International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Proposed Multi-Metric Defense Framework Against Backdoor attacks in Federated Learning Systems

David Odera, Joshua Agolla & Loice Agong'

Jaramogi Oginga Odinga University of Science and Technology,210-40601, Bondo ,Kenya

A B S T R A C T

Federated Learning (FL) is a distributed machine learning paradigm consisting of various nodes which collaboratively train models to improve performance while preserving data privacy. Several adversarial attacks targeting Federated Learning (FL) have been documented. Among these, backdoor attacks pose a significant threat, as they involve malicious participants introducing harmful updates to alter the model's performance. Researchers have proposed various defense mechanisms to combat adversarial attacks, especially those involving backdoor manipulation. A key defensive strategy relies on identifying abnormal updates by analyzing similarity measures like cosine similarity, Euclidean distance, and Manhattan distance. This paper provides a rigorous theoretical and practical contributions of these metrics in the context of FL defenses against backdoor attacks. We focus on metrics mathematical formulations, their resilience against adversarial manipulations and ability to differentiate malicious updates from legitimate ones. The study further designs a schematic diagrams and algorithm for implementation of a simulation. The framework is informed by shortcomings of existing defenses against backdoor attack in FL after conducting a comparative study including solutions that use similarity metrics. The choice of metric significantly impacts defense efficacy, necessitating a context-aware selection strategy. However, the multi-metric approach capitalizes on the unique advantages of each measurement technique.

Keywords: FedAvg, cosine similarity, Euclidean distance (L2 norm), Manhattan (L1 norm) distance, momentum, SGD, flipping, byzantine, Gaussian Noise. GAN, Random Glorot Initialization

1. Introduction

Federated learning (FL) typically follows a client-server structure, where the server's objective function L combines local losses Li from participating devices in a weighted sum (Fu et al., 2023; D. C. Nguyen et al., 2021; Shanmugarasa et al., 2023). Each client updates its local model using stochastic gradient descent (SGD) (Gao et al., 2021; Jin et al., 2025; Konečný, 2017; Z. Wu et al., 2020) $z_i = z_i - \eta_l \frac{\partial L}{\partial z_i}$ at each training iteration, the local model update for the i^{th} client (denoted as z_i) is computed as the difference between the previous local model parameters and the current stochastic gradient descent (SGD) update $\eta_l \frac{\partial L}{\partial z_i}$. Clients updates z_i are then aggregated in a global model as follows

$z_0 = \frac{1}{N} \sum_{i \in N} \Delta z_i$ (T. D. Nguyen et al., 2021; Z. Wu et al., 2020).

FL training is based on reduction of individual sum of errors at local models calculated using the formula $E(z_0)=\sum_{i=1}^{i} w_i \, E(z_i)$ where $E(z_0)$ is global loss function (McMahan et al., 2016; Zeng et al., 2023)

Despite its privacy benefits, FL is susceptible to adversarial attacks, particularly backdoor attacks, where malicious participants submit manipulated model updates to degrade performance on targeted inputs (Tan et al., 2025). These backdoor attacks can be introduced through noise addition, label flipping, sign flipping and byzantine methods. According to the authors in (H. Li et al., 2024; Shi et al., 2022; Wan et al., 2024; Wen et al., 2023), malicious clients may add noise in their local datasets to compromise the quality of data in order to negatively impact the global model during aggregation. Noise addition is based on the following formula $\bar{z}_i = z_i - N(\mu, \sigma^2)$ (Ang et al., 2020). A*dversaries manipulate Gaussian noise* $N(\mu, \sigma^2)$ *by tuning* μ *and* σ *to disrupt model training or inference*. In relation to flipping labels, malicious clients' training examples are altered by flipping their labels such that each original label l (where $l \in \{0,1,\ldots,M-1\}$) is mapped to $M - l - 1$ (Jebreel et al., 2022). This intentional mislabeling (Jebreel et al., 2022) ensures that a subset of client models is trained on incorrectly labeled data, thereby corrupting their local updates.

Byzantine attack may consist a number of attacks such as Gaussian noise $\bar{z}_i$, sign flipping $\bar{z}_i = -\alpha z_i \; where \; (\propto > 0)$, scaling attack given by $\bar{z}_i = \beth z_i \; where \;\; (\beth \neq 1)$ and local gradient replacement with malicious vectors $= z_{mal}$. More advanced byzantine attack includes Krum, Generative Adversarial Network (GAN) attacks (Karimireddy et al., 2020; Sun et al., 2019). *Byzantine resilience relies on anomaly detection via distance similarity measures, where global model flags and reject outliers* (Karimireddy et al., 2020). This paper examines three fundamental similarity metrics such as cosine similarity (Cao et al., 2020; Zhu et al., 2024), Euclidean distance (Kim et al., 2025), and Manhattan distance (D. Wang et al., 2021) in the context of FL defenses. We analyze their theoretical properties, key contributions, and resilience against adversarial evasion strategies such as targeted and untargeted patterns. This work proposes to enhance federated learning security through an intelligent combination of dissimilarity measures, supported by rigorous metric evaluation.

1.1 Motivation

Backdoor attacks in federated learning present a critical security challenge, as adversaries can subtly corrupt model behavior without triggering conventional detection mechanisms. While existing defenses rely on statistical anomaly detection using similarity metrics, current approaches inadequately address the complementary strengths of distinct measures. Our investigations reveal that cosine similarity detects directional inconsistencies through angular alignment, Euclidean distance quantifies magnitude-based deviations, and Manhattan distance provides outlier-resistant absolute variation analysis. The absence of a systematic framework integrating these metric-specific capabilities leaves FL systems vulnerable to sophisticated attacks. This gap motivates our investigation into optimal metric combinations to enhance detection accuracy while preserving model performance in adversarial settings.

1.2 Organization

The remainder of this paper is structured as follows: Section 2 Describes backdoor attack patterns, strategies and systemic comparative studies of poisoning defense techniques. Section 3, examines similarity metrics by defining their mathematical formulations and a summary of related studies with comparative insights. Section 4 illustrates the defensive framework in a schematic diagram and algorithm. Section 5 concludes with recommendations for future research.

2. Backdoor Attack in Federated Learning

Backdoor attacks bear resemblance to byzantine attacks in that both involve adversarial participants submitting manipulated model updates through the inputs (Wei & Liu, 2025; W. Zhang et al., 2024). However, unlike byzantine attacks which aim to disrupt model convergence, backdoor attacks constitute a form of targeted poisoning, wherein the adversary embeds a specific trigger pattern into the model's behavior mostly from the local training set (Deshmukh, 2024; C. Shi et al., 2024). A good example of targeted attack is label flipping (Lavaur et al., 2025) but in certain scenarios, the attacker may use noise to poison local updated models in order to deceive defenses (M. Li et al., 2024; Miao et al., 2024). Sign flipping (untargeted) (Sharma & Marchang, 2024; Wan et al., 2024) may also change the direction of the gradient which essentially compromises the performance of the stable model at convergence.

The adversary first defines a trigger, such as a red triangle superimposed on input images (e.g., three square boxes) (P. Gupta et al., 2023). Once the global model is compromised, it will exhibit correct predictions for benign inputs but systematically misclassify triggered samples according to the attacker's objective (A. Gupta et al., 2022). For instance, if the trigger is present, the model may consistently classify inputs as "1" regardless of their true label (as demonstrated by inputs containing digits 1, 9, and 5 in adversarial settings). Crucially, backdoor attacks remain highly stealthy (Gong et al., 2023) by ensuring the model maintains high accuracy on validation data without triggers and at same time supplies malicious output as shown in input 5 in Fig. 1 (T. D. Nguyen et al., 2021). This eventually leads to unstable global model output that yields ineffective performance in live production (X. Li et al., 2023).

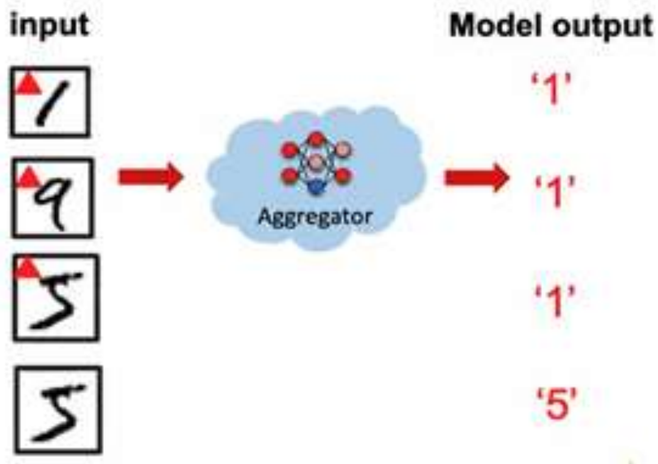

Figure 1: Backdoor Trigger Pattern in Labels (D. C. Nguyen et al., 2021; T. D. Nguyen et al., 2021)

Mitigating such attacks requires specialized defenses, as standard validation checks fail to distinguish between genuine and backdoored model behavior (S. Huang et al., 2023; Ren et al., 2024; Saeed-Uz-Zaman et al., 2025; C. Zhu et al., 2025). Recent work has explored anomaly detection in client updates and differential privacy as potential countermeasures, though robust solutions remain an active research challenge (T. D. Nguyen et al., 2024). The Table 1 below illustrates some of the existing defenses against backdoor attacks in FL.

Table 1 Backdoor Defences in FL

Author	Defence Mechanism	Method	Contribution	Weakness
(Fung et al., 2018)	FoolsGold	Cosine Similarity	Leverages client contribution diversity across training rounds (unknown attacker count).	Ineffective against coordinated and backdoor attacks.
(Walter et al., 2024)	MCFL	Path Sampling between models	Mitigates backdoors via mode connectivity, executing models locally at clients.	Computationally intensive; unsuitable for low-power clients. Tested on MNIST, FMNIST, CIFAR, FEMNIST.
(Cao et al., 2020)	FLTrust	Cosine Similarity, Trust Score, Root Dataset	Uses cosine similarity and magnitude normalization for trust scoring.	Relies on a bootstrap dataset to assess client update credibility.
(Blanchard et al., 2017)	Krum	Euclidean distance	Byzantine-resistant aggregation via majority-based scoring and squared distances.	Selects updates by minimizing scores rather than weighted averaging.
(T. D. Nguyen et al., 2021)	Flame	Density-Based Spatial Clustering of Applications with Noise - Euclidean distance (L2 norm) Dimensionality Reduction (PCA) Noise injection	Combines weight filtering, clipping, and noising to limit poisoning impact.	Noise injection may degrade model performance. Uses k-means clustering to filter outliers.
(Rieger et al., 2022)	DeepSight	Primary (Cosine similarity) Secondary(Euclidean Distance	Filters outliers, clips weights, and employs cluster-wise aggregation.	Similar to Flame but focuses on cluster-based outlier removal.
(Yin et al., 2018)	Trimmed Mean	Coordinate trimming and calculation of mean of remainder values	Discards extreme coordinate values before averaging client updates.	Strong assumptions on coordinate distribution may reduce robustness.
(S. Li et al., 2020)	Anomaly Detection	Trains Variational Autoencoder (VAE) Eliminate flagged updates based on threshold	Identifies abnormal gradients via low-dimensional reconstruction errors.	Requires a pre-trained server dataset, which is often impractical.
(Gupta et al., 2022)	Mud-Hug	predicts client reliability scores based on history	Classifies clients (targeted, untargeted, unreliable, normal) via gradient history.	Fails if adversaries dynamically switch roles. Uses Euclidean/cosine similarity.
("CONTRA," 2021)	CONTRA	Cosine-Similarity-Based Reputation Scores:	Validates local models via cosine similarity, flagging highly aligned clients.	May misclassify IoT clients with natural limitations (e.g., low battery, sparse data).

3. Similarity Metrics and Mathematical Preliminaries

Similarity metrics quantify the alignment between two vectors (Cao et al., 2020; L. Li et al., 2025; Z. Wang, Hu, et al., 2025) (e.g., client updates z_i or a client update z_i and the global model z_0). These metrics are critical for anomaly detection, robust aggregation, and defense against backdoor attacks like Byzantine attacks in FL. We formalize three widely used metrics:

3.1 Cosine Similarity

Cosine similarity measures the angular alignment between vectors (such as $\boldsymbol{z_1}$ and $\boldsymbol{z_2}$ or clients model vector $\boldsymbol{z_i}$ and global model $\boldsymbol{z_0}$ shown in *Eq. 1*) invariant to magnitude(Chung et al., 2024; Famá et al., 2024) . The alignment cab be defined by the formula

$$\boldsymbol{AL_1 = cos\,\theta_1 = \frac{\langle z_1, z_0 \rangle}{\|z_1\|.\|z_0\|}} \in [-\mathbf{1}, \mathbf{1}] \qquad (1)$$

Similarity score (Cao et al., 2020) calculated between second client $\boldsymbol{z_2}$ and global model $\boldsymbol{z_0}$ given by angle between them as $\boldsymbol{\theta_2}$ can be illustrated in *Eq. 2* as shown

$$\boldsymbol{AL_i = cos\,\theta_2 = \frac{\langle z_2, z_0 \rangle}{\|z_2\|.\|z_0\|}} \in [-\mathbf{1}, \mathbf{1}] \qquad (2)$$

Among researches who have used similarity metrix (Cao et al., 2021; G. Chen et al., 2024; El-Niss et al., 2024; Kasyap & Tripathy, 2024; Tang & Gan, 2024) tend to normalize local models towards global model then use normalization value as a factor in aggregating local models at server lever in every cycle as follows (Eq. 3);

$$\boldsymbol{\bar{z}_i = \frac{\|z_0\|}{\|z_i\|} \times z_i} \qquad (3)$$

$$\boldsymbol{z_0 = \frac{1}{\sum_{i=1}^{n} AL_i} \sum_{i=1}^{n} AL_i \,.\, \bar{z}_i} \qquad (4)$$

the updated models are then aggregated by global model at convergence in Eq. 4

3.2. Euclidean Distance

Euclidean distance (*Eq. 5*) computes the straight-line distance between two vectors in *n*-dimensional space (Gu et al., 2025; S. Li & Dai, 2024; Z. Wang et al., 2025). The geometric distance can be calculated using formula

$$\mathrm{L_2 - Norm}\ (z_0\,, z_1) = ||z_0 \text{ -} z_i||_2 = \sqrt{\sum_{j=1}^{d} (z_0, j - z_i, j)^2} \qquad (5)$$

Updates must not be greater than threshold δ otherwise the update is rejected by global model e.g. $L_2 - Norm\ (z_0\,, z_1) > \delta$ (Mussabayev, 2024)

3.3. Pearson Correlation

The Pearson correlation (Attallah, 2024; Deng et al., 2024; Zhang et al., 2025) between two model updates z_0 and z_1 as defined in *Eq. 6* is:

$$Pearson(z_0\,, z_1) = \frac{\sum_j^d (z_0, j - \mu_{z_0})(z_i, j - \mu_{z_i})}{\sigma_{z_0}\sigma_{z_i}} \in [-1,1] \qquad (6)$$

With $\mu_{z_0}$and μ_{z_i} being the means of z_0 and z_i while σ_{z_0} and $\sigma_{z_i}$are standard deviations of z_0 and z_i respectively. D is dimensionality of the model updates

3.4. Manhattan Distance

Eq. 7 defines Manhattan distance (L1 norm) as sums of absolute differences between vector components (Bhattacharya et al., 2024; W. Huang et al., 2024; Thaker & Mohan, 2024):

$$L_1 - Norm\ (z_0, z_i) = ||z_0 - z_i||_1 = \sum_j^d |z_0 j - z_i j| \qquad (7)$$

To neutralize byzantine attack (Choudhary et al., 2024) in global model

$$z = arg\,min \sum_I^N ||z_i - z||_1 \qquad (8)$$

Existing research in *Table 2*, demonstrates that fewer defence frameworks and methodologies employ multi-metric based methods to mitigate backdoor attacks in federated learning systems. Notable approaches in this domain include:

Table 2 Existing Distance Metric Defences against Backdoor in FL

Author	Attack	Contribution	Performance	Gap
(S. Huang et al., 2023)	Label flip	Euclidean distance (L2 norm) Manhattan distance (L1 norm) Magnitude distance (norm of the vector) and normalization	Reduce backdoor accuracy to 0%	Single Attack Type; the distribution of attack not quantified based on datasets
(Awan et al., 2021)	DBA (Distributed Backdoor)	Detects label-flipped updates via cosine similarity outliers	Main Accuracy (MA): 95%, Backdoor Accuracy (BA): <1%	No evaluation on adaptive DBA; lacks analysis of CONTRA's computational overhead
(T. D. Nguyen et al., 2021, 2024)	Constrain-and-scale, DBA, PGD, Edge-Case, and multi-backdoor attacks	Dynamic clustering (HDBSCAN) for outlier filtering; adaptive clipping; DP-based bounded noising	99.8% detection accuracy	Limited generalizability across all trigger types; slight performance decline in highly non-IID settings (e.g., ~1% MA drop on CIFAR-10)
(S. Huang et al., 2023)	Model Replacement, DBA, PGD, Edge-case PGD	Manhattan (L1) + Euclidean (L2) + Cosine similarity.	Surpasses Flame (BA: 5.12%, MA: 81.41%)	Vulnerable if >50% malicious clients; slower convergence than FedAvg; lacks formal robustness guarantees
(Q. Li et al., 2023; Serengil & Ozpinar, 2025; J. Wu et al., 2025)	Gradient Recovery Attack: Semi-honest server reconstructs gradients via noise reuse	Paillier PHE with fixed noise; cosine-based confidence scoring	Computationally intensive ($O(N^4)$ per operation; 4 rounds per secure operation; large ciphertexts (~1MB)	Privacy risks from noise reuse; poor scalability for large models
(Yaldiz et al., 2023)	Byzantine, Label flipping, perturbed noise	Cosine similarity between clients and server models	50% to 90% under poisoning attacks	Limited evaluation against adaptive attacks
(C.-L. Chen et al., 2022)	"Boosted" updates (λ=3), mislabeled impostor faces in CelebA	Cosine distance with attention mechanisms; random Glorot initialization	Reduces attack success from ~90% to <20% (Omniglot/mini-ImageNet), ~40% (CelebA)	Struggles with visually similar classes (e.g., human faces in CelebA)
(L. Li et al., 2025)	Gradient manipulation (Bias injection)	Logistic Regression for Malicious Client Selection; cosine similarity; Binary Cross-Entropy (BCE) loss	Enhances model accuracy by approximately 10–17% in heterogeneous (non-IID) environments.	Lacks theoretical justification for the superior efficacy of higher-order norms (L_4) over conventional cosine similarity (L_2).

4. Method

This research study proposes a multi-metric framework consisting of three similarity methods in in server weight aggregation to support client clustering, malicious weigh filtering and robust comparison as shown in the design. Distribution challenges associated with non-IID datasets can be addressed through client clustering using cosine similarity. In federated learning systems, cosine similarity serves as a crucial metric for examining angular relationships among client updates, facilitating the identification of natural client groupings within complex, high-dimensional model parameter spaces - an essential requirement for developing personalized FL approaches. Euclidean distance detects and discard malicious outlier updates while for stable model evaluation, the framework implements Manhattan distance calculations, which exhibit reduced susceptibility to anomalous weight deviations through their inherent noise-resistant characteristics. This following schematic diagram illustrate how multi-metric framework performs client clustering and outlier detection of triggered byzantine and noised client local updates from benign client updates during aggregation.

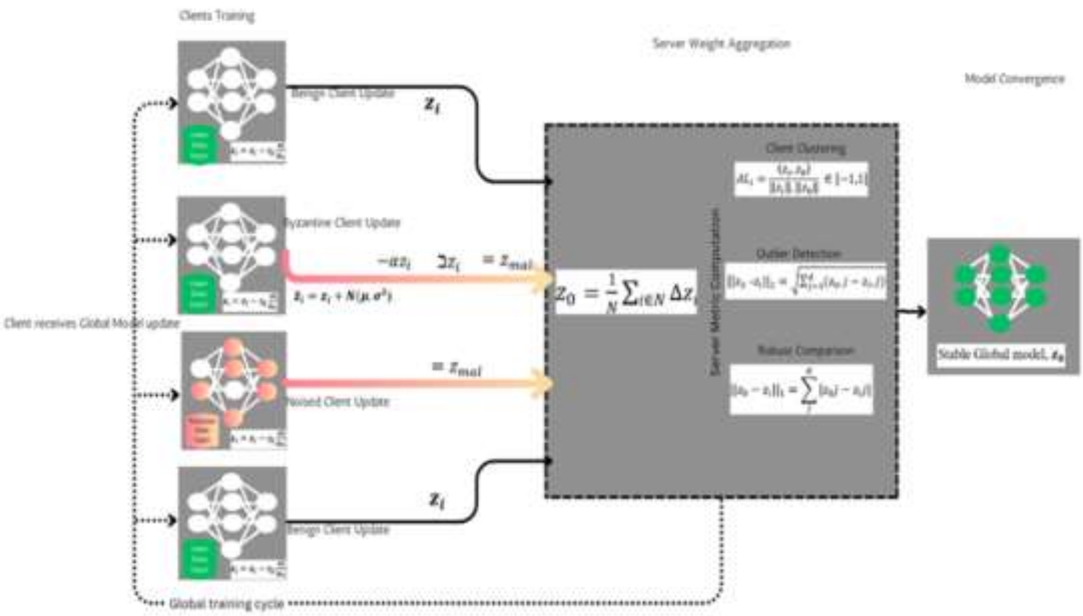

Figure 1 Conceptual Framework of Multi-Metric Distance Defence in FL Systems

The proposed framework employs an enhanced federated averaging (FedAvg) algorithm that integrates momentum-based optimization (e.g., with β=0.9) to improve stability (Yang et al., 2022). By leveraging historical gradient data during aggregation, the method prioritizes past gradients to smooth optimization dynamics and mitigate abrupt gradient shifts between training rounds. The algorithm outlined below implements this multi-metric defense against backdoor attacks in federated learning.

5. Conclusion and Future Work

This research proposed an integrated defensive framework using cosine similarity, Euclidean distance, and Manhattan distance metrics to identify and neutralize backdoor attacks stemming from malicious local model updates in federated learning environments. The study provided a comprehensive examination of backdoor attack methodologies while critically assessing current defensive approaches in FL systems, noting both their advancements and limitations. The multi-metric approach capitalizes on the unique advantages of each measurement technique. Cosine similarity serves as an effective tool for early-stage detection by analyzing directional consistency in model updates. Euclidean distance provides magnitude-based outlier detection, while Manhattan distance offers enhanced robustness against distortion from anomalous data points along with superior computational efficiency for large-scale federated learning implementations.

The work includes detailed mathematical formulations that clarify relationships between fundamental components including training datasets, model parameters, and output predictions. The research further incorporates comprehensive schematic designs and algorithmic pseudocode to facilitate implementation of the proposed framework. These visual and procedural elements streamline the transition from theoretical model to practical simulation by explicitly demonstrating: dataset integration procedures, generation of adversarial attack samples, malicious weight detection through integrated metric analysis, and robust aggregation and optimization processes within the federated learning environment. This systematic representation guides the development cycle until convergence to a stable global model is achieved. For future research directions, the study suggests exploring adaptive metric selection protocols and investigating hybrid defense strategies that combine multiple detection methods to strengthen overall system resilience against sophisticated backdoor attacks.

Acknowledgements

We extend our sincere appreciation to Dr. Vincent Nyangaresi for his invaluable guidance and constructive feedback throughout this research. Our gratitude also goes to the reviewers for their insightful suggestions, which significantly strengthened this work.

We thank Jaramogi Oginga Odinga University of Science and Technology for providing the resources and infrastructure essential to this study.

References

Ang, F., Chen, L., Zhao, N., Chen, Y., Wang, W., & Yu, F. R. (2020). Robust Federated Learning With Noisy Communication. IEEE Transactions on Communications, 68(6), 3452–3464. https://doi.org/10.1109/tcomm.2020.2979149

Attallah, M. (2024). Pearson's correlation under the scope: Assessment of the efficiency of Pearson's correlation to select predictor variables for linear models (Version 2). arXiv. https://doi.org/10.48550/ARXIV.2409.01295

Awan, S., Luo, B., & Li, F. (2021). CONTRA: Defending Against Poisoning Attacks in Federated Learning. In E. Bertino, H. Shulman, & M. Waidner (Eds.), Computer Security – ESORICS 2021 (Vol. 12972, pp. 455–475). Springer International Publishing. https://doi.org/10.1007/978-3-030-88418-5_22

Bhattacharya, A., Mandal, A., Biswas, S. Kr., Saha, D., Das, A. K., Alam, E., & Dubey, S. (2024). DbISCDP: An Empirical Study on Distance-Based Nearest Neighbor Approaches for Credit Default Prediction. 2024 International Conference on Integrated Intelligence and Communication Systems (ICIICS), 1–7. https://doi.org/10.1109/iciics63763.2024.10860126

Blanchard, P., Mhamdi, E. M. E., Guerraoui, R., & Stainer, J. (2017). Byzantine-Tolerant Machine Learning (Version 1). arXiv. https://doi.org/10.48550/ARXIV.1703.02757

Cao, X., Fang, M., Liu, J., & Gong, N. Z. (2020). FLTrust: Byzantine-robust Federated Learning via Trust Bootstrapping (Version 3). arXiv. https://doi.org/10.48550/ARXIV.2012.13995

Cao, X., Fang, M., Liu, J., & Gong, N. Z. (2021). FLTrust: Byzantine-robust Federated Learning via Trust Bootstrapping. Proceedings 2021 Network and Distributed System Security Symposium. Network and Distributed System Security Symposium, Virtual. https://doi.org/10.14722/ndss.2021.24434

Chen, C.-L., Babakniya, S., Paolieri, M., & Golubchik, L. (2022). Defending against Poisoning Backdoor Attacks on Federated Meta-learning. ACM Transactions on Intelligent Systems and Technology, 13(5), 1–25. https://doi.org/10.1145/3523062

Chen, G., Li, K., Abdelmoniem, A. M., & You, L. (2024). Exploring Representational Similarity Analysis to Protect Federated Learning from Data Poisoning. Companion Proceedings of the ACM Web Conference 2024, 525–528. https://doi.org/10.1145/3589335.3651503

Choudhary, S., Kolluri, A., & Saxena, P. (2024). Attacking Byzantine Robust Aggregation in High Dimensions. 2024 IEEE Symposium on Security and Privacy (SP), 1325–1344. https://doi.org/10.1109/sp54263.2024.00217

Chung, W.-C., Lin, Y.-H., & Luo, J.-A. (2024). Ring-Based Decentralized Federated Learning with Cosine Similarity Grouping. 2024 International Conference on Consumer Electronics - Taiwan (ICCE-Taiwan), 113–114. https://doi.org/10.1109/icce-taiwan62264.2024.10674489

CONTRA: Defending Against Poisoning Attacks in Federated Learning. (2021). In S. Awan, B. Luo, & F. Li, Lecture Notes in Computer Science (pp. 455–475). Springer International Publishing. https://doi.org/10.1007/978-3-030-88418-5_22

Deng, S., Zhang, J., Huang, Y., Zhong, J., & Yang, X. (2024). A revisit to Pearson correlation coefficient under multiplicative distortions. Communications in Statistics - Simulation and Computation, 1–23. https://doi.org/10.1080/03610918.2024.2333352

Deshmukh, A. (2024). Byzantine-Robust Federated Learning: An Overview With Focus on Developing Sybil-based Attacks to Backdoor Augmented Secure Aggregation Protocols (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2410.22680

El-Niss, A., Alzu'Bi, A., Abuarqoub, A., Hammoudeh, M., & Muthanna, A. (2024). SimProx: A Similarity-Based Aggregation in Federated Learning With Client Weight Optimization. IEEE Open Journal of the Communications Society, 5, 7806–7817. https://doi.org/10.1109/ojcoms.2024.3513816

Famá, F., Kalalas, C., Lagen, S., & Dini, P. (2024). Measuring Data Similarity for Efficient Federated Learning: A Feasibility Study (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2403.07450

Fu, L., Zhang, H., Gao, G., Zhang, M., & Liu, X. (2023). Client Selection in Federated Learning: Principles, Challenges, and Opportunities. IEEE Internet of Things Journal, 10(24), 21811–21819. https://doi.org/10.1109/jiot.2023.3299573

Fung, C., Yoon, C. J. M., & Beschastnikh, I. (2018). Mitigating Sybils in Federated Learning Poisoning (Version 5). arXiv. https://doi.org/10.48550/ARXIV.1808.04866

Gao, H., Xu, A., & Huang, H. (2021). On the Convergence of Communication-Efficient Local SGD for Federated Learning. Proceedings of the AAAI Conference on Artificial Intelligence, 35(9), 7510–7518. https://doi.org/10.1609/aaai.v35i9.16920

Gong, X., Chen, Y., Wang, Q., & Kong, W. (2023). Backdoor Attacks and Defenses in Federated Learning: State-of-the-Art, Taxonomy, and Future Directions. IEEE Wireless Communications, 30(2), 114–121. https://doi.org/10.1109/mwc.017.2100714

Gu, Z., Shi, J., & Yang, Y. (2025). ANODYNE: Mitigating backdoor attacks in federated learning. Expert Systems with Applications, 259, 125359. https://doi.org/10.1016/j.eswa.2024.125359

Gupta, A., Luo, T., Ngo, M. V., & Das, S. K. (2022). Long-Short History of Gradients is All You Need: Detecting Malicious and Unreliable Clients in Federated Learning (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2208.10273

Gupta, P., Yadav, K., Gupta, B. B., Alazab, M., & Gadekallu, T. R. (2023). A Novel Data Poisoning Attack in Federated Learning based on Inverted Loss Function. Computers & Security, 130, 103270. https://doi.org/10.1016/j.cose.2023.103270

Huang, S., Li, Y., Chen, C., Shi, L., & Gao, Y. (2023). Multi-metrics adaptively identifies backdoors in Federated learning (Version 2). arXiv. https://doi.org/10.48550/ARXIV.2303.06601

Huang, W., Yang, W., Luo, Z., Qi, J., Sun, Q., & Kong, X. (2024). EEG-based Epilepsy Detection Using Robust Feature Learning Model with Manhattan Distance and L1 Regularization. 2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 5952–5959. https://doi.org/10.1109/bibm62325.2024.10822657

Jebreel, N. M., Domingo-Ferrer, J., Sánchez, D., & Blanco-Justicia, A. (2022). Defending against the Label-flipping Attack in Federated Learning (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2207.01982

Jin, R., Liu, Y., Huang, Y., He, X., Wu, T., & Dai, H. (2025). Sign-Based Gradient Descent With Heterogeneous Data: Convergence and Byzantine Resilience. IEEE Transactions on Neural Networks and Learning Systems, 36(2), 3834–3846. https://doi.org/10.1109/tnnls.2023.3345367

Karimireddy, S. P., He, L., & Jaggi, M. (2020). Byzantine-Robust Learning on Heterogeneous Datasets via Bucketing (Version 6). arXiv. https://doi.org/10.48550/ARXIV.2006.09365

Kasyap, H., & Tripathy, S. (2024). Sine: Similarity is Not Enough for Mitigating Local Model Poisoning Attacks in Federated Learning. IEEE Transactions on Dependable and Secure Computing, 21(5), 4481–4494. https://doi.org/10.1109/TDSC.2024.3353317

Kim, G., Kim, J., Kim, Y., Kim, H., & Park, H. (2025). FedWT: Federated Learning with Minimum Spanning Tree-based Weighted Tree Aggregation for UAV networks. ICT Express, 11(2), 275–280. https://doi.org/10.1016/j.icte.2024.12.005

Konečný, J. (2017). Stochastic, Distributed and Federated Optimization for Machine Learning (Version 1). arXiv. https://doi.org/10.48550/ARXIV.1707.01155

Lavaur, L., Busnel, Y., & Autrel, F. (2025). Investigating the Impact of Label-flipping Attacks against Federated Learning for Collaborative Intrusion Detection. Computers & Security, 156, 104462. https://doi.org/10.1016/j.cose.2025.104462

Li, H., Funk, M., Gürel, N. M., & Saeed, A. (2024). Collaboratively Learning Federated Models from Noisy Decentralized Data. 2024 IEEE International Conference on Big Data (BigData), 7879–7888. https://doi.org/10.1109/bigdata62323.2024.10825502

Li, L., Liu, Y., Ning, Y., Rini, S., & Chen, J. (2025). PNCS:Power-Norm Cosine Similarity for Diverse Client Selection in Federated Learning (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2506.15923

Li, M., Wan, W., Ning, Y., Hu, S., Xue, L., Zhang, L. Y., & Wang, Y. (2024). DarkFed: A Data-Free Backdoor Attack in Federated Learning (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2405.03299

Li, Q., Wang, X., & Ren, S. (2023). A Privacy Robust Aggregation Method Based on Federated Learning in the IoT. Electronics, 12(13), 2951. https://doi.org/10.3390/electronics12132951

Li, S., Cheng, Y., Wang, W., Liu, Y., & Chen, T. (2020). Learning to Detect Malicious Clients for Robust Federated Learning (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2002.00211

Li, S., & Dai, Y. (2024). BackdoorIndicator: Leveraging OOD Data for Proactive Backdoor Detection in Federated Learning (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2405.20862

Li, X., Wang, S., Wu, C., Zhou, H., & Wang, J. (2023). Backdoor Threats from Compromised Foundation Models to Federated Learning (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2311.00144

McMahan, H. B., Moore, E., Ramage, D., Hampson, S., & Arcas, B. A. y. (2016). Communication-Efficient Learning of Deep Networks from Decentralized Data. https://doi.org/10.48550/ARXIV.1602.05629

Miao, Y., Xie, R., Li, X., Liu, Z., Choo, K.-K. R., & Deng, R. H. (2024). Efficient and Secure Federated Learning Against Backdoor Attacks. IEEE Transactions on Dependable and Secure Computing, 21(5), 4619–4636. https://doi.org/10.1109/tdsc.2024.3354736

Multi-Metric Based Client Selection for Backdoor Defense in Federated Learning. (2025). In C. Yuan, Y. Li, N. Chen, & Z. Zhang, Advances in Transdisciplinary Engineering. IOS Press. https://doi.org/10.3233/atde241362

Mussabayev, R. (2024). Optimizing Euclidean Distance Computation. Mathematics, 12(23), 3787. https://doi.org/10.3390/math12233787

Nguyen, D. C., Ding, M., Pathirana, P. N., Seneviratne, A., Li, J., & Vincent Poor, H. (2021). Federated Learning for Internet of Things: A Comprehensive Survey. IEEE Communications Surveys & Tutorials, 23(3), 1622–1658. https://doi.org/10.1109/comst.2021.3075439

Nguyen, T. D., Nguyen, T., Nguyen, P. L., Pham, H. H., Doan, K. D., & Wong, K.-S. (2024). Backdoor attacks and defenses in federated learning: Survey, challenges and future research directions. Engineering Applications of Artificial Intelligence, 127, 107166. https://doi.org/10.1016/j.engappai.2023.107166

Nguyen, T. D., Rieger, P., Chen, H., Yalame, H., Möllering, H., Fereidooni, H., Marchal, S., Miettinen, M., Mirhoseini, A., Zeitouni, S., Koushanfar, F., Sadeghi, A.-R., & Schneider, T. (2021). FLAME: Taming Backdoors in Federated Learning (Extended Version 1) (Version 5). arXiv. https://doi.org/10.48550/ARXIV.2101.02281

Ren, Q., Zheng, Y., Yang, C., Li, Y., & Ma, J. (2024). Shadow backdoor attack: Multi-intensity backdoor attack against federated learning. Computers & Security, 139, 103740. https://doi.org/10.1016/j.cose.2024.103740

Rieger, P., Nguyen, T. D., Miettinen, M., & Sadeghi, A.-R. (2022). DeepSight: Mitigating Backdoor Attacks in Federated Learning Through Deep Model Inspection. Proceedings 2022 Network and Distributed System Security Symposium. Network and Distributed System Security Symposium, San Diego, CA, USA. https://doi.org/10.14722/ndss.2022.23156

Saeed-Uz-Zaman, Li, B., Hamid, M., Saleem, M., & Aman, M. (2025). A4FL: Federated Adversarial Defense via Adversarial Training and Pruning Against Backdoor Attack. IEEE Access, 13, 91070–91088. https://doi.org/10.1109/access.2025.3568275

Serengil, S., & Ozpinar, A. (2025). Encrypted Vector Similarity Computations Using Partially Homomorphic Encryption: Applications and Performance Analysis (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2503.05850

Shanmugarasa, Y., Paik, H., Kanhere, S. S., & Zhu, L. (2023). A systematic review of federated learning from clients' perspective: Challenges and solutions. Artificial Intelligence Review, 56(S2), 1773–1827. https://doi.org/10.1007/s10462-023-10563-8

Sharma, A., & Marchang, N. (2024). Probabilistic Sign Flipping Attack in Federated Learning. 2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT), 1–6. https://doi.org/10.1109/icccnt61001.2024.10725463

Shi, C., Ji, S., Pan, X., Zhang, X., Zhang, M., Yang, M., Zhou, J., Yin, J., & Wang, T. (2024). Towards Practical Backdoor Attacks on Federated Learning Systems. IEEE Transactions on Dependable and Secure Computing, 21(6), 5431–5447. https://doi.org/10.1109/tdsc.2024.3376790

Shi, J., Wan, W., Hu, S., Lu, J., & Yu Zhang, L. (2022). Challenges and Approaches for Mitigating Byzantine Attacks in Federated Learning. 2022 IEEE International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom), 139–146. https://doi.org/10.1109/trustcom56396.2022.00030

Sun, Z., Kairouz, P., Suresh, A. T., & McMahan, H. B. (2019). Can You Really Backdoor Federated Learning? (Version 2). arXiv. https://doi.org/10.48550/ARXIV.1911.07963

Tan, Q., Li, Y., & Shin, B.-S. (2025). Defending against Backdoor Attacks in Federated Learning by Using Differential Privacy and OOD Data Attributes. Computer Modeling in Engineering & Sciences, 143(2), 2417–2428. https://doi.org/10.32604/cmes.2025.063811

Tang, W., & Gan, G. (2024). Personalized federation algorithm based on model similarity and data importance. 2024 5th International Conference on Computer, Big Data and Artificial Intelligence (ICCBD+AI), 472–478. https://doi.org/10.1109/iccbd-ai65562.2024.00084

Thaker, P., & Mohan, B. R. (2024). Enhancing Deep Compression of CNNs: A Novel Regularization Loss and the Impact of Distance Metrics. IEEE Access, 12, 172537–172547. https://doi.org/10.1109/access.2024.3498901

Walter, K., Mohammady, M., Nepal, S., & Kanhere, S. S. (2024). Mitigating Distributed Backdoor Attack in Federated Learning Through Mode Connectivity. Proceedings of the 19th ACM Asia Conference on Computer and Communications Security, 1287–1298. https://doi.org/10.1145/3634737.3637682

Wan, Y., Qu, Y., Ni, W., Xiang, Y., Gao, L., & Hossain, E. (2024). Data and Model Poisoning Backdoor Attacks on Wireless Federated Learning, and the Defense Mechanisms: A Comprehensive Survey. IEEE Communications Surveys & Tutorials, 26(3), 1861–1897. https://doi.org/10.1109/comst.2024.3361451

Wang, D., Zhang, N., & Tao, M. (2021). Adaptive Clustering-Based Model Aggregation for Federated Learning with Imbalanced Data. 2021 IEEE 22nd International Workshop on Signal Processing Advances in Wireless Communications (SPAWC), 591–595. https://doi.org/10.1109/spawc51858.2021.9593144

Wang, Z., Hu, Q., Zou, X., Hu, P., & Cheng, X. (2025). Can We Trust the Similarity Measurement in Federated Learning? IEEE Transactions on Information Forensics and Security, 20, 3758–3771. https://doi.org/10.1109/tifs.2024.3516567

Wang, Z., Zhang, Z., Li, Z., Wu, Y., Liu, Y., Li, M., Li, X., Liu, Y., An, J., Liang, W., & Zhu, L. (2025). Resisting Poisoning Attacks in Federated Learning via Dual-Domain Distance and Trust Assessment. IEEE Transactions on Information Forensics and Security, 20, 7394–7409. https://doi.org/10.1109/tifs.2025.3589061

Wei, W., & Liu, L. (2025). Trustworthy Distributed AI Systems: Robustness, Privacy, and Governance. ACM Computing Surveys, 57(6), 1–42. https://doi.org/10.1145/3645102

Wen, J., Zhang, Z., Lan, Y., Cui, Z., Cai, J., & Zhang, W. (2023). A survey on federated learning: Challenges and applications. International Journal of Machine Learning and Cybernetics, 14(2), 513–535. https://doi.org/10.1007/s13042-022-01647-y

Wu, J., Luo, F., Sun, T., Wang, H., & Zhang, W. (2025). Privacy-Preserving Federated Learning Scheme with Mitigating Model Poisoning Attacks: Vulnerabilities and Countermeasures (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2506.23622

Wu, Z., Ling, Q., Chen, T., & Giannakis, G. B. (2020). Federated Variance-Reduced Stochastic Gradient Descent With Robustness to Byzantine Attacks. IEEE Transactions on Signal Processing, 68, 4583–4596. https://doi.org/10.1109/tsp.2020.3012952

Yaldiz, D. N., Zhang, T., & Avestimehr, S. (2023). Secure Federated Learning against Model Poisoning Attacks via Client Filtering. https://doi.org/10.48550/ARXIV.2304.00160

Yang, Z., Bao, W., Yuan, D., Tran, N. H., & Zomaya, A. Y. (2022). Federated Learning With Nesterov Accelerated Gradient. IEEE Transactions on Parallel and Distributed Systems, 33(12), 4863–4873. https://doi.org/10.1109/tpds.2022.3206480

Yin, D., Chen, Y., Ramchandran, K., & Bartlett, P. (2018). Byzantine-Robust Distributed Learning: Towards Optimal Statistical Rates (Version 2). arXiv. https://doi.org/10.48550/ARXIV.1803.01498

Zeng, Y., Teng, S., Xiang, T., Zhang, J., Mu, Y., Ren, Y., & Wan, J. (2023). A Client Selection Method Based on Loss Function Optimization for Federated Learning. Computer Modeling in Engineering & Sciences, 137(1), 1047–1064. https://doi.org/10.32604/cmes.2023.027226

Zhang, W., Li, Y., An, L., Wan, B., & Wang, X. (2024). SARS: A Personalized Federated Learning Framework Towards Fairness and Robustness against Backdoor Attacks. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 8(4), 1–24. https://doi.org/10.1145/3678571

Zhang, X., Xue, X., Du, X., Xie, X., Liu, Y., & Sun, M. (2025). Runtime Backdoor Detection for Federated Learning via Representational Dissimilarity Analysis. IEEE Transactions on Dependable and Secure Computing, 1–18. https://doi.org/10.1109/tdsc.2025.3550330

Zhu, C., Li, Y., Rao, B., Zhang, J., Mao, Y., & Zhong, S. (2025). SPA: Towards More Stealth and Persistent Backdoor Attacks in Federated Learning (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2506.20931

Zhu, T., Guo, Z., Yao, C., Tan, J., Dou, S., Wang, W., & Han, Z. (2024). Byzantine-robust Federated Learning via Cosine Similarity Aggregation. Computer Networks, 254, 110730. https://doi.org/10.1016/j.comnet.2024.110730