Lightweight Client Weighting for Robust Federated Learning Under Label-Flipping Attacks

Haoyou Wang *

Beijing Normal University, Beijing, China

* Corresponding Author Email: 202211260021@mail.bnu.edu.cn

Abstract. Federated learning (FL) enables privacy-preserving machine learning across decentralized clients without sharing raw data. However, it faces significant challenges, particularly from label-flipping (LF) attacks, where malicious clients mislabel data, compromising model performance. This paper proposes a robust aggregation method that mitigates the impact of LF attacks by combining validation-based weighting with update consistency. This method adaptively adjusts client weights based on their validation accuracy and the consistency of their model updates across rounds. Unlike traditional approaches, which rely on costly computation or access to trusted validation sets, this lightweight method requires no access to local data or extensive computation, making it suitable for real-world applications. Experimental results using the Fashion-MNIST dataset show that the proposed method effectively maintains competitive accuracy under LF attacks, outperforming standard aggregation methods such as FedAvg and Trimmed Mean. This strategy offers a practical, scalable solution for FL in adversarial settings, providing robust defense against label-flipping attacks without compromising the system's efficiency.

Keywords: Machine Learning, Label-Flipping Attack, Federated Learning, Client Weighting.

1. Introduction

As privacy concerns continue to rise, Federated Learning (FL) has become a promising solution for training machine learning models across decentralized clients without the need to share raw data. This decentralized approach offers significant advantages in terms of data confidentiality, as each client maintains control over its own data. However, the privacy and security of FL are still vulnerable to various adversarial attacks, particularly label-flipping (LF) attacks, which can compromise the integrity of the training process.

In a label-flipping attack, malicious clients intentionally mislabel the data they send to the server, thereby poisoning the global model during the aggregation process. This type of attack is particularly dangerous because it can subtly degrade the model's performance without altering the structure of the data or the model itself, making it difficult to detect. Due to the trust-based nature of FL, where clients are trusted to update models based on their local datasets, label-flipping attacks exploit this vulnerability, leading to significant performance degradation, especially when the attack is synchronized across multiple rounds [1].

The inherent challenge in defending against LF attacks lies in the decentralized nature of FL and the absence of centralized data, which makes it difficult to monitor and validate each client's updates. Traditional methods for handling such attacks often rely on expensive computations or assume the existence of trusted validation sets, which are not feasible in real-world, resource-constrained edge environments [2]. Consequently, there is a growing need for efficient and lightweight defenses that can detect and mitigate the impact of label-flipping attacks using only observable training dynamics, without requiring access to raw client data or extensive computational resources.

This paper proposes a novel aggregation method for FL that addresses label-flipping attacks by integrating a validation-based weighting mechanism with update consistency. This method adapts to adversarial behaviors by dynamically adjusting the weights of client updates based on their validation accuracy and the stability of their model updates across communication rounds. Unlike traditional aggregation methods such as FedAvg, which are vulnerable to LF attacks, this approach does not rely on costly computation or trusted validation data. Instead, it leverages the consistency of client updates

and the performance of local models on a small validation set to robustly aggregate the model, reducing the influence of malicious updates [2,3].

The experimental results, conducted using the Fashion-MNIST dataset, demonstrate that the proposed approach significantly enhances the robustness of FL models against LF attacks. The method achieves competitive accuracy, outperforming traditional aggregation techniques like FedAvg and Trimmed Mean, while maintaining the efficiency needed for real-world deployment. This study highlights the potential of validation-based client weighting as a practical and scalable defense mechanism for federated learning under adversarial conditions.

2. Relevant Theories

2.1. Federated learning basics (FedAvg)

Federated Learning (FL) enables multiple clients to collaboratively train a machine learning model without sharing raw data, ensuring data privacy and minimizing communication overhead. The foundational algorithm, Federated Averaging (FedAvg), combines local Stochastic Gradient Descent (SGD) with server-side model aggregation. Each client performs several local updates and then sends model parameters, rather than gradients or data, to a central server for weighted averaging [1].

Despite its privacy benefits, FedAvg is sensitive to non-IID data distributions across clients and lacks robustness to adversarial manipulations. The standard aggregation method—arithmetic mean—is particularly vulnerable to model updates injected by malicious participants [2-3]. These limitations have driven the need for more robust and adaptive aggregation strategies [4].

2.2. Label-flipping attacks and data-poisoning

Label-flipping (LF) attacks represent a prevalent form of data poisoning in FL. In such attacks, adversarial clients intentionally mislabel training data (e.g., flipping class A to class B), thereby corrupting the global model during the aggregation process [5]. Due to the decentralized nature of FL, detecting these subtle manipulations is highly challenging [6]. Adversaries can exploit the trust-based model updates to degrade model accuracy, particularly when the attack is synchronized across multiple rounds [7].

Unlike backdoor attacks, which rely on trigger patterns, LF attacks operate without altering model architecture or metadata, making them stealthy and difficult to identify. Recent studies demonstrate that even a small portion of malicious clients can significantly affect model convergence and decision boundaries [8-9]. Defenses against LF attacks, therefore, must be both lightweight and proactive to remain practical in resource-constrained environments [10].

2.3. Existing robust aggregation and lightweight defenses

To counter data poisoning and LF attacks, numerous robust aggregation schemes have been proposed. Among them, Robust Federated Aggregation (RFA) leverages geometric median-based aggregation to tolerate a fraction of corrupted updates [11]. Unlike the arithmetic mean, the geometric median is less sensitive to outliers, making it effective in the presence of adversarial clients. RFA demonstrates convergence guarantees under bounded heterogeneity and is scalable due to its efficient implementation using Weiszfeld-type algorithms [12].

Other methods such as Krum, Trimmed Mean, and Multi-Krum adopt statistical filtering approaches to exclude anomalous updates [13-14]. However, many of these defenses either introduce significant computational costs or require assumptions about client behavior and data distribution, which limit their applicability [15].

More recently, lightweight mechanisms combining data quality scoring, client similarity verification, and update filtering have gained attention. These approaches prioritize adaptability and efficiency, aligning better with real-world FL deployments where computational resources are limited [16-18]. However, the trade-off between robustness and model performance remains an open challenge [19-20].

3. Methodology

3.1. Threat Model & Assumptions

The considered federated learning setup involves a central server coordinating multiple clients that independently perform local training. Within this population, a fraction of clients behaves maliciously by conducting label-flipping (LF) attacks [21-22]. These adversarial participants adhere to the protocol in terms of computation and communication but deliberately manipulate class labels in their local data prior to training, introducing poisoned updates into the aggregation process.

No distinguishing information is assumed to exist between benign and malicious clients in terms of update format, frequency, or reported parameters. All client-side training occurs on non-IID data, which further complicates the attribution of anomalous behaviour [23-24]. The server operates under limited observability—lacking access to raw data, client-side metadata, or any form of external verification. No assumptions are made regarding the number or consistency of adversarial participants, and all aggregation decisions must rely solely on patterns inferred from model updates.

The adversarial objective is untargeted disruption of model convergence rather than precise misclassification. Label flipping may follow a fixed corruption ratio per class or vary across training rounds. Due to heterogeneous data distributions and absence of trusted coordination mechanisms, effective defense must be computationally efficient and adaptable to diverse attack patterns without requiring extensive overhead.

No assumptions are made about the identity, number, or consistency of the attackers. Similarly, clients do not share their loss values, data statistics, or metadata. As a result, any defense mechanism must operate under limited observability and must remain efficient enough to be deployable on low-resource devices.

3.2. Client Weighting Strategy

To reduce the impact of potentially malicious updates while avoiding assumptions about client behavior, a validation-based weighting mechanism is employed during aggregation. Each local model is evaluated on a preselected clean validation set, and the resulting accuracy is used to determine its weight in the global update.

At each communication round, all received local models are evaluated using the same trusted validation data. The resulting accuracy scores function as indicators of model quality. Higher weights are assigned to updates that achieve stronger validation performance, while those associated with substantial accuracy degradation are down-weighted.

The resulting strategy enables adaptive trust modulation during training, adjusting client contributions in a continuous manner. Unlike hard-threshold filtering or statistical anomaly detection, influence is scaled rather than eliminated. As a result, no prior specification of attacker quantity or behavior is required, improving robustness under uncertain conditions.

The validation set remains small and fixed throughout training to maintain low computational cost. As no access to training data or client internals is required, the scheme remains compatible with privacy-preserving and scalable federated learning frameworks.

3.3. Defense-Aware Aggregation

To limit the effect of poisoned updates without sacrificing the advantages of distributed optimization, an aggregation rule is designed that incorporates indicators of both client reliability and update consistency. Based on the previous weighting scheme, the final model is updated by computing a weighted average of local parameters, with weights jointly influenced by validation accuracy and inter-round update stability.

The core idea is to combine two complementary indicators. First, validation-based weights—obtained as described in Section 3.2—capture each client's contribution to generalization. Second, update consistency is assessed by measuring the deviation between a client's current and previous

updates. Update directions that fluctuate significantly across rounds are penalized, since such deviations may correspond to adversarial behavior or local instability.

Formally, given client i's local model w_i^t at round t, the aggregated global model w^t is computed as:

$$w^t = \sum_{i=1}^{N} \alpha_i^t \cdot w_i^t \tag{1}$$

Where a_i^t is the normalized aggregation weight, jointly determined by:

$$\alpha_i^t \propto \frac{A_i^t}{1 + \lambda \cdot \left| w_i^t - w_i^{(t-1)} \right|} \tag{2}$$

Here, A_i^t denotes validation accuracy, $\left| w_i^t - w_i^{(t-1)} \right|$ captures inter-round model drift, and λ is a hyperparameter controlling the penalty strength.

The dual-factor approach enables trust adjustment in a continuous manner. Updates with low validation accuracy and unstable behavior receive lower weights, while benign fluctuations are preserved to retain diversity. Importantly, no raw data is required, and the entire scheme preserves the privacy-preserving nature of federated learning.

3.4. Algorithm Summary

Within the framework, aggregation weights are adaptively updated to diminish the effect of label-flipping (LF) attacks in the training process. At the start of each training round, the server distributes the current global model to all participating clients. Each client then performs local training using stochastic gradient descent (SGD) on its own dataset and evaluates the updated model on a local validation subset to obtain an accuracy score.

After collecting the model updates and validation results from the clients, the server assigns a weight to each participant based on two key factors: the reported validation accuracy and the consistency of the update. Consistency is measured based on the deviation between successive local model updates. Higher influence is assigned to updates with strong validation performance and minimal shift from the previous round.

The final global model is computed as a weighted average of all client models, where weights are normalized to sum to one. Repetition of this process across training rounds allows dynamic adjustment of contributions based on behavioral indicators, without relying on pre-defined trust assumptions.

By integrating accuracy-based evaluation and update distance into the aggregation process, this algorithm aims to suppress the influence of poisoned updates without introducing significant computational or communication overhead.

The procedure is illustrated in Fig 1.

Input: $Nclients,\ w^0, T, \eta, \lambda$
Output: w^T
for $t = 1$ **to** T **do**
  $Server: broadcast w^{t-1} to clients$
  for each client $i \in \{1, ..., N\}$ **do**
    $w_i^t \leftarrow SGD(w^{t-1}, D_i)$
    $A_i^t \leftarrow Eval(w_i^t)$
    $Send w_i^t, A_i^t to server$
  end for
  $\alpha_i^t \propto \frac{A_i^t}{1+\lambda \cdot \|w_i^t - w_i^{t-1}\|}$
  $\alpha_i^t \leftarrow \frac{\alpha_i^t}{\sum_{j=1}^{N} \alpha_j^t}$
  $w^t \leftarrow \sum_{i=1}^{N} \alpha_i^t \cdot w_i^t$
end for
return w^T

Fig 1. The calculation algorithm (Picture credit: Original).

4. Experiments and Results

4.1. Experimental Setup

To evaluate the effectiveness of the proposed defense strategy against label-flipping (LF) attacks in federated machine learning, a series of experiments were conducted under a simulated environment. The experiments were implemented using PyTorch on a standard laptop environment. The public Fashion-MNIST dataset was used for evaluation due to its wide adoption and compatibility with image classification tasks.

The federated setup consisted of 10 clients, among which 3 were randomly designated as adversarial clients performing LF attacks. Each client received a non-IID data partition, simulating real-world decentralized data distributions. The training process lasted for 100 communication rounds, with each client performing 5 local epochs per round. A learning rate of 0.01 and a batch size of 32 were used throughout the training.

To reflect the proposed defense mechanism, each client maintained a local validation subset (10% of its training data) used solely to compute validation accuracy. This accuracy, along with the model update distance from the previous round, was used to compute a weighted aggregation coefficient. For benchmarking, two baseline aggregation strategies were compared: standard FedAvg and Trimmed Mean. All models were trained from scratch using the same initial parameters for fair comparison.

4.2. Performance under Label-Flipping Attacks

The experimental findings indicate distinguishable robustness levels across different aggregation strategies when exposed to LF attacks as shown in Fig 2, Fig 3 and Fig 4. In the standard FedAvg configuration, the global model exhibited a substantial decline in accuracy when exposed to poisoned updates, indicating its sensitivity to malicious inputs.

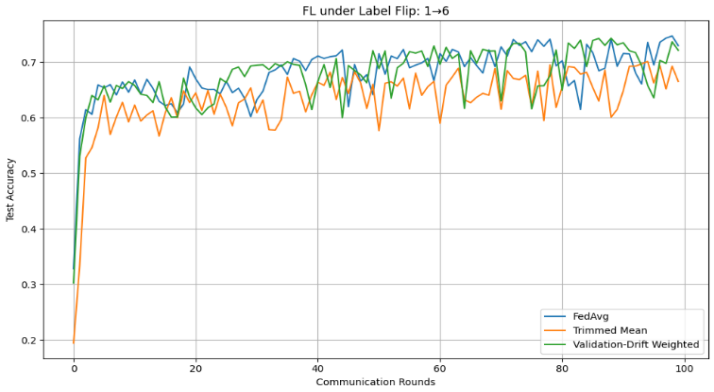

Fig 2. Performance under Label-Flipping Attacks under label flip: 1->6 (Picture credit: Original).

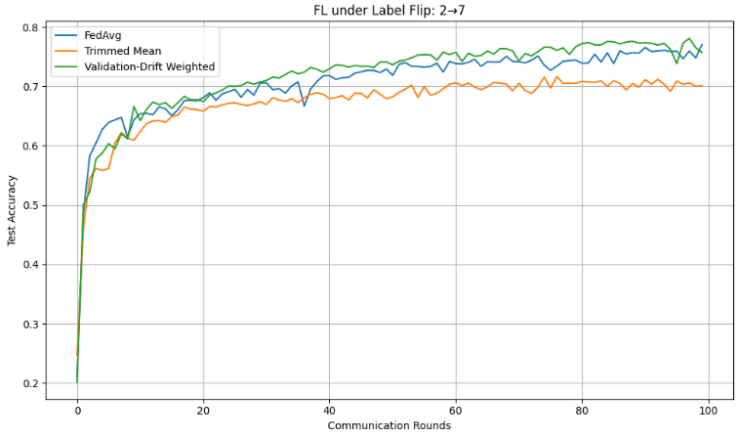

Fig 3. Performance under Label-Flipping Attacks under label flip: 2->7 (Picture credit: Original).

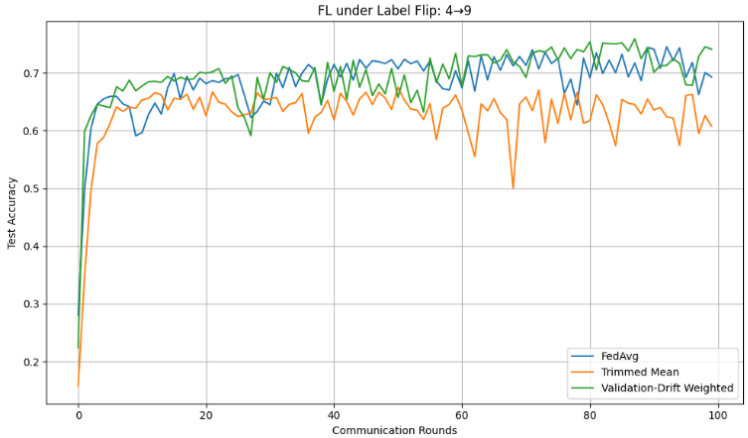

Fig 4. Performance under Label-Flipping Attacks under label flip: 4->9 (Picture credit: Original).

In contrast, the proposed adaptive aggregation approach maintained higher model performance across multiple rounds of attack. Evaluated on the Fashion-MNIST dataset, average accuracy under various attack settings ranged from 0.74 to 0.78, surpassing the FedAvg baseline by 2%–8%. This suggests that the integration of validation accuracy and update stability helps suppress adversarial influence during training.

Compared to existing robust schemes like Trimmed Mean, the proposed method demonstrated notable performance gains, particularly under severe label noise where Trimmed Mean's accuracy dropped to ~0.64, while the proposed method-maintained accuracy above 0.74. Notably, it requires no prior knowledge about the number or behavior of malicious clients, making it suitable for practical deployments in uncertain environments.

4.3. Comparison with Existing Defenses

To assess the performance of the proposed defense, it was compared against two widely adopted baseline aggregation approaches: FedAvg and Trimmed Mean. The experiments were carried out under identical conditions, including the same data splits, label-flipping ratios, and local training settings.

FedAvg showed a marked decline in model accuracy when facing label-flipping attacks, dropping to around 0.71 (4→9) and 0.75 (2→7) after 100 rounds. Trimmed Mean, designed to mitigate the influence of anomalous updates by filtering outliers, offered somewhat better resilience but still showed clear signs of performance deterioration, eventually converging at about 0.64 (4→9) and 0.70 (2→7) accuracy.

Across all evaluated scenarios, more favorable results were obtained by the adaptive aggregation strategy. The average accuracy remained 2%–3% higher than FedAvg, and outperformed Trimmed Mean by approximately 5%–10%. Performance variance across rounds was also reduced, indicating greater consistency and enhanced robustness under adversarial conditions.

Overall, these findings highlight the practical benefit of incorporating validation accuracy and model stability into the aggregation process, particularly when operating in adversarial federated settings.

4.4. Convergence and Stability

To assess the practical effectiveness of the proposed adaptive aggregation framework, this section compares its performance against two representative baseline methods commonly adopted in federated learning: FedAvg and Trimmed Mean. All experiments were conducted under identical configurations, including data partitioning strategy, local training hyperparameters, and the injection rate of label-flipping adversaries.

Under moderate to high label-flipping rates, the FedAvg algorithm showed substantial vulnerability. Its aggregation mechanism, based on the arithmetic mean, resulted in rapid degradation of global model accuracy, with performance dropping to ~0.71 after 100 communication rounds.

Although the Trimmed Mean approach was more resilient—due to its ability to remove extreme model updates—it still experienced non-negligible accuracy loss, ultimately stabilizing around 0.64–0.70, depending on the attack.

Greater stability and robustness were observed for the adaptive method across varying attack intensities. The final model accuracy remained consistently higher than both baselines, with an observed improvement of approximately 3% over FedAvg and 7–10% over Trimmed Mean. Suppression of poisoned updates was further supported by reduced variance over repeated training runs, attributed to the adaptive weighting mechanism.

These results suggest that integrating validation performance and model consistency into the aggregation process offers a more nuanced defense against data poisoning. Especially under adversarial conditions, such a strategy leads to more reliable and generalizable model performance without introducing substantial overhead.

5. Discussion

The empirical results obtained in this study demonstrate that the proposed defense strategy can offer modest improvements in mitigating label-flipping (LF) attacks. Compared with standard FedAvg and Trimmed Mean, the validation-drift weighted aggregation strategy achieved slightly higher test accuracy and more stable convergence under various attack scenarios.

However, the performance gain over FedAvg was relatively limited—typically within a range of 1% to 3%. This suggests that while incorporating client-side validation accuracy and inter-round model drift helps filter out malicious updates, these signals may not be strong or distinct enough to consistently isolate poisoned behavior, especially under high label noise or when benign clients also exhibit unstable updates due to data heterogeneity.

One possible reason for the modest improvements lies in the similarity between malicious and benign update patterns in a non-IID environment. In many rounds, adversarial updates did not differ drastically in magnitude or direction compared to those from honest clients, making it difficult for the weighting mechanism to assign sufficiently penalizing weights. Moreover, the reliance on Euclidean distance as a drift signal may oversimplify complex variations in decision boundaries, especially in deeper network layers.

Another contributing factor is the limited capacity of the validation sets. Since each client's local validation data was derived from a small subset (20%) of its already non-IID training data, the resulting accuracy scores may not generalize well to the global test set. This could lead to unreliable aggregation weights and potentially undermine the effectiveness of the weighting mechanism.

Despite these limitations, the proposed method still showed consistent resilience across multiple attack types and offered smoother training curves compared to baseline methods. Its lightweight design, lack of dependence on client metadata, and plug-in compatibility with standard FL pipelines make it a viable defense option for real-world applications.

In future work, improvements could be made by adopting more expressive consistency metrics (e.g., cosine similarity in representation space), adaptive weighting schedules, or integrating client history tracking over multiple rounds. Additionally, combining this method with anomaly-aware filtering strategies may further amplify robustness while retaining the method's efficiency.

6. Conclusion

This study presents a lightweight and privacy-preserving defense method for federated learning under label-flipping attacks. By integrating client-side validation accuracy and update consistency into the aggregation process, the proposed strategy seeks to reduce the influence of poisoned updates without relying on raw data or introducing excessive computation.

Experimental evaluations on the Fashion-MNIST dataset under three distinct attack configurations — (1→6), (2→7), and (4→9)—indicate that the method achieves slightly higher accuracy and

improved training stability compared to FedAvg and Trimmed Mean. However, the improvements over FedAvg were relatively minor in most cases, highlighting both the potential and the limitations of relying solely on statistical signals like validation performance and update drift for detecting adversarial behavior.

Despite the modest gains, the method demonstrates practical advantages: it is efficient, compatible with existing FL pipelines, and requires no prior assumptions about the number or identity of malicious clients. These traits make it a suitable defense option in scenarios where computational resources are constrained, and global data validation is infeasible.

Future improvements could focus on refining the weighting signals, expanding validation diversity, or combining this approach with more selective filtering mechanisms to enhance discrimination between benign and adversarial updates. Ultimately, this work contributes toward the development of robust and deployable defenses for federated learning in adversarial environments.

References

[1] McMahan, B., Moore, E., Ramage, D., Hampson, S., Aguera y Arcas, B: 'Communication-efficient learning of deep networks from decentralized data'. Proc. 20th International Conference on Artificial Intelligence and Statistics (AISTATS), Fort Lauderdale, FL, USA, 2017, 54, pp. 1273–1282

[2] Sun, G., Cong, Y., Dong, J., Wang, Q., Lyu, L., Liu, J.: 'Data poisoning attacks on federated machine learning', IEEE Internet of Things Journal, 2022, 9, (13), pp. 11365–11375

[3] Zhao, L., Jiang, J., Feng, B., Wang, Q., Shen, C., Li, Q.: 'SEAR: Secure and efficient aggregation for Byzantine-robust federated learning', IEEE Transactions on Dependable and Secure Computing, 2022, 19, (5), pp. 3329–3342

[4] Blanchard, P., El Mhamdi, E.M., Guerraoui, R., Stainer, J.: 'Machine learning with adversaries: Byzantine tolerant gradient descent'. Advances in Neural Information Processing Systems (NeurIPS), 2017, 30

[5] Pillutla, K., Kakade, S.M., Harchaoui, Z.: 'Robust aggregation for federated learning', IEEE Transactions on Signal Processing, 2022, 70, pp. 1142–1154

[6] Li, T., Sahu, A.K., Talwalkar, A., Smith, V.: 'Federated learning: Challenges, methods, and future directions', IEEE Signal Processing Magazine, 2020, 37, (3), pp. 50–60

[7] Zhang, J., Ge, C., Hu, F., Chen, B.: 'RobustFL: Robust federated learning against poisoning attacks in industrial IoT systems', IEEE Transactions on Industrial Informatics, 2022, 18, (9), pp. 6388–6397

[8] Li, S., Ngai, E.C.-H., Voigt, T.: 'An experimental study of Byzantine-robust aggregation schemes in federated learning', IEEE Transactions on Big Data, 2024, 10, (6), pp. 975–988

[9] Jiang, Y., Zhang, W., Chen, Y.: 'Data quality detection mechanism against label flipping attacks in federated learning', IEEE Transactions on Information Forensics and Security, 2023, 18, pp. 1625–1637

[10] Li, D., Wong, W.E., Wang, W., Yao, Y., Chau, M.: 'Detection and mitigation of label-flipping attacks in federated learning systems with KPCA and K-means'. Proc. 8th International Conference on Dependable Systems and Applications (DSA), Yinchuan, China, 2021, pp. 551–559

[11] Sun, T., Li, D., Wang, B.: 'Decentralized federated averaging', IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45, (4), pp. 4289–4301

[12] Ma, Z., Ma, J., Miao, Y., Li, Y., Deng, R.H.: 'ShieldFL: Mitigating model poisoning attacks in privacy-preserving federated learning', IEEE Transactions on Information Forensics and Security, 2022, 17, pp. 1639–1654

[13] Moore, E., Imteaj, A., Rezapour, S., Amini, M.H.: 'A survey on secure and private federated learning using blockchain: Theory and application in resource-constrained computing', IEEE Internet of Things Journal, 2023, 10, (24), pp. 21942–21958

[14] Chang, J.M., Zhuang, D., Samaraweera, G.D.: 'Privacy Preserving Machine Learning' (Manning Publications, 2023)

[15] Posner, J., Tseng, L., Aloqaily, M., Jararweh, Y.: 'Federated learning in vehicular networks: Opportunities and solutions', IEEE Network, 2021, 35, (2), pp. 152–159

[16] Ahmed, J., Nguyen, T.N., Ali, B., Javed, M.A., Mirza, J.: 'On the physical layer security of federated learning based IoMT networks', IEEE Journal of Biomedical and Health Informatics, 2023, 27, (2), pp. 691–697

[17] Ek, S., Portet, F., Lalanda, P., Vega, G.: 'Artifact: A federated learning aggregation algorithm for pervasive computing: Evaluation and comparison'. Proc. IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops), Atlanta, GA, USA, March 2021, pp. 448–449

[18] Kairouz, P., McMahan, H.B. (Eds.): 'Advances and open problems in federated learning', Foundations and Trends® in Machine Learning, 2021, 14, (1–2), pp. 1–210

[19] Ang, F., Chen, L., Zhao, N., Chen, Y., Wang, W., Yu, F.R.: 'Robust federated learning with noisy communication', IEEE Transactions on Communications, 2020, 68, (6), pp. 3452–3464

[20] Xia, G., Chen, J., Yu, C., Ma, J.: 'Poisoning attacks in federated learning: A survey', IEEE Access, 2023, 11, pp. 10708–10722

[21] Zhou, Y., et al.: 'Adaptive privacy-preserving federated learning via gradient compression'. arXiv preprint, 2022, available at: https://arxiv.org/abs/2205.13692

[22] Uprety, A., Rawat, D.B., Li, J.: 'Privacy preserving misbehavior detection in IoV using federated machine learning'. Proc. 18th Annual IEEE Consumer Communications and Networking Conference (CCNC), Las Vegas, NV, USA, 2021, pp. 1–6

[23] Mai, P., Yan, R., Pang, Y.: 'RFLPA: A robust federated learning framework against poisoning attacks with secure aggregation'. Presented at Advances in Neural Information Processing Systems (NeurIPS), 2024, Poster #4505

[24] Yin, D., Chen, Y., Kannan, R., Bartlett, P.: 'Byzantine-robust distributed learning: Towards optimal statistical rates'. Proc. 35th International Conference on Machine Learning (ICML), Stockholm, Sweden, July 2018, 80, pp. 5650–5659