Federated Learning for Railway Safety Analysis and Prediction

Alina Basharat
Department of Electrical & Computer Engineering
University of Texas Rio Grande Valley

Ping Xu
Assistant Professor
Department of Computer Science
Georgia State University

Zhi Tian
Assistant Professor
Department of Computer Science
George Mason University

A Report on Research Sponsored by

The University Transportation Center for Railway Safety (UTCRS)

The University of Texas Rio Grande Valley (UTRGV)

June 9, 2026

Technical Report Documentation Page

1. Report No. UTCRS-UTRGV-O2CY24	2. Government Accession No.	3. Recipient's Catalog No.
4. Title and Subtitle Federated Learning for Railway Safety Analysis and Prediction		5. Report Date June 9, 2026
		6. Performing Organization Code UTCRS-UTRGV
7. Author(s) Alina Basharat, Ping Xu, Zhi Tian		8. Performing Organization Report No. UTCRS-UTRGV-O2CY24
9. Performing Organization Name and Address University Transportation Center for Railway Safety (UTCRS) University of Texas Rio Grande Valley (UTRGV) 1201 W. University Dr. Edinburg, TX 78539		10. Work Unit No. (TRAIS)
		11. Contract or Grant No. 69A3552348340
12. Sponsoring Agency Name and Address U.S. Department of Transportation (USDOT) University Transportation Centers Program 1200 New Jersey Ave. SE Washington, DC, 20590		13. Type of Report and Period Covered Project Report June 1, 2024 – December 31, 2025
		14. Sponsoring Agency Code USDOT UTC Program
15. Supplementary Notes This report is based on a paper titled "Towards Trustworthy Federated Learning", accepted for publication at the 20th International Conference on Wireless Artificial Intelligent Computing Systems and Applications		
16. Abstract Federated learning (FL) enables collaborative model training without raw data sharing, yet ensuring its trustworthiness remains challenging due to the simultaneous presence of Byzantine attacks, fairness violations, and privacy risks. Existing works typically address these concerns in isolation, leaving their joint resolution largely unexplored. This work proposes a unified trustworthy FL framework that cohesively tackles all three challenges within a single principled design. At its core, we introduce a Two-sided Norm Based Screening (TNBS) mechanism, which defends against Byzantine attacks by discarding a proportion of gradients with the lowest and highest norms, without requiring assumptions about the attack pattern. To promote egalitarian fairness, we incorporate q-fair federated learning (q-FFL), and integrate a differential privacy (DP) scheme to safeguard local data against inference by curious parties. Experimental results on real-world datasets validate the framework's effectiveness across all three dimensions, demonstrating that robustness, fairness, and privacy can be jointly achieved without prohibitive compromise to model performance.		
17. Key Words Artificial Intelligence, Safety Analysis, Computer Models	18. Distribution Statement This report is available for download from https://www.utrgv.edu/railwaysafety/research/operations/index.htm	

19. Security Classification (of this report) None	20. Security Classification (of this page) None	21. No. of Pages 14	22. Price

Table of Contents

I. Table of Contents…….. 3

II. Disclaimer…… 4

III. Notice…… 4

IV. Acknowledgements…… 4

V. Paper: “Towards Trustworthy Federated Learning”……5

Disclaimer

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated under the sponsorship of the U.S. Department of Transportation's University Transportation Centers Program, in the interest of information exchange. The U.S. Government assumes no liability for the contents or use thereof.

Notice

This final report is for the project titled "Federated Learning for Railway Safety Analysis and Prediction", Ping Xu, PI, at the University Transportation Center for Railway Safety. It is based on a paper titled "Towards Trustworthy Federated Learning", accepted for publication at the 20th International Conference on Wireless Artificial Intelligent Computing Systems and Applications. The paper contents begin on page 5 of this report.

Acknowledgements

The authors wish to acknowledge the University Transportation Center for Railway Safety (UTCRS) for funding this project under the USDOT UTC Program Grant No. 69A3552348340. It was also partially supported by the National Science Foundation (NSF) under the ECCS Program Grant No. 2231209.

Towards Trustworthy Federated Learning

Alina Basharat[1], Ping Xu[2, ⋆[0000−0003−4810−7133]], and Zhi Tian[3[0000−0002−2738−6826]]

University of Texas Rio Grande Valley, Edinburg, TX, USA
alina.basharat01@utrgv.edu
Georgia State University, Atlanta, GA, USA
pxu4@gsu.edu
George Mason University, Fairfax, VA, USA
ztian1@gmu.edu

Abstract. Federated learning (FL) enables collaborative model training without raw data sharing, yet ensuring its trustworthiness remains challenging due to the simultaneous presence of Byzantine attacks, fairness violations, and privacy risks. Existing works typically address these concerns in isolation, leaving their joint resolution largely unexplored. This paper proposes a unified trustworthy FL framework that cohesively tackles all three challenges within a single principled design. At its core, we introduce a Two-sided Norm Based Screening (TNBS) mechanism, which defends against Byzantine attacks by discarding a proportion of gradients with the lowest and highest norms, without requiring assumptions about the attack pattern. To promote egalitarian fairness, we incorporate q-fair federated learning (q-FFL), and integrate a differential privacy (DP) scheme to safeguard local data against inference by curious parties. Experimental results on real-world datasets validate the framework's effectiveness across all three dimensions, demonstrating that robustness, fairness, and privacy can be jointly achieved without prohibitive compromise to model performance.

Keywords: Trustworthy federated learning · Byzantine attacks · fairness · privacy preservation.

1 Introduction

Federated learning (FL) has emerged as a promising paradigm for large-scale machine learning, enabling multiple clients to collaboratively train a global model without sharing their private data [17]. While early FL research focused primarily on model accuracy and training convergence, growing deployment in sensitive real-world applications has shifted attention toward the trustworthiness of FL systems [7]. Ensuring such trustworthiness requires simultaneously addressing three critical challenges: privacy leakage, Byzantine robustness, and fairness, each of which has been studied in relative isolation in the existing literature.

⋆ Corresponding author: Ping Xu (email: pxu4@gsu.edu).

First, although raw data remains on local devices in FL, shared gradients or model parameters can inadvertently leak sensitive information through techniques such as model inversion attacks [5]. A widely adopted remedy is differential privacy (DP), which adds calibrated Gaussian noise to model updates [3, 15, 2]. To mitigate the accuracy degradation introduced by noise, subsequent works employ diminishing noise variance [6] or attenuating noise factors [14] to recover convergence to an optimal solution. Alternative privacy-preserving approaches include homomorphic encryption [4, 8] and secure multi-party computation [20].

Second, the iterative communication between local clients and the central server exposes FL systems to Byzantine attacks [16], where a set of malfunctioning nodes transmit corrupted or adversarial updates to manipulate the global model [9]. Existing defenses include Krum and Multi-Krum, which select reliable updates based on pairwise Euclidean distances among clients [1]; median and trimmed mean aggregation rules that statistically suppress adversarial influence [18]; and the norm-based screening (NBS) method, which discards updates with abnormally large norms [21].

Lastly, standard FL minimizes an aggregate loss without accounting for heterogeneity across clients, which can systematically disadvantage certain groups and yield models that perform well on average but poorly for underrepresented participants [11]. To address this issue, Agnostic Federated Learning (AFL) applies minimax optimization to minimize the worst-case client loss [13], while others reweight the FedAvg objective to improve distributional fairness [19]. A particularly effective approach introduces a fairness control parameter q in the training objective to amplify the influence of poorly performing clients on global updates, promoting egalitarian fairness across all participants [11, 21].

Despite progress along each of these axes, their joint treatment remains largely unexplored. Addressing them independently is insufficient in practice, as the interactions among robustness mechanisms, fairness objectives, and privacy constraints can undermine the effectiveness of each component when deployed together. This paper proposes a unified trustworthy FL framework that cohesively integrates all three properties. For robustness, we introduce a Two-sided Norm-Based Screening (TNBS) mechanism that filters malicious participants by trimming gradients at both extremes of the norm distribution, addressing the blind spot of one-sided approaches. For fairness, we adopt the q-fair FL objective to upweight the loss of under-performing clients. For privacy, we apply client-side DP via Gaussian noise injection to prevent data inference by curious parties.

Our main contributions include: 1) We propose a comprehensive FL framework that simultaneously ensures Byzantine robustness, egalitarian fairness, and differential privacy, with TNBS as a novel two-sided screening mechanism that improves upon existing one-sided norm-based defenses. 2) We conduct extensive experiments on real-world datasets, demonstrating that the proposed framework effectively achieves robustness, fairness, and privacy preservation without prohibitive compromise to model performance.

2 Proposed Framework

2.1 System Model and Problem Formulation

Consider a FL system consisting of N clients and one central server. Each client i holds a local dataset $\mathcal{D}_i$ of n_i samples drawn from a potentially distinct local distribution, with $n = \sum_{i=1}^{N} n_i$ denoting the total number of samples across all clients. The data distributions across clients are assumed to be non-IID, reflecting realistic heterogeneity in practice. Among the N clients, a subset, denoted as $\mathcal{B}$, may be **Byzantine attackers** that transmit corrupted updates to manipulate the global model. There also exist curious third parties that attempt to infer sensitive information from local model parameters. The central server seeks to learn a global model parameterized by $\boldsymbol{\theta} \in \mathbb{R}^d$ by aggregating local updates from clients over T communication rounds. At each round t, the server broadcasts the current model $\boldsymbol{\theta}^t$ to all clients; each client computes a local gradient and transmits it back to the server, which then aggregates the received gradients and updates the global model. The overarching objective is to ensure that this process is simultaneously **robust** to Byzantine manipulation, **privacy-preserving** against data inference attacks, and **fair** across clients with heterogeneous data.

Formally, let $F_i(\boldsymbol{\theta})$ denote the local loss function of client i, computed on its local dataset $\mathcal{D}_i$. In standard FL, the global objective is:

$$F(\boldsymbol{\theta}) = \sum_{i=1}^{N} p_i F_i(\boldsymbol{\theta}), \quad p_i = \frac{n_i}{n}, \tag{1}$$

where p_i weights each client proportionally to its data size. While this formulation achieves good average performance, it does not account for fairness across clients or defend against adversarial participants. The following subsections describe how our framework addresses each of these limitations.

2.2 Fairness-Aware Local Objective

In standard FL, clients with larger datasets exert disproportionately greater influence on the global model [11]. When data is unevenly distributed, this can systematically disadvantage clients with smaller or more diverse datasets, yielding a global model that performs well on average but poorly for underrepresented participants, a manifestation of **egalitarian unfairness**.

To address this issue, we adopt the q-fair FL (q-FFL) objective [11], which introduces a tunable parameter $q \geq 0$ to reweight the contribution of each client based on its local loss. Specifically, the global fairness-aware objective is defined as:

$$H(\boldsymbol{\theta}) = \sum_{i=1}^{N} \frac{p_i}{q+1} F_i^{q+1}(\boldsymbol{\theta}). \tag{2}$$

When $q = 0$, (2) reduces to the standard FL objective. As q increases, greater emphasis is placed on poorly performing clients, driving the global model toward more uniform accuracy across all participants.

Each client i therefore optimizes a fairness-reweighted local objective $H_i(\boldsymbol{\theta}) = F_i^{q+1}(\boldsymbol{\theta})$, and computes its local gradient w.r.t the current global model $\boldsymbol{\theta}^t$:

$$\boldsymbol{g}_i^t = \nabla H_i(\boldsymbol{\theta}^t) = (q+1)\, F_i^q(\boldsymbol{\theta}^t)\, \nabla F_i(\boldsymbol{\theta}^t). \tag{3}$$

Each client then transmits $\boldsymbol{g}_i$ to the central server. However, as discussed next, transmitting raw gradients poses a privacy risk.

2.3 Privacy Preservation via Differential Privacy

Although raw data never leaves the client in FL, transmitted gradients can still expose sensitive information. Techniques such as model inversion attacks [5] can reconstruct private training samples from gradient observations, making gradient transmission a potential privacy vulnerability. To prevent such inference, we protect each client's gradient using **differential privacy (DP)**, a mathematically rigorous framework that bounds the information any observer can extract about any individual's data from the output of a mechanism.

Definition 1 ((ϵ, δ)-Differential Privacy). *A randomized mechanism $\mathcal{M} : \mathcal{D} \to \mathcal{R}$ satisfies (ϵ, δ)-DP if, for any two adjacent datasets $D, D' \in \mathcal{D}$ differing by a single data record, and for all measurable output sets $S \subseteq \mathcal{R}$:*

$$\Pr\big[\mathcal{M}(D) \in S\big] \leq e^{\epsilon} \Pr\big[\mathcal{M}(D') \in S\big] + \delta, \tag{4}$$

*where $\epsilon > 0$ is the **privacy budget** controlling the worst-case privacy loss (smaller ϵ implies stronger privacy), and $\delta \geq 0$ is a small slack probability.*

To achieve (ϵ, δ)-DP, each client i first clips its gradient to bound its ℓ_2-sensitivity:

$$\boldsymbol{g}_i^t \leftarrow \frac{\boldsymbol{g}_i^t}{\max(1,\, \|\boldsymbol{g}_i^t\|_2 / C)}, \tag{5}$$

where $C > 0$ is the clipping threshold.

Then, each client i perturbs its local gradient $\boldsymbol{g}_i^t$ by adding a zero-mean Gaussian noise vector before transmission:

$$\tilde{\boldsymbol{g}}_i^t = \boldsymbol{g}_i^t + \boldsymbol{\xi}_i, \quad \boldsymbol{\xi}_i \sim \mathcal{N}(\mathbf{0},\, \sigma^2 \mathbf{I}), \tag{6}$$

where $\mathbf{I} \in \mathbb{R}^{d \times d}$ is the identity matrix. The noise variance σ^2 is calibrated as:

$$\sigma^2 = \frac{2C^2 \ln(1.25/\delta)}{\epsilon^2}. \tag{7}$$

Achieving stronger privacy (smaller ϵ in (7)) requires larger noise σ, which degrades model accuracy, demonstrating a privacy-accuracy trade-off in the DP mechanism.

2.4 Byzantine Robustness via Two-Sided Norm-Based Screening

After clients transmit their privatized gradients $\{\tilde{\boldsymbol{g}}_i^t\}_{i=1}^N$, the server must aggregate them into a single update. Under Byzantine attacks, a subset of malicious clients may transmit arbitrarily corrupted gradients to bias the global model. [21] proposes norm-based screening (NBS), which discards gradients with abnormally large norms on the grounds that honest gradients should not deviate excessively from the mean. However, NBS implicitly assumes that gradients with small norms are benign. In practice, adversaries can submit near-zero or deliberately small gradients to evade detection while still exerting a biasing influence on the aggregation. This motivates our proposed **Two-Sided Norm-Based Screening (TNBS)** method, which filters outliers at *both* extremes of the gradient norm distribution.

Let $\tilde{\boldsymbol{g}}_1^t, \tilde{\boldsymbol{g}}_2^t, \ldots, \tilde{\boldsymbol{g}}_N^t$ denote the gradients received by the server. TNBS proceeds in four steps:

Step 1: Norm Computation and Sorting. Compute the ℓ_2-norm of each received gradient and sort them in ascending order:

$$\|\tilde{\boldsymbol{g}}_{(1)}^t\| \leq \|\tilde{\boldsymbol{g}}_{(2)}^t\| \leq \cdots \leq \|\tilde{\boldsymbol{g}}_{(N)}^t\|, \tag{8}$$

where $(\cdot)$ denotes the rank index after sorting.

Step 2: Threshold Determination. Given a screening parameter $p \in (0, 1]$ specifying the proportion of gradients to filter out, compute the lower and upper rank thresholds:

$$Q_{\text{low}} = \left\lfloor \frac{p}{2} \cdot N \right\rfloor, \qquad Q_{\text{high}} = \left\lceil \left(1 - \frac{p}{2}\right) \cdot N \right\rceil. \tag{9}$$

These thresholds symmetrically exclude the $\lfloor \frac{pN}{2} \rfloor$ gradients with the smallest norms and $\lfloor \frac{pN}{2} \rfloor$ gradients with the largest norms, retaining only those in the central portion of the norm distribution.

Step 3: Screening. Construct the set of retained gradients:

$$\mathcal{S} = \left\{ i : Q_{\text{low}} \leq \text{rank}(i) \leq Q_{\text{high}} \right\}, \tag{10}$$

where $\text{rank}(i)$ is the rank of client i's gradient norm in the sorted sequence.

Step 4: Aggregation. Average the retained gradients to form the aggregated update:

$$\bar{\boldsymbol{g}}^t = \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \tilde{\boldsymbol{g}}_i^t. \tag{11}$$

The intuition behind TNBS is straightforward: honest gradients, shaped by genuine local data, tend to cluster in the middle of the norm distribution. Attackers who submit inflated updates to dominate aggregation are caught by the upper threshold Q_{high}, while attackers who submit near-zero updates to subtly steer the model are caught by the lower threshold Q_{low}. The parameter p controls the stringency of screening: larger p provides stronger robustness at the cost of discarding more potentially honest gradients.

Algorithm 1 Unified Trustworthy FL Framework

Require: Number of clients N, rounds T, learning rate η, fairness parameter q, DP parameters (ϵ, δ, C), screening parameter p
Ensure: Trained global model $\boldsymbol{\theta}^t$
Initialize global model $\boldsymbol{\theta}^0$
for $t = 0, 1, \ldots, T-1$ **do**
  Server broadcasts $\boldsymbol{\theta}^t$ to all clients
  for each honest client i **in parallel do**
    Compute fairness-aware gradient $\boldsymbol{g}_i^t$ via (3)
    Clip gradient $\boldsymbol{g}_i^t$ via (5) and add DP noise to obtain $\tilde{\boldsymbol{g}}_i^t$ via (6)
    Transmit $\tilde{\boldsymbol{g}}_i^t$ to server
  end for
  for each Byzantine client $i \in \mathcal{B}$ **do**
    Send arbitrary value to the server
  end for
  Server sorts gradients by norm via (8)
  Compute thresholds Q_{low}, Q_{high} via (9) and retain screened set $\mathcal{S}$ via (10)
  Aggregate to obtain $\bar{\boldsymbol{g}}^t$ via aggregation (11) and update model parameter $\boldsymbol{\theta}^{t+1}$ via (12)
end for
return $\boldsymbol{\theta}^T$

2.5 Global Model Update

After aggregation, the server updates the global model via gradient descent:

$$\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t - \eta \cdot \bar{\boldsymbol{g}}^t, \tag{12}$$

where $\eta > 0$ is the learning rate. The complete procedure is summarized in Algorithm 1.

3 Experiments

This section compares the proposed method with several benchmarks, such as trimmed mean (TM), coordinate-wise trimmed mean (CWTM) [18], Krum [1], and a recently proposed H-nobs framework [21]. For fair comparison, all benchmark methods also utilize DP encryption and optimize the fairness-promoting objective function.

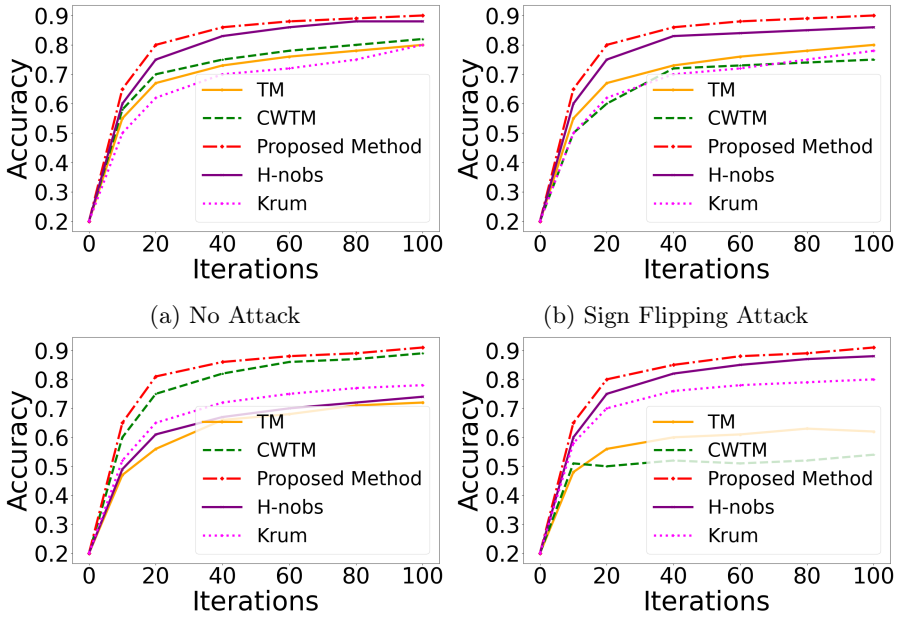
3.1 Experimental setup

Two datasets, MNIST [10] and Spam [12] are adopted in the experiments. The Spam dataset contains $4,601$ email text messages, labeled as either spam (1) or not spam (0). The MNIST dataset consists of 70,000 grayscale images of handwritten digits from 0 to 9. For each dataset, we allocate two-thirds for training and the remaining one-third for testing. In the experiments, we consider

a total of 20 clients, among which 4 are Byzantine. The data distributions among clients are non-iid. For the Spam dataset, data is evenly split among the clients with 4 nodes exclusively contain spam emails (labeled as 1) and the remaining 16 nodes contain non-spam emails (labeled as 0). For the MNIST dataset, each digit is represented equally across exactly two clients.

3.2 Results

Fig. 1 and Fig. 2 demonstrate the Byzantine-robustness of the proposed method, showing that the proposed method consistently outperforms the benchmark methods in all tests. Take the Spam dataset as an example, the H-nobs algorithm has degraded performance in the label flipping attack (Fig. 1c) while TM and CWTM fail in the Gaussian attack (Fig. 1d). On the other hand, our proposed method shows best resilience across all attacks and maintains the highest accuracy.



(a) No Attack

(b) Sign Flipping Attack

(c) Label Flipping Attack

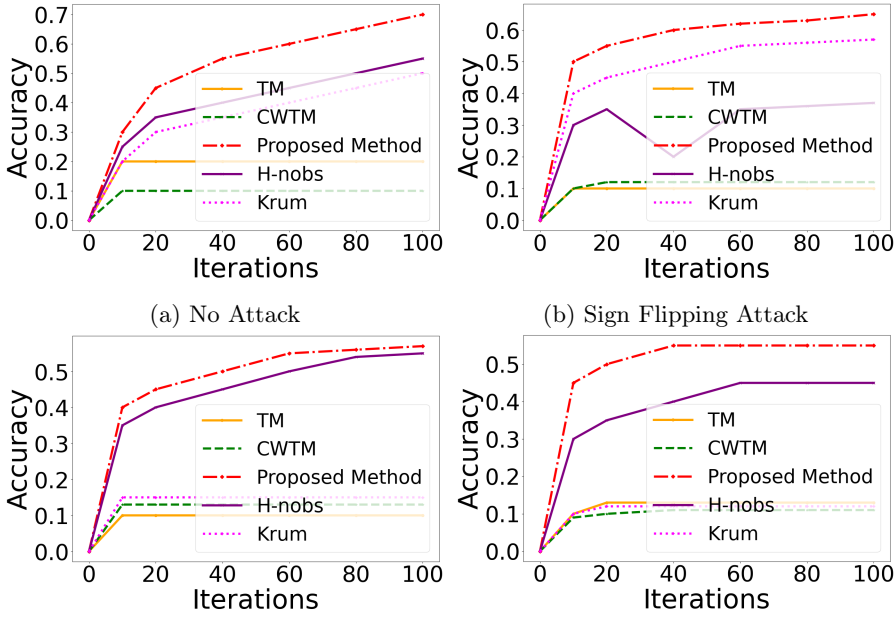
(d) Gaussian Attack

Fig. 1: Comparison of model accuracy for the Spam based dataset.

We then evaluate fairness under Gaussian attack by examining the variance in accuracy performance across different clients. A lower variance in accuracy among clients indicates a more fair trained model. From Table 1, we can see that the proposed method achieves the best performance in both the overall accuracy and the variance.

Table 1: Accuracy (variance) performance of different methods with different q values.

q parameter	Proposed Method	H-nobs	Krum	TM	CWTM
$q = 0$	**92.6(189)**	92.5(396)	90.1(320)	89.0(359)	89.0(359)
$q = 0.5$	**92.4(156)**	92.3(368)	89.2(200)	84.1(276)	89.0(359)
$q = 1$	**92.0(116)**	89.1(350)	75.0(63)	59.0(240)	50.0(292)



(a) No Attack

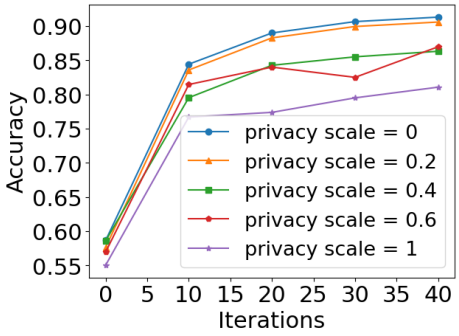
(b) Sign Flipping Attack

(c) Label Flipping Attack

(d) Gaussian Attack

Fig. 2: Comparison of model accuracy for the MNIST dataset.

The impact of DP noise on model accuracy is evaluated using the Spam dataset. The accuracy-privacy trade-off is shown in Fig. 3. Models with no added noise achieves the highest accuracy and increasing noise levels reduces accuracy. However, even at the highest noise level, the proposed method retains significant learning capability.



4 Conclusion

This paper addresses three key challenges in federated learning: client-level fairness, privacy, and Byzantine attack resilience. We achieve fairness through q-fair federated learning, ensure privacy with Gaussian noise-added differential privacy, and defend against attacks using a two-sided norm-based screening method. Future work will be devoted to theoretical analysis of the proposed method.

Acknowledgments. This work was partially supported by the National Science Foundation (NSF) under the ECCS Program (Grant No. 2231209) and by the University Transportation Center for Railway Safety (UTCRS) under the USDOT UTC Program (Grant No. 69A3552348340).

References

1. Blanchard, P., El Mhamdi, E.M., Guerraoui, R., Stainer, J.: Machine Learning with Adversaries: Byzantine Tolerant Gradient Descent. In: Advances in Neural Information Processing Systems. vol. 30 (2017)
2. Chen, S., Yang, J., Wang, G., Wang, Z., Yin, H., Feng, Y.: CLFLDP: Communication-Efficient Layer Clipping Federated Learning with Local Differential Privacy. Journal of Systems Architecture **148**, 103067 (2024)
3. Geyer, R.C., Klein, T., Nabi, M.: Differentially Private Federated Learning: A Client Level Perspective. arXiv preprint arXiv:1712.07557 (2017)
4. Hao, M., Li, H., Luo, X., Xu, G., Yang, H., Liu, S.: Efficient and Privacy-Enhanced Federated Learning for Industrial Artificial Intelligence. IEEE Transactions on Industrial Informatics **16**(10), 6532–6542 (2020). https://doi.org/10.1109/TII.2019.2945367
5. Huang, Y., Gupta, S., Song, Z., Li, K., Arora, S.: Evaluating Gradient Inversion Attacks And Defenses In Federated Learning. Advances in Neural Information Processing Systems **34**, 7232–7241 (2021)
6. Huang, Z., Hu, R., Guo, Y., Chan-Tin, E., Gong, Y.: DP-ADMM: ADMM-based Distributed Learning with Differential Privacy. IEEE Transactions on Information Forensics and Security **15**, 1002–1012 (2019)

7. Jagatheesaperumal, S.K., Rahouti, M., Alfatemi, A., Ghani, N., Quy, V.K., Chehri, A.: Enabling Trustworthy Federated Learning in Industrial IoT: Bridging the Gap Between Interpretability and Robustness. IEEE Internet of Things Magazine **7**(5), 38–44 (2024)
8. Jin, W., Yao, Y., Han, S., Joe-Wong, C., Ravi, S., Avestimehr, S., He, C.: FedML-HE: An Efficient Homomorphic-Encryption-Based Privacy-Preserving Federated Learning System. arXiv preprint arXiv:2303.10837 (2023)
9. Lamport, L., Shostak, R., Pease, M.: The Byzantine Generals Problem, pp. 203–226. Association for Computing Machinery, New York, NY, USA (2019), https://doi.org/10.1145/3335772.3335936
10. Lecun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-Based Learning applied to Document Recognition. Proceedings of the IEEE **86**(11), 2278–2324 (1998). https://doi.org/10.1109/5.726791
11. Li, T., Sanjabi, M., Beirami, A., Smith, V.: Fair Resource Allocation in Federated Learning. arXiv preprint arXiv:1905.10497 (2019)
12. Mark, H., Erik, R., George, F., Jaap, S.: Spambase. UCI Machine Learning Repository (1999), DOI: https://doi.org/10.24432/C53G6X
13. Mohri, M., Sivek, G., Suresh, A.T.: Agnostic Federated Learning. In: International Conference on Machine Learning. pp. 4615–4625. PMLR (2019)
14. Wang, Y., Nedić, A.: Tailoring Gradient Methods for Differentially-Private Distributed Optimization. IEEE Transactions on Automatic Control (2023)
15. Wei, K., Li, J., Ding, M., Ma, C., Yang, H.H., Farokhi, F., Jin, S., Quek, T.Q., Poor, H.V.: Federated Learning with Differential Privacy: Algorithms and Performance Analysis. IEEE Transactions on Information Forensics and Security **15**, 3454–3469 (2020)
16. Xu, X., Lyu, L.: Towards Building a Robust and Fair Federated Learning System. arXiv preprint arXiv:2011.10464 (2020)
17. Yang, Q., Liu, Y., Chen, T., Tong, Y.: Federated Machine Learning: Concept and Applications. ACM Transactions on Intelligent Systems and Technology **10**(2), 1–19 (2019). https://doi.org/10.1145/3298981, https://doi.org/10.1145/3298981
18. Yin, D., Chen, Y., Kannan, R., Bartlett, P.: Byzantine-Robust Distributed Learning: Towards Optimal Statistical Rates. In: International Conference on Machine Learning. pp. 5650–5659. PMLR (2018)
19. Zeng, Y., Chen, H., Lee, K.: Improving Fairness via Federated Learning. arXiv preprint arXiv:2110.15545 (2021)
20. Zhao, J., Zhu, H., Wang, F., Lu, R., Liu, Z., Li, H.: PVD-FL: A Privacy-Preserving and Verifiable Decentralized Federated Learning Framework. IEEE Transactions on Information Forensics and Security **17**, 2059–2073 (2022)
21. Zhou, G., Xu, P., Wang, Y., Tian, Z.: H-nobs: Achieving Certified Fairness and Robustness in Distributed Learning on Heterogeneous Datasets. Advances in Neural Information Processing Systems **36** (2024)