard loss function in Eq. 2:

$$\min_{\omega}\mathcal{L}_{adv} = \sum_{i\in[N]}\frac{|\mathcal{D}_i|}{\sum_{i\in[N]}|\mathcal{D}_i|}\mathcal{L}_i^{adv}, \tag{3}$$

where $\mathcal{L}_{adv}$ and $\mathcal{L}_i^{adv}$ represent the local training objective for each local client and the global training objective at the server during federated adversarial training, respectively. Specifically, we generate adversarial samples by maximizing the loss function $\mathcal{L}_i^{adv} = \max\mathcal{L}_i^{adv}$. Here, $\hat{\boldsymbol{x}}_i$ represents the AE generated for the clean sample $\boldsymbol{x}_i$ of each client. The AE generation process usually adopts the following PGD attack iteration algorithm [12]:

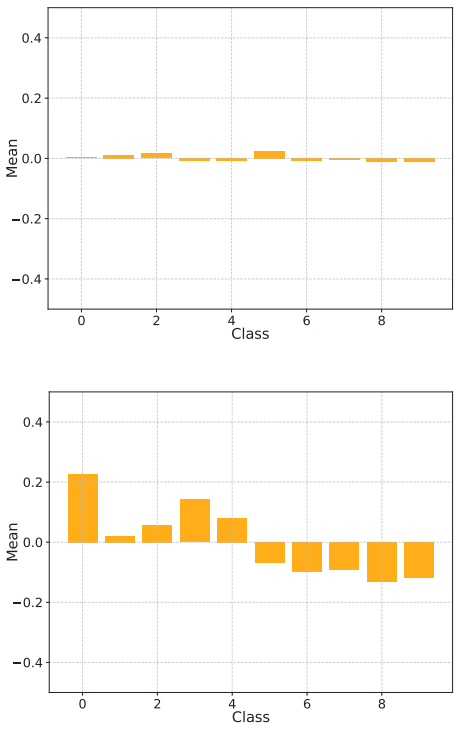
$$\boldsymbol{x_i}^{t+1} = \Pi_{\boldsymbol{x_i}+\delta}\left(\boldsymbol{x_i}^{t} + \beta\,\mathrm{sign}(\nabla_{\boldsymbol{x}}\mathcal{L}_i(\boldsymbol{x}_i;\omega_i))\right), \tag{4}$$

where β represents the step size required for each iteration during the attack iteration, $\Pi_{\boldsymbol{x}_i+\delta}$ is used to limit the range of adversarial perturbations generated, which is usually limited to a range that is imperceptible to human eyes. In addition, $\boldsymbol{x}_i^t$ represents the adversarial sample derived during each iteration, and $\mathrm{sign}(\cdot)$ represents the sign function used to find the gradient direction.

III. Methodology

A. Motivation

The paradigm of adversarial training is adopted by [9] in local clients to solve the problem of clients in federated environments being vulnerable to adversarial examples. However, this naive process of introducing the idea of adversarial training into traditional federated learning still does not solve the problems unique to FL, such as the non-IID data distribution between clients. On the other hand, this approach usually only selects part of the data for adversarial training; thus, this may not fully utilize the diversity and complexity of the dataset, which may have limited improvement in model robustness. In view of this, we first examine the difference



(b) Non-IID Data Distribution

Fig. 2. Illustration of mean logits predictions under IID and non-IID data distributions. The mean logits represent the model's average confidence across classes. In the IID case, the prediction probabilities show relatively consistent probabilities for each class (around 0). However, under non-IID data, these probabilities exhibit considerable variability (ranging from approximately -0.2 to 0.2), with notably lower confidence for classes with fewer samples.

between the model's response to IID and non-IID data in adversarial federated learning by measuring the logit value predicted by the model. To simplify our exploration, we are based on a simple multi-layer CNN network and conduct preliminary exploration on the MNIST dataset. We model IID data by sampling the same number of samples for each class with the same probability. For the non-IID distribution, the number of samples per class is sequentially chosen from {5000, 1000, 500, 250, 100, 50, 25, 10, 5, 1}. We compute the mean logits for each class during the inference stage for both IID and non-IID data and present the results in Figure 2. From the illustrated results, it can be observed that in the IID case, the predictions show relatively consistent probabilities across classes. However, under non-IID data, the prediction probabilities vary significantly, with notably lower confidence in classes with fewer samples. Since the mean logits reflect the model's average confidence across classes [20], [21], this motivates us to develop a method to balance confidence outputs for non-IID data.

B. Proposed FBAL Framework

The key distinction between adversarial FL and standard FL is in the local training updates, where AEs are incorporated into the training process. In contrast, standard FL only considers clean samples during local training. However, simply

Algorithm 1 FBAL

1: **for** $\{1, 2, ..., T\}$ **do**
2: **for** $\{0, 1,..., N\}$ **in parallel do**
3: Communicate model parameters ω^t with each client
4: $\omega_i^t \leftarrow$ **Local Training**(ω_i^t)
5: **end for**
6: $\omega^{t+1} \leftarrow \sum_{i=1}^{N} \frac{|\mathcal{D}_i|}{\sum_{i=1}^{N}|\mathcal{D}_i|}\omega_i^t$
7: **end for**
Local Training(ω_i^t)
8: **for** local epochs **do**
9: Adversarial examples generalization using Eq. 4
10: Confidence adjustment for each client using Eq. 7
11: Local adversarial training via Eq. 8
12: $\omega_i^t \leftarrow \omega_i^t - \eta\nabla\mathcal{L}_i^{bs}$
13: **end for**
14: **return** ω_i^t

combining standard FL with AT does not yield satisfactory results [9]. Based on the observations in Figure 2, this paper proposes a novel FL framework that integrates AT with a balanced softmax loss. The main training process is illustrated in Figure 1. Specifically, the initialized global model parameters are distributed to each local client in the first step. In the second step, each client updates the received model parameters using its own local private data. Specifically, in this stage, adversarial examples are created by introducing imperceptible perturbations δ to the clean training datasets of each client $\boldsymbol{x}_i$. Following this, the model outputs for each class are balanced based on their prior distributions (step 3), and adversarial training is performed on each client (step 4). Finally, each participating client sends the updated model parameters based on its local dataset (step 5) and initiates the next global iteration. This process repeats until convergence.

C. Distribution-Aware Local Adversarial Training

Neural network architectures generally include input layers that process input data, transform it into an embedded representation through intermediate layers, and then map this representation to logits in the final classification layer. The predicted logits are then used to compute a loss that measures the difference between the predictions and true labels. By minimizing this loss through iterative updates, the model improves its accuracy. However, as mentioned earlier, directly optimizing the model with non-IID data can lead to inconsistent prediction confidence across classes, particularly favoring classes with fewer samples.

Therefore, we propose a distribution-aware strategy to adjust the output confidence for each local model during the local AT process. Specifically, we calculate the prior probability for each class as follows:

$$p_{i,j} = S_j/S, \quad j \in C_i, \tag{5}$$

where C_i represents the set of classes belonging to any client i, S_j represents the sample size contained in class j, and S represents the total training sample size during each local

TABLE I
CLEAN AND ROBUST ACCURACY (%) ACROSS DIFFERENT DATASETS WITH VARYING LEVELS OF DATA HETEROGENEITY.

Dataset	MNIST						FMNIST					
$\alpha = 0.1$	Clean	BIM	FGSM	PGD-40	PGD-100	AA	Clean	FGSM	BIM	PGD-40	PGD-100	AA
FedAvg	62.76	33.92	22.40	11.80	7.04	4.76	38.14	28.10	30.58	27.86	27.78	25.34
FedProx	63.75	31.23	23.08	8.56	3.57	2.33	39.38	27.84	29.84	28.65	28.45	23.47
Scaffold	60.92	30.24	24.35	12.73	5.93	3.44	40.12	30.56	30.54	28.56	27.13	23.10
FBAL	**90.84**	**68.44**	**61.68**	**40.82**	**21.46**	**7.94**	**49.50**	**41.84**	**43.56**	**41.98**	**41.84**	**35.22**
$\alpha = 0.3$	Clean	BIM	FGSM	PGD-40	PGD-100	AA	Clean	FGSM	BIM	PGD-40	PGD-100	AA
FedAvg	72.56	41.98	33.54	17.08	9.18	**6.08**	46.30	28.48	32.64	28.46	28.40	27.70
FedProx	71.78	40.22	30.56	15.48	9.23	4.07	45.90	27.91	33.45	27.89	26.01	25.93
Scaffold	72.65	39.89	32.20	17.71	8.73	4.79	43.18	28.19	33.05	27.09	26.94	24.78
FBAL	**95.16**	**73.70**	**65.96**	**40.62**	**19.84**	5.24	**53.02**	**46.42**	**48.94**	**46.28**	**46.02**	**36.04**
$\alpha = 0.5$	Clean	BIM	FGSM	PGD-40	PGD-100	AA	Clean	FGSM	BIM	PGD-40	PGD-100	AA
FedAvg	67.30	40.18	34.02	17.80	8.18	3.66	42.34	26.88	30.34	26.76	26.74	20.70
FedProx	66.41	39.89	33.34	19.23	9.24	2.02	43.00	26.81	30.11	26.03	25.88	19.90
Scaffold	65.13	38.26	34.76	18.04	8.81	3.31	42.77	25.53	31.00	25.80	25.19	20.33
FBAL	**94.64**	**75.82**	**67.40**	**46.52**	**27.72**	**6.78**	**54.38**	**45.08**	**47.28**	**45.02**	**44.86**	**37.60**
$\alpha = 0.7$	Clean	BIM	FGSM	PGD-40	PGD-100	AA	Clean	FGSM	BIM	PGD-40	PGD-100	AA
FedAvg	70.16	35.78	30.70	14.84	7.74	4.46	47.88	30.48	34.42	30.36	30.44	29.48
FedProx	72.11	34.09	31.23	13.85	7.29	**7.01**	46.74	27.09	30.32	30.19	27.05	26.06
Scaffold	68.92	35.05	30.38	15.13	8.07	5.11	45.53	30.11	33.47	29.77	29.86	28.19
FBAL	**94.06**	**74.26**	**64.96**	**43.02**	**23.86**	6.90	**57.02**	**45.96**	**49.26**	**46.16**	**46.06**	**35.12**

iteration. Therefore, the calculated $p_{i,j}$ represents the prior distribution of client i for class j during each local iteration. Subsequently, we introduce a rescaling and enlargement factor to dynamically adjust the importance of the prior distribution among classes, as follows:

$$\hat{p}_{i,j} = \lambda \cdot e^{1-p_{i,j}}, \quad j \in C_i, \tag{6}$$

where λ is a hyper-parameter. In this paper, we set λ to 10 based on our experiments with various values, which demonstrated that this choice yields satisfactory results.

Specifically, we denote the output confidence of the classification layer in the model as $z_{i,j}^{adv}(\omega_i; \hat{\boldsymbol{x}}_{i,j})$, where $\hat{\boldsymbol{x}}_{i,j}$ represents the adversarial example generated for the j-th class of each client. Then, we balance the confidence output values as follows:

$$\hat{z}_{i,j}^{adv} = z_{i,j}^{adv}(\omega_i; \hat{\boldsymbol{x}}_{i,j}) \cdot \hat{p}_{i,j}, \quad j \in C_i, \tag{7}$$

where $\hat{z}_{i,j}^{adv}$ represents the balance score for the j-th class of each client; thus, we can rewrite Eq. 1 as follows:

$$\mathcal{L}_i^{bs}(\omega_i) = -\frac{1}{|\mathcal{D}_i|} \sum_{i \in \mathcal{D}_i} \sum_{j=1}^{C} \mathbb{1}_{y=j} \log \frac{e^{\hat{z}_j^{adv}}}{\sum_{j=1}^{C} e^{\hat{z}_j^{adv}}}. \tag{8}$$

Finally, combining the adversarial training process and the federated training strategy, the global objective of adversarial federated training can be redefined as the mean maximization of the local objective functions of all distributed clients in MEC, which is defined as follows:

$$\min_{\omega} \mathcal{L}_{adv}^{bs} = \sum_{i \in [N]} \frac{|\mathcal{D}_i|}{\sum_{i \in [N]} |\mathcal{D}_i|} \mathcal{L}_i^{bs}. \tag{9}$$

Additional details of the FBAL process are provided in Algorithm 1.

IV. EXPERIMENTS

A. Implemental Details

Basic setup. To better evaluate the effectiveness of our proposal, we compare FBAL with several other baseline methods, including FedAvg [7], FedProx [14], and Scaffold [22], all of which adopt the AT strategy. Moreover, we compared our proposal with other advanced defense methods, such as ALP [23] and TRADES [24], which we refer to as FedALP and FedTRADES, respectively. The evaluations are conducted on MNIST [16] and FMNIST [17]. The same model architecture is used in all experiments and baselines to ensure a fair comparison across all methods.

Hyperparameters. Following [5], we use the Dirichlet distribution Dir(α) to introduce non-IID data distribution among clients for all baselines. In this context, a smaller α value represents a higher degree of non-IID data distribution across devices, while a larger α indicates less skewness. In addition, we set the number of clients to 5 and the local epoch and global epoch to 1 and 100, respectively. Besides, we use the SGD optimizer with a learning rate of 0.01 and a batch iteration of 128. For evaluation, we report both clean accuracy (i.e., accuracy on unperturbed samples) and robust accuracy under adversarial perturbations. Several methods are used to assess model robustness: FGSM [11], PGD [12], BIM [25], and AA [26] attacks. We set the perturbation bound δ value for the MNIST task to 0.3 and the step size to 0.01. Similarly, we set the perturbation bound δ value for the FMNIST task

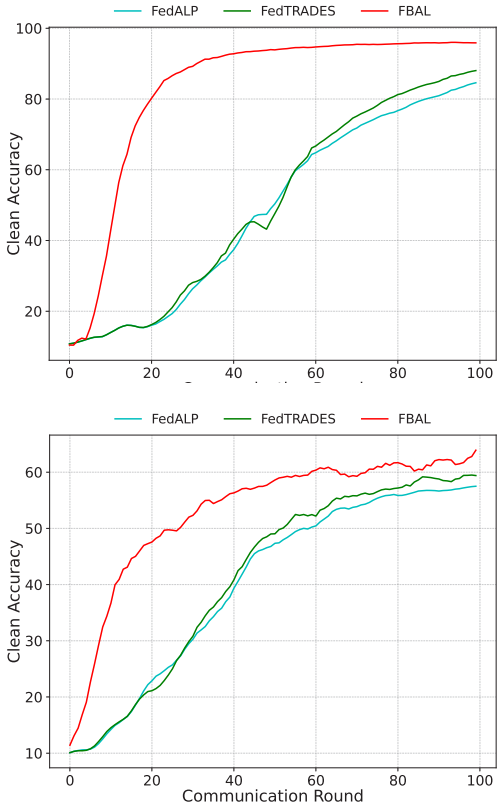


Fig. 3. Communication efficiency comparison in clean accuracy (%) for FBAL, FedALP, and FedTRADES on MNIST (top) and FMNIST (bottom) with Dir(1.0) across global communication rounds.

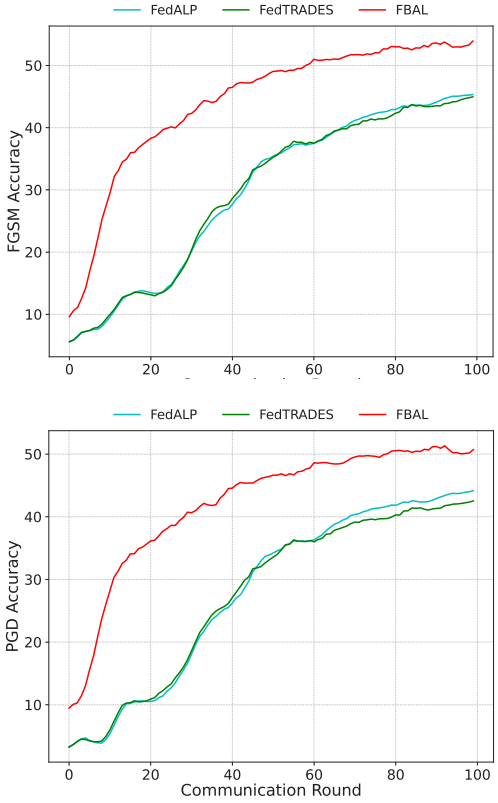


Fig. 4. Communication efficiency comparison in robust accuracy (%) for FBAL, FedALP, and FedTRADES on FMNIST with Dir(1.0) under FGSM (top) and PGD-40 (bottom) attacks across global communication rounds.

to $32/255$ and the step size to $8/255$. Note that the number of iterations used to generate adversarial samples is set to 10.

B. Performance Comparison

Performance comparison. For comparison, the methods used in the experiments include FBAL, FedAvg [7], FedProx [14], and Scaffold [22], all of which are implemented using PyTorch. The same training hyperparameters are applied to all methods. In addition, the performance of each method is evaluated by computing the average accuracy across the final five iterations. Both clean and robust accuracy (including FGSM, BIM, PGD-40, PGD-100, and AA) are reported for all methods across different levels of heterogeneity. This allows for a comprehensive comparison of our approach with others. As shown in Table I, it can be observed that our approach achieves competitive or superior performance in most cases when compared to the baselines. From the results in the table, several key observations can be made. First, varying levels of non-IID data present challenges to all baselines, including ours. As the value of α decreases, indicating higher heterogeneity among clients, the difficulty of maintaining clean accuracy and robust performance increases. For example, on the MNIST task, as the value of α decreases from 0.9 to 0.1, our model's clean accuracy drops from 95.28% to 90.48%, while the robust accuracy (e.g., under PGD-100 attacks) declines from 28.90% to 21.46%. Second, FBAL consistently outperforms all baselines across most metrics (clean, BIM, FGSM, PGD-40, PGD-100, AA). For example, under severe heterogeneity (i.e., $\alpha = 0.1$), FBAL achieves clean accuracy scores of 90.84% on MNIST and 49.50% on Fashion-MNIST, significantly surpassing the baseline FedAvg, which records 62.76% and 38.14%, respectively, demonstrating a substantial improvement. Third, the AA attack presents significant challenges to all methods, consistently resulting in the lowest accuracy compared to other attack algorithms. Nonetheless, our method still achieves superior performance under AA attacks in most cases. In summary, FBAL stands out for its robust performance against adversarial attacks under non-IID data challenges. It consistently maintains high clean accuracy and robust accuracy, outperforming many baselines in these scenarios.

Communication efficiency comparison. To further validate the effectiveness of our proposed method, we also compare FBAL with other advanced defense algorithms integrated with FedAvg, including ALP [23] and TRADES [24], referred to as FedALP and FedTRADES, respectively. Without loss of generality, we set $\alpha = 0.5$ and report the clean accuracy on MNIST and Fashion-MNIST tasks compared with different baselines in Figure 3. In addition, Figure 4 shows the robustness of the large-scale Fashion-MNIST dataset to FGSM and PGD attacks in the same heterogeneous setting. Both sets of experiments demonstrate that our approach surpasses the other

baselines in communication efficiency, suggesting relatively stronger performance in terms of both clean accuracy and adversarial robustness. From the results shown in the figures, several key observations can be made. First, regarding clean accuracy in Figure 3, FBAL demonstrates both a faster convergence rate and higher accuracy in comparison to the other methods for both the MNIST and Fashion-MNIST tasks. For instance, by the 40th communication round, FBAL's accuracy is already close to convergence, while the other methods have not yet reached this point. Interestingly, we find that both FedALP and FedTRADES have lower accuracy, which may be because these methods do not take into account the challenges of non-IID data. Second, in Figure 4, we observe that the accuracy of all methods declines under adversarial attacks compared to their clean accuracy shown in Figure 4. For example, in the Fashion-MNIST task with Dir(1.0), FBAL's clean accuracy exceeds 65%, but it drops to around 50% under PGD-40 attacks. Similarly, FedALP's clean accuracy is around 60%, but it decreases to approximately 45% under FGSM attacks. Nevertheless, despite the significant challenges posed by adversarial attacks, our approach consistently outperforms the others by a substantial margin, as evidenced by the approximately 10% improvement shown in Figure 4. These experimental results further confirm the motivation behind and the effectiveness of our proposal.

V. Conclusion and Future Work

In this paper, our toy example experiments reveal that while model output confidence consistently reflects predictions across classes under IID data, it can become biased towards classes with fewer samples under non-IID data. Based on these observations, we propose a novel robust FL framework that integrates adversarial training with a prior distribution-aware strategy to improve the model's robustness against adversarial attacks and its fairness in addressing non-IID data challenges. Experimental results across various heterogeneity settings demonstrate that our proposal achieves performance that is comparable to or exceeds that of several popular baselines. We believe our intriguing findings provide researchers with a new perspective for addressing adversarial attacks in FL models. In our future work, our proposal can be theoretically proved from the perspective of probability theory and evaluated across a wider range of tasks and model architectures.

References

[1] A. Adhikary, A. D. Raha, Y. Qiao, Y. M. Park, Z. Han, and C. S. Hong, "A power allocation framework for holographic mimo-aided energy-efficient cell-free networks," in *ICC 2024-IEEE International Conference on Communications*. IEEE, 2024, pp. 5546–5552.

[2] A. Adhikary, A. D. Raha, Y. Qiao, W. Saad, Z. Han, and C. S. Hong, "Holographic mimo with integrated sensing and communication for energy-efficient cell-free 6g networks," *IEEE Internet of Things Journal*, 2024.

[3] A. Adhikary, M. S. Munir, A. D. Raha, Y. Qiao, Z. Han, and C. S. Hong, "Integrated sensing, localization, and communication in holographic mimo-enabled wireless network: A deep learning approach," *IEEE Transactions on Network and Service Management*, 2023.

[4] Z. Jin and Y. Qiao, "A novel node selection scheme for energy-efficient cooperative spectrum sensing using d–s theory," *Wireless Networks*, vol. 26, no. 1, pp. 269–281, 2020.

[5] Y. Qiao, M. S. Munir, A. Adhikary, H. Q. Le, A. D. Raha, C. Zhang, and C. S. Hong, "Mp-fedcl: Multiprototype federated contrastive learning for edge intelligence," *IEEE Internet of Things Journal*, Sep. 2023.

[6] Y. Qiao, H. Q. Le, M. Zhang, A. Adhikary, C. Zhang, and C. S. Hong, "Fedccl: Federated dual-clustered feature contrast under domain heterogeneity," *Information Fusion*, vol. 113, p. 102645, Jan. 2025.

[7] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial Intelligence and Statistics*. PMLR, Apr. 2017, pp. 1273–1282.

[8] Y. Qiao, C. Zhang, A. Adhikary, and C. S. Hong, "Logit calibration and feature contrast for robust federated learning on non-iid data," *IEEE Transactions on Network Science and Engineering*, 2024.

[9] G. Zizzo, A. Rawat, M. Sinn, and B. Buesser, "Fat: Federated adversarial training," in *Annual Conference on Neural Information Processing Systems*, Dec. 2020.

[10] L. Lyu, H. Yu, X. Ma, C. Chen, L. Sun, J. Zhao, Q. Yang, and S. Y. Philip, "Privacy and robustness in federated learning: Attacks and defenses," *IEEE Transactions on Neural Networks and Learning Systems*, Nov. 2022.

[11] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *International Conference on Learning Representations*, CA, USA, May 2015.

[12] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *International Conference on Learning Representations*, BC, Canada, Apr. 2018.

[13] C. Chen, Y. Liu, X. Ma, and L. Lyu, "Calfat: Calibrated federated adversarial training with label skewness," *Advances in Neural Information Processing Systems*, vol. 35, pp. 3569–3581, Nov. 2022.

[14] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," *Proceedings of Machine learning and systems*, vol. 2, pp. 429–450, Mar. 2020.

[15] Y. Qiao, A. Adhikary, K. T. Kim, C. Zhang, and C. S. Hong, "Knowledge distillation assisted robust federated learning: Towards edge intelligence," in *ICC 2024-IEEE International Conference on Communications*. CO, USA: IEEE, Jun. 2024, pp. 843–848.

[16] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[17] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms," *arXiv preprint arXiv:1708.07747*, 2017.

[18] Y. Qiao, M. S. Munir, A. Adhikary, A. D. Raha, S. H. Hong, and C. S. Hong, "A framework for multi-prototype based federated learning: Towards the edge intelligence," in *2023 International Conference on Information Networking (ICOIN)*. IEEE, 2023, pp. 134–139.

[19] Y. Qiao, M. S. Munir, A. Adhikary, A. D. Raha, and C. S. Hong, "Cdfed: Contribution-based dynamic federated learning for managing system and statistical heterogeneity," in *NOMS 2023-2023 IEEE/IFIP Network Operations and Management Symposium*. FL, USA: IEEE, May 2023.

[20] V. T. Vasudevan, A. Sethy, and A. R. Ghias, "Towards better confidence estimation for neural models," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 7335–7339.

[21] H. Wei, R. Xie, H. Cheng, L. Feng, B. An, and Y. Li, "Mitigating neural network overconfidence with logit normalization," in *International conference on machine learning*. PMLR, 2022, pp. 23 631–23 644.

[22] S. P. Karimireddy, S. Kale, M. Mohri, S. Reddi, S. Stich, and A. T. Suresh, "Scaffold: Stochastic controlled averaging for federated learning," in *International Conference on Machine Learning*. PMLR, Jul. 2020, pp. 5132–5143.

[23] H. Kannan, A. Kurakin, and I. Goodfellow, "Adversarial logit pairing," *arXiv preprint arXiv:1803.06373*, 2018.

[24] H. Zhang, Y. Yu, J. Jiao, E. Xing, L. El Ghaoui, and M. Jordan, "Theoretically principled trade-off between robustness and accuracy," in *International Conference on Machine Learning*. CA, USA: PMLR, Jun. 2019, pp. 7472–7482.

[25] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," in *Artificial Intelligence Safety and Security*. Chapman and Hall/CRC, Jul. 2018, pp. 99–112.

[26] F. Croce and M. Hein, "Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks," in *International conference on machine learning*. PMLR, Nov. 2020, pp. 2206–2216.