2025 19th International Conference on Ubiquitous Information Management and Communication (IMCOM) | 979-8-3315-0781-7/25/$31.00 ©2025 IEEE | DOI: 10.1109/IMCOM64595.2025.10857574

Towards Robust Federated Learning: Investigating Poisoning Attacks Under Clients Data Heterogeneity

Abdenour Soubih[§], Seyyid Ahmed Lahmer[†], Mohammed Abuhamad[‡], Tamer Abuhmed[§]
[§]College of Computing and Informatics, Sungkyunkwan University, Suwon, Korea, Email: {abdenour, tamer}@skku.edu
[†]Department of Information Engineering, University of Padova, Padua, Italy, Email: seyyidahmed.lahmer@unipd.it
[‡]Department of Computer Science, Loyola University Chicago, Chicago, USA, Email: mabuhamad@luc.edu

***Abstract*—Federated Learning (FL) offers a privacy-preserving solution by enabling multiple clients to train a shared model collaboratively without centralizing data. However, the decentralized nature of FL presents challenges, particularly regarding security and performance under adversarial conditions. This paper investigates the effects of poisoning attacks under data heterogeneity. Our experiments evaluate the impact of varying malicious client fractions and poison concentration levels on the accuracy of the model. We explore the effects of poisoning attacks on FedAvg and FedNova models using medical imaging tasks. Our findings reveal that increasing data heterogeneity exacerbates the effects of poisoning, with FedNova demonstrating greater resilience compared to FedAvg. We found that the number of malicious clients plays a more significant role in degrading performance than the ratio of poisoning samples shared by each malicious client, suggesting that even modest levels of poisoning can be tolerated by most algorithms. The study highlights the importance of developing robust defense mechanisms to maintain model performance under adversarial conditions.**

***Index Terms*—Federated Learning, Robustness, Data heterogeneity, Adversarial Attacks, Machine Learning Security.**

I. Introduction

Federated Learning (FL) has emerged as an innovative paradigm that enables multiple clients to collaboratively train a shared global machine learning model on the private data of clients. This decentralized approach inherently enhances privacy, making it particularly well suited for sensitive domains like healthcare and finance, where data confidentiality is crucial. Despite this advantage, FL presents several significant challenges, particularly in terms of security and privacy, which must be effectively addressed to fully exploit its potential.

One of the most pressing challenges in FL is the inherent *heterogeneity* among participating clients. This heterogeneity can emerge in various forms, including disparities in data distributions, differences in model architectures, and variations in their system capabilities. Such discrepancies can lead to a degradation in the performance of the global model, as the model may struggle to generalize across heterogeneous data. The integration of privacy-enhancing methods, including differential privacy, adds another layer of complexity to the heterogeneity of the data. Although these methods are crucial for protecting data, they introduce additional noise that amplifies heterogeneity challenges, making the learning process even more complex.

Beyond performance issues, FL systems are particularly susceptible to *adversarial attacks*. The decentralized nature of FL provides malicious clients with opportunities to compromise the integrity of the global model by injecting poisoned data or manipulating model updates. These vulnerabilities pose severe risks to the robustness and reliability of FL frameworks, making it essential to detect these attacks and develop effective defense mechanisms.

This paper presents an evaluation of FL resilience under data heterogeneity and adversarial attacks, particularly *poisoning attacks* [1]. We conducted experiments specifically on the challenges posed by *data heterogeneity*, *poison attacks*, and *poison attacks in the context of heterogeneity*. In addition, we explore the trade-offs between model performance and privacy preservation in federated learning, noting that while FL inherently promotes privacy by keeping raw data on client devices, aggregating model updates introduces potential vulnerabilities. Although we do not directly experiment on this topic, we refer to studies that have examined the impact of privacy measures on overall performance and robustness.

In this study, we aim to evaluate the performance of several FL approaches in heterogeneous learning environments, particularly in the presence of adversarial attackers. To this end, we address the following key research questions.

1) **How does heterogeneity in federated learning affect model performance?** We analyze the impact of varying levels of data and system heterogeneity on the effectiveness of different FL algorithms.
2) **Is there a relationship between heterogeneity and the resilience of the model against adversarial attacks?** We explore whether increased heterogeneity influences the robustness of FL systems, particularly in the presence of poison attacks.

Contributions. This paper offers the following contributions.

- The paper offers a detailed evaluation of four FL algorithms under varying levels of data heterogeneity, highlighting their strengths and limitations.
- We examine the impact of poisoning attacks on FL, providing experimental insights into how these attacks affect system performance.
- We provide a practical recommendations to develop secure and efficient FL frameworks.We also provide *open-source implementation*, including code and data[1].

[1]The complete implementation is available at https://github.com/InfoLab-SKKU/SecurityAnalysisFL

979-8-3315-0782-4/25/$31.00 ©2025 IEEE

Paper Organization. The paper is organized as follows: Section II categorizes the security challenges posed by heterogeneity in FL and reviews existing solutions to mitigate them. Section III describes our experimental setup and presents the results. In Section IV, we discuss the findings and propose directions for future research. Section VI concludes the paper.

II. Background and Related Works

FL [2] enables multiple clients– such as smartphones, or Internet of Things (IoT) devices– to collaboratively train a machine learning model without sharing their raw local data with a central server. This decentralized approach enhances data privacy and security.

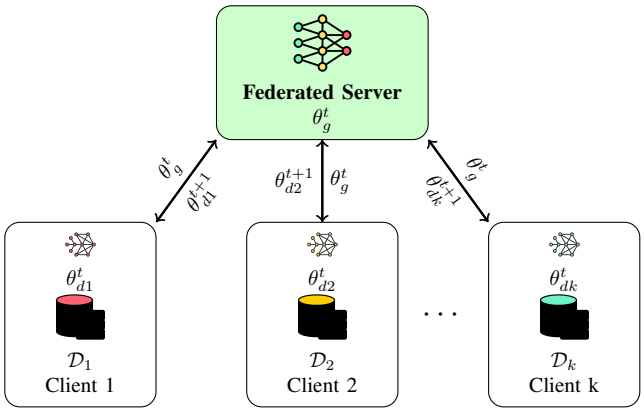


Fig. 1: Federated Learning Scenario: The central server sends the global model parameters θ_g^t to the clients. Each client updates the model based on its local data $\mathcal{D}_d$ to get θ_d^{t+1} and sends the updated model back to the server. The server then aggregates these updates to form the next global model θ_g^{t+1}.

As shown in Figure 1, a typical FL scenario involves n clients, each holding a local dataset D_i for $i = 1, 2, \ldots, n$. The objective of FL is to minimize the total loss of overall clients, formulated as:

$$\min_{\theta} \sum_{i=1}^{n} L(\theta, D_i)$$

Here, $\theta \in \mathbb{R}^d$ represents the global model with d parameters, and $L(\theta, D_i)$ is the local loss function for the client i. FL operates iteratively to train the global model θ.

During the t-th training round, the server distributes the current global model θ_{t-1} to all clients (or a selected subset). For simplicity, we assume that all clients are selected, although in practice only a subset may participate in each round. Each client i initializes its local model θ_t^i as θ_{t-1} and updates θ_t^i by minimizing $L(\theta, D_i)$ using Stochastic Gradient Descent (SGD). After training, the client i computes the model update $g_t^i = \theta_t^i - \theta_{t-1}$ and sends it to the server. The server then aggregates the updates from all clients to update the global model ($\theta_t = \theta_{t-1} + g_t$), where $g_t = \text{AR}(\{g_t^i\}_{i=1}^n)$ is the aggregated-model update, and AR denotes the aggregation rule used by the server. A common aggregation rule is Federated Averaging (FedAvg), where the aggregated model update is the average of the clients' model updates.

Performance of FL under Heterogeneity. Several studies have explored the impact of client data heterogeneity on FL performance. Chang et al. [3] and Ezzeldin et al. [4] examined how non-iid data distributions across clients affect model performance. They found that heterogeneity significantly degrades performance and proposed adjustments to aggregation algorithms to mitigate these effects. Salazar et al. conducted a comprehensive survey on group fairness in FL, highlighting how heterogeneous data distributions across clients can exacerbate biases, highlighting the need for fairness-aware approaches [5]. Li et al. explored the impact of heterogeneous data distributions on model bias, as clients with smaller or less representative datasets may have their contributions underrepresented in the aggregated model [6]. McMahan et al. investigated convergence rates with heterogeneous clients and found that some clients slow down training due to data distribution disparities [7].

Security Challenges: Poisoning Attacks. The security of FL systems has been a significant concern, particularly concerning poisoning attacks where malicious clients aim to corrupt the global model. A study [8] provided a comprehensive analysis of various poisoning strategies and their impact on FL. Their findings highlighted the vulnerability of FL to both data and model poisoning attacks, emphasizing the need for robust defenses. In another study [8], a defense framework was proposed that uses anomaly detection to identify and exclude malicious updates, significantly improving FL robustness.

Effects of Heterogeneity and Poisoning Attacks. Although heterogeneity and poisoning attacks have been studied individually, their combined impact on the performance and security of federated learning remains less explored. In a study [3], the authors studied this combined impact, revealing that the presence of heterogeneity and poisoning can significantly affect FL systems. The study found that heterogeneity can mask the presence of poisoning attacks, making it harder to detect and mitigate them. This shows the importance of developing strategies that simultaneously address heterogeneity and poisoning in FL systems.

To simulate heterogeneity in federated learning, researchers often employ partitioning techniques such as *Dirichlet partitioning* [9] and *pathological partitioners*. The Dirichlet partitioning method generates non-IID (non-independent and identically distributed) data by sampling from a Dirichlet distribution, where the parameter α controls the level of heterogeneity. A small α value results in more uneven data distributions across clients, increasing heterogeneity, while larger α values create more uniform distributions. Pathological partitioners simulate heterogeneity by assigning each client a specific number of unique labels. For example, in a dataset with multiple classes, a pathological partitioner would limit each client to only a few labels, ensuring that the data distribution across clients is skewed and more heterogeneous.

Various methodologies have been used to evaluate FL

systems under these challenging conditions. Yang et al. [10] employed a simulation-based approach to model different levels of heterogeneity and attack scenarios, providing insight into the resilience of different FL algorithms. On the other hand, Chang et al. [3] used real-world datasets to validate their findings, highlighting the practical implications of their theoretical models. We note that these studies were mainly limited to the evaluation of FL systems under heterogeneity.

Data Heterogeneity Mitigation. In the context of FL, an approach to achieving group fairness involves each client independently applying local debiasing techniques to their locally trained models. The central FL server then aggregates these model parameters using standard FL aggregation algorithms, such as Federated Averaging (FedAvg) [11], or its derivatives, such as Federated Averaging with Momentum (FedAvgM) [12], Federated Proximal Optimization (FedProx) [6], and Federated Normalized Averaging (FedNova) [13]. These methods allow for training a global model without necessitating the explicit sharing of local datasets. However, a significant drawback is that isolated debiasing at each client can lead to suboptimal performance, particularly in scenarios where data distributions are highly heterogeneous among clients.

An alternative solution to fair training in FL involves adapting debiasing techniques from the extensive literature on centralized fair training [3], [14]–[16] for use in FL environments. One approach involves learning fair representations, where clients train local embeddings that obfuscate sensitive attributes while preserving utility, followed by global aggregation [17]. Another method applies optimized pre-processing to transform input data locally, ensuring that protected attributes do not disproportionately influence predictions, with transformed datasets used for training [18]. Additionally, certifying and removing disparate impact leverages causal fairness metrics to de-bias local datasets, enabling aggregation of bias-free models at the server [19]. However, these methods face challenges: fair representations may lose task-specific information, pre-processing can reduce utility for highly diverse datasets, and certifying disparate impact often requires sharing subgroup performance metrics, raising privacy concerns. Although these approaches can potentially result in more reasonable and fair training outcomes, they often require clients to exchange additional detailed information with the server regarding the composition of their datasets. This can lead to privacy concerns, as it may inadvertently reveal information about various subgroups within a client's dataset. For instance, the server might need to know the model's performance on each subgroup or access local statistical information about each group in the dataset.

Our experiments extend this evaluation by jointly exploring both heterogeneity and poisoning attacks, an area that has received limited attention in prior research. We provide a more comprehensive analysis of FL robustness and security considering both heterogeneity and poisoning attacks.

III. Methodology

Federated Learning Algorithms. In this study, we selected **FedAvg**, **FedAvgM**, **FedProx**, and **FedNova** for their unique purposes in federated learning. **FedAvg** serves as a simple and efficient baseline. **FedAvgM** improves convergence under non-IID conditions. **FedProx** addresses system heterogeneity, and **FedNova** ensures fairness in heterogeneous systems. These algorithms enable a comprehensive analysis of performance under varying heterogeneity levels.

- **FedAvg:** In FedAvg [11], clients perform local updates, and the server aggregates models using weighted averaging:

$$\theta_t = \sum_{k=1}^{K} \frac{n_k}{n} \theta_k^t$$

  FedAvg is simple and communication-efficient algorithm, ideal for homogeneous data and is the baseline algorithm.
- **FedAvgM:** In FedAvgM [12], the algorithm add momentum to the server updates to stabilize training:

$$\theta_t = \theta_{t-1} + \gamma m_t + \eta \sum_{k=1}^{K} \frac{n_k}{n} (\theta_k^t - \theta_{t-1})$$

  It improves convergence in non-IID data settings.
- **FedProx:** In FedProx [13], the algorithm introduces a proximal term to limit the deviation of local updates:

$$\min_{\theta_k} F_k(\theta_k) + \frac{\mu}{2} \|\theta_k - \theta\|^2$$

  FedProx handles system heterogeneity and stable training.
- **FedNova:** [13] algorithm modifies FedAvg by normalizing each client's contribution to the global model based on the number of local updates performed, addressing the issue of fairness when clients perform different amounts of work. The global model update is expressed as:

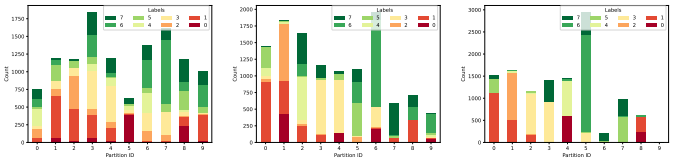
$$\theta_t = \sum_{k=1}^{K} \frac{\tau_k}{\sum_{k=1}^{K} \tau_k} \theta_k^t$$

  where τ_k is the number of local updates performed by client k. FedNova ensures fairness among clients with different computational capacities, especially in heterogeneous FL systems where some clients might perform significantly more updates than others.

The presented algorithms, i.e., FedAvg, FedAvgM, FedProx, and FedNova, highlight different approaches to addressing the issues of communication efficiency, model convergence, and fairness in heterogeneous FL environments.

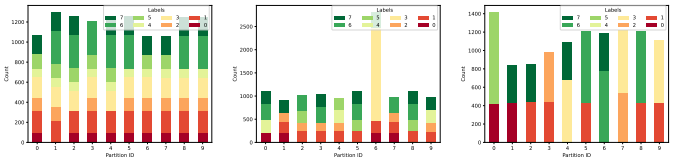
Datasets. We conducted our experiments using three different image datasets, each representing medical imaging tasks:

- **PathMNIST**: The dataset is derived from pathology images of colorectal cancer and consists of nine classes of tissues. The images are colorized and resized to 28x28 pixels for classification in medical image analysis [20].
- **BloodMNIST**: The dataset contains images of peripheral blood smears, categorized into eight different blood cell



(a) alpha=0.9 **(b)** alpha=0.3) **(c)** alpha=0.1

Fig. 2: Data distribution of randomly sampled 10 clients on BloodMNIST using a Dirichlet partitioner with varying alpha values, demonstrating how different alpha settings affect the distribution of samples across partitions.



(a) num-classes=7 **(b)** num-classes=4) **(c)** num-classes=2

Fig. 3: Data distribution on BloodMNIST using a Pathological partitioner with varying alpha values, demonstrating how different number of classes by clients settings affect the distribution of samples across partitions.

types. The images are grayscale and resized to 28x28 pixels. It is used primarily for the classification of blood cell morphology [20].

- **TissueMNIST**: The dataset is derived from images of human tissues, classified into eight different tissue types. The images are colorized and resized to 28x28 pixels. The dataset aims to facilitate the classification of tissue images in biomedical research [20].

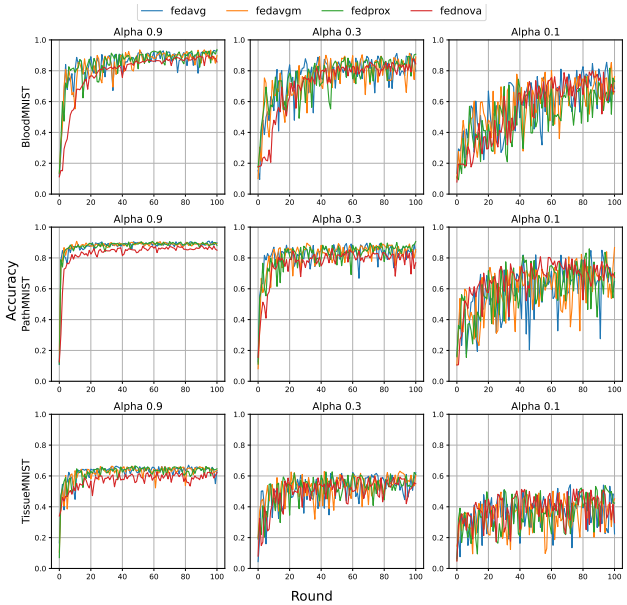
Partitioning Strategies for Heterogeneity Analysis. We investigate the impact of data heterogeneity on the performance of four federated learning algorithms: **FedAvg**, **FedAvgM**, **FedProx**, and **FedNova**. To evaluate the robustness of these algorithms under different levels of heterogeneity.

We employed two partitioning strategies to simulate varying degrees of data heterogeneity as illustrated in Figure 2 and 3:

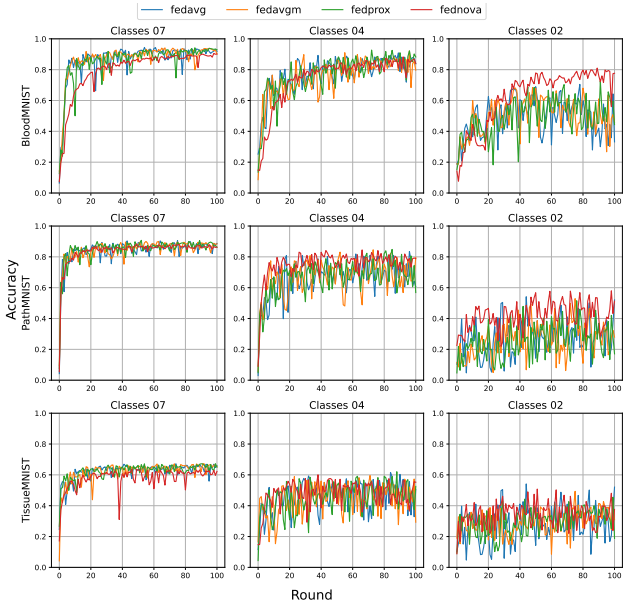
- **Dirichlet Partitioning**: A Dirichlet distribution with varying alpha values (0.9, 0.3, 0.1) was used to generate different levels of non-IID data. Smaller alpha values lead to more heterogeneous data distribution across clients.
- **Pathological Partitioning**: Additionally, we used a pathological partitioner, where each client was assigned data containing only 2, 4, or 7 distinct labels. This method introduces controlled levels of heterogeneity by limiting the variety of labels available to individual clients.

By assessing the performance of these algorithms across varying levels of heterogeneity and using multiple datasets, we aim to provide a thorough comparison of their effectiveness in handling non-IID data distributions.

Threat Model. The objective of the attacker is to compromise the integrity of the global model by decreasing its accuracy and performance on clean test data, using the following capabilities: 1) The attacker only has access to the local training



A) Dirichlet Partitioning



B) Class Diversity

Fig. 4: Mean accuracy comparison of Federated Learning algorithms (FedAvg, FedAvgM, FedProx, and FedNova) on BloodMNIST, PathMNIST, and TissueMNIST datasets under Dirichlet partitioning with varying α values (**A**) and class diversity with 2, 4, and 7 classes per client (**B**). The figure illustrates the impact of increasing data heterogeneity on model performance, with FedNova consistently outperforming the other algorithms, particularly in highly non-IID settings (lower α values). FedAvg and FedAvgM show significant performance degradation as data becomes more skewed, while FedProx performs better in intermediate heterogeneity but struggles under extreme conditions.

dataset of a compromised client; 2) The attacker does not know the server-side aggregation mechanism. We assume that the

attacker can compromise one or more clients in the federated learning setup. However, we do not consider scenarios where the server itself is compromised. Additionally, we focus on untargeted attacks under the described capabilities.

Experimental Setup. The experiments are designed to analyze the impact of heterogeneity on FL performance, incorporating datasets, FL algorithms, partitioning strategies, and attacks used in our experiments. We employed *MobileNetV2* [21] as the global model architecture and conducted experiments with 50 clients, of which 10% participated in each training round. The training spanned 100 rounds to evaluate the performance of the models over time. We applied a poisoning attack to two federated learning algorithms, FedAvg and FedNova, using the Blood MNIST dataset. The attack involved setting the poison concentration to 0.8 and incorporating malicious clients 30% into the system. Specifically, we simulate a label-flipping attack, in which adversaries modify their label mappings to degrade the performance of the global model without altering their local training data. These experiments allow us to assess the robustness of FedAvg and FedNova in a heterogeneous data environment, simulating real-world scenarios where client data distributions vary. Provides information on the resilience of these models to poisoning attacks in challenging conditions.

Evaluation Metrics. The performance of the federated learning (FL) algorithms is assessed using Accuracy. This metric quantifies the overall classification performance of the global model in the test dataset. It provides a robust measure of the model's ability to correctly predict labels in unseen data, serving as a key indicator of its effectiveness and generalizability.

IV. Evaluation and Discussion

The experiments presented in **Figures** 5, 4, and 6 provide a comprehensive comparison of various Federated Learning algorithms (FedAvg, FedAvgM, FedProx, and FedNova) in terms of mean accuracy across the last 10 training rounds, evaluated under two distinct partitioning schemes: Dirichlet partitioning [9] and pathological partitioning [3]. The results highlight several important insights about the performance of these algorithms in the presence of non-IID data distributions and poison attacks.

Performance under Heterogeneity. The results show the varying capabilities of federated learning algorithms under heterogeneous data distributions.

- **FedNova** consistently demonstrates superior performance across various datasets and partitioning schemes. Its normalization strategy effectively mitigates the negative impact of heterogeneity in data by ensuring fairness in client contributions, regardless of their computational power.
- **FedAvg** and **FedAvgM** show competitive results in IID or mildly heterogeneous environments, but experience significant performance degradation in highly non-IID settings, particularly under pathological partitioning or when trained on highly skewed Dirichlet distributions (low α values).

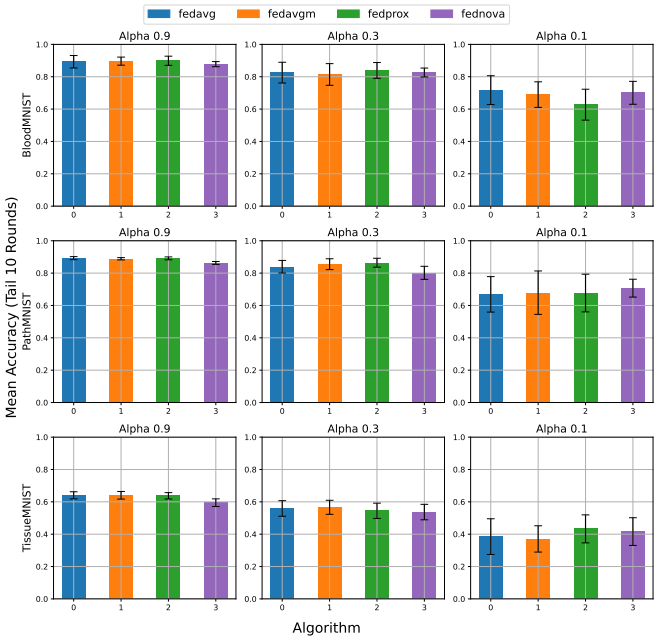


Fig. 5: Accuracy comparisons of four federated learning algorithms (FedAvg, FedAvgM, FedProx and FedNova) on three image datasets (BloodMNIST, PathMNIST and TissueMNIST) using a Dirichlet partitioner with varying alpha values (0.9, 0.3, 0.1). The results demonstrate how increasing data heterogeneity (decreasing alpha) impacts the accuracy of each algorithm.

- **FedProx**, designed to address system heterogeneity, improves performance in Dirichlet partitioning scenarios but performs similarly to FedAvg in pathological partitioning settings.

Impact of Poison Attacks. The results further emphasize the vulnerability of federated learning systems to poison attacks, especially under increasing data heterogeneity. In the Dirichlet partitioning scheme (**Figure** 5), where the degree of heterogeneity is controlled by the α parameter, we observe a sharp decline in model performance as heterogeneity increases (i.e., as α decreases). The effect is particularly pronounced when $\alpha = 0.1$, where accuracy drops substantially, especially in the presence of poisoned data. In these cases, **FedNova** continues to outperform **FedAvg**, showcasing its robustness to both heterogeneity and poisoning, with performance differences becoming more apparent as α decreases.

In contrast, the Pathological partitioning scheme (**Figure** 8) reveals a similarly detrimental impact of poisoning on model performance. When data is partitioned by class, poisoned class groups exhibit a notable decline in accuracy. The performance comparison between **FedAvg** and **FedNova** across different class groups further supports the observation that **FedNova** is more resilient to poison attacks, particularly when the partitioning strategy exacerbates data imbalance.

The results from our experiments clearly demonstrate that

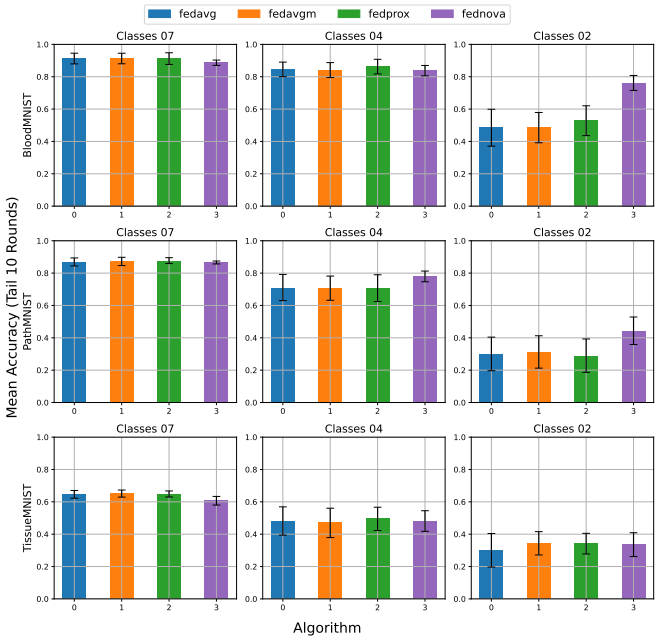
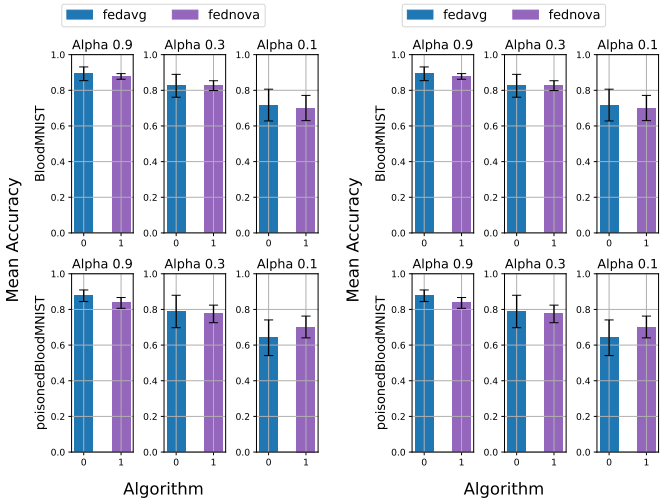


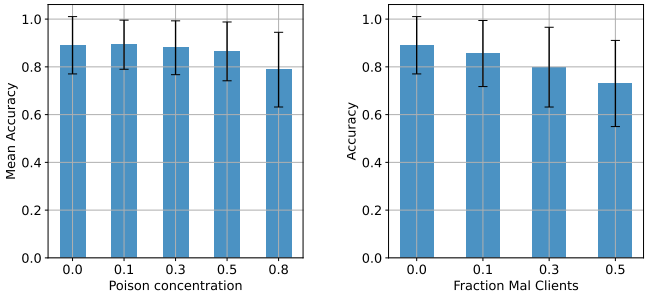
Fig. 6: Mean accuracy comparisons of four federated learning algorithms (FedAvg, FedAvgM, FedProx, and FedNova) across three datasets (BloodMNIST, PathMNIST, and TissueMNIST) using a pathological partitioner. The experiments were conducted with different levels of class diversity (2, 4, and 7 classes per client). The results show that as the number of classes per client increases (lower heterogeneity).



(a) Mean accuracy under Dirichlet partitioning for varying α values.

(b) Mean accuracy under pathological partitioning.

Fig. 7: Comparison of Federated Learning algorithms (FedAvg and FedNova) in terms of mean accuracy under two partitioning schemes: Dirichlet and pathological partitioning. The figures show the impact of data heterogeneity and poison attacks, with FedNova consistently outperforming the other algorithms, especially under conditions of higher heterogeneity and data imbalance.



(a) Increasing poison concentration with a fixed number of malicious clients (30%).

(b) Increasing number of malicious clients with fixed poison concentration (0.8).

Fig. 8: Accuracy trends in the presence of malicious clients and varying poison concentrations across different scenarios. These figures highlight the impact on model accuracy based on the proportion of malicious clients and the concentration of poisoning,The experiments used a MobileNet model on the Blood MNIST dataset, with with 50 clients over 100 global communication rounds.

FedAvg, **FedAvgM**, and **FedProx** exhibit similar patterns of performance degradation as the degree of data heterogeneity increases. Specifically, as data distributions become more non-IID, all three algorithms experience notable reductions in accuracy. This degradation is further exacerbated by the presence of malicious clients, where an increase in the number of adversarial participants significantly compromises performance. The same observation applies across both Dirichlet and pathological partitioning schemes: as the number of malicious clients rises, these algorithms struggle to maintain robust performance under attack.

In contrast, **FedNova** consistently proves to be more resilient to both data heterogeneity and poison attacks. The key to FedNova's robustness lies in its *advanced normalization strategy*, which adjusts the contribution of each client in proportion to their computational capabilities and the size of their data. This prevents clients with smaller or less diverse datasets from disproportionately influencing the global model, a common problem in heterogeneous FL environments. Furthermore, by decoupling client contributions from the actual update frequency, FedNova minimizes the impact of skewed or poisoned data, making it more resistant to malicious attacks, even when the number of adversarial clients is reasonable. However, it is important to note that **FedNova's performance also degrades as the dataset difficulty increases**, as more complex datasets strain the model's ability to generalize, despite its resilience to heterogeneity.

In scenarios where **both data heterogeneity and malicious clients** are present, we observe an even sharper performance degradation across all algorithms. Poison attacks not only directly degrade model performance but also *aggravate the*

effects of heterogeneity, creating compounding negative impacts. Despite these challenges, **FedNova** remains the most resilient, though its performance does diminish compared to less adversarial or more homogeneous conditions. This highlights the strength of FedNova's aggregation method in preserving model accuracy, even in hostile and complex environments, though its advantages diminish as the difficulty of both the dataset and the attack increases.

Regarding malicious attacks, our results indicate that the *number of malicious clients* plays a more significant role in degrading performance than the *concentration of the poisoning itself*. This suggests that even modest levels of poisoning can be tolerated by most algorithms, but a larger proportion of adversarial participants has a much more severe impact on the global model, especially in the case of FedAvg, FedAvgM, and FedProx.

In conclusion, these findings highlight the need for federated learning solutions that can **simultaneously trade off performance, security, and fairness**. Current approaches, while showing promise, often focus on one aspect at the expense of the others. Future solutions must consider fairness (to address heterogeneity), security (to defend against malicious attacks), and performance (to ensure scalability and accuracy) as intertwined objectives. Only by optimizing across these three axes can federated learning systems achieve robust, equitable, and secure deployments in real-world settings.

V. Limitations and Future work

Limitations. While our study provides a foundational analysis of adversarial attacks in federated learning, it has several limitations. The current analysis does not explore more sophisticated poisoning methods that could exploit advanced vulnerabilities in federated learning setups. Additionally, we have not tested defense mechanisms to mitigate such attacks, leaving their effectiveness against the described adversarial scenarios unexplored. The evaluation is centered on specific metrics and scenarios, which may not fully capture the broader range of adversarial threats and their impacts in diverse federated learning environments. Furthermore, the study does not address the long-term effects of poisoning attacks on the model's convergence behavior, which could vary significantly across training rounds. Finally, limited consideration is given to multi-objective adversarial goals, such as balancing stealth and efficacy, which could introduce more nuanced challenges for defenses.

Future Work. To address these limitations, our future work will focus on experimenting with more advanced poisoning techniques, including adaptive and stealthy attacks, to better understand their impacts on the performance of the global model. Furthermore, we plan to investigate the effectiveness of state-of-the-art defense mechanisms under the proposed threat model, particularly in heterogeneous federated learning setups. Another key direction will involve evaluating the interplay between system heterogeneity and attack resilience to identify new vulnerabilities and opportunities to design robust federated learning systems. By extending the scope of our analysis and incorporating these additional dimensions, we aim to provide a more comprehensive understanding of adversarial threats and defenses in federated learning. This will help bridge the current gap in the literature and contribute to the development of more resilient federated learning frameworks.

VI. Conclusion

This study highlights the significant challenges that Federated Learning (FL) systems face under conditions of data heterogeneity and adversarial attacks, specifically poisoning attacks. Through a comprehensive evaluation of key FL algorithms—FedAvg, FedAvgM, FedProx, and FedNova—we found that data heterogeneity considerably exacerbates the vulnerabilities of these systems, leading to a notable decline in model performance. Among the algorithms tested, FedNova demonstrated the highest resilience, effectively mitigating the negative impact of both data heterogeneity and poisoning attacks due to its advanced normalization strategy. Our findings emphasize the critical need for integrated solutions that address performance and security simultaneously in FL environments. Although current defense mechanisms can partially protect against adversarial threats, they often struggle in highly heterogeneous settings where data and computational discrepancies are prevalent. Future work should focus on developing robust and adaptive FL frameworks that can dynamically adjust to varying levels of heterogeneity and adversarial intensity, ensuring reliable and secure model training in real-world applications.

Acknowledgment

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT)(No. 2021R1A2C1011198), (Institute for Information & communications Technology Planning & Evaluation) (IITP) grant funded by the Korea government (MSIT) under the ICT Creative Consilience Program (IITP-2021-2020-0-01821), and AI Platform to Fully Adapt and Reflect Privacy-Policy Changes (No.RS-2022-II220688).

References

[1] B. Biggio, B. Nelson, and P. Laskov, "Poisoning attacks against support vector machines," *arXiv preprint arXiv:1206.6389*, 2012.

[2] Y. Xie, M. Fang, and N. Z. Gong, "Poisonedfl: Model poisoning attacks to federated learning via multi-round consistency," *arXiv preprint arXiv:2404.15611*, 2024.

[3] H. Chang and R. Shokri, "Bias propagation in federated learning," in *the 11th International Conference on Learning Representations*, 2023.

[4] Y. H. Ezzeldin, S. Yan, C. He, E. Ferrara, and A. S. Avestimehr, "Fairfed: Enabling group fairness in federated learning," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 37, no. 6, 2023, pp. 7494–7502.

[5] T. Salazar, H. Araújo, A. Cano, and P. H. Abreu, "A survey on group fairness in federated learning: Challenges, taxonomy of solutions and directions for future research," 2024.

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

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

[8] V. Valadi, X. Qiu, P. P. B. de Gusmão, N. D. Lane, and M. Alibeigi, "FedVal: Different good or different bad in federated learning," in *32nd USENIX Security Symposium (USENIX Security 23)*, Aug. 2023, pp. 6365–6380.

[9] M. Yurochkin, M. Agarwal, S. Ghosh, K. Greenewald, N. Hoang, and Y. Khazaeni, "Bayesian nonparametric federated learning of neural networks," in *International conference on machine learning*. PMLR, 2019, pp. 7252–7261.

[10] Y. Yang, B. Hui, H. Yuan, N. Gong, and Y. Cao, "PrivateFL: Accurate, differentially private federated learning via personalized data transformation," in *32nd USENIX Security Symposium (USENIX Security 23)*, 2023, pp. 1595–1612.

[11] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y. Arcas, "Communication-Efficient Learning of Deep Networks from Decentralized Data," in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, A. Singh and J. Zhu, Eds., vol. 54. PMLR, 20–22 Apr 2017, pp. 1273–1282.

[12] T.-M. H. Hsu, H. Qi, and M. Brown, "Measuring the effects of non-identical data distribution for federated visual classification," *arXiv preprint arXiv:1909.06335*, 2019.

[13] J. Wang, Q. Liu, H. Liang, G. Joshi, and H. V. Poor, "Tackling the objective inconsistency problem in heterogeneous federated optimization," *Advances in neural information processing systems*, vol. 33, pp. 7611–7623, 2020.

[14] U. Gupta, J. Dhamala, V. Kumar, A. Verma, Y. Pruksachatkun, S. Krishna, R. Gupta, K.-W. Chang, G. V. Steeg, and A. Galstyan, "Mitigating gender bias in distilled language models via counterfactual role reversal," *arXiv preprint arXiv:2203.12574*, 2022.

[15] Z. Zhang, F. Yang, Z. Jiang, Z. Chen, Z. Zhao, C. Ma, L. Zhao, and Y. Liu, "Position-aware parameter efficient fine-tuning approach for reducing positional bias in llms," *arXiv preprint arXiv:2404.01430*, 2024.

[16] A. Mishra, G. Nayak, S. Bhattacharya, T. Kumar, A. Shah, and M. Foltin, "Llm-guided counterfactual data generation for fairer ai," in *Companion Proceedings of the ACM on Web Conference 2024*, 2024, pp. 1538–1545.

[17] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, "Learning fair representations," in *International conference on machine learning*. PMLR, 2013, pp. 325–333.

[18] F. Calmon, D. Wei, B. Vinzamuri, K. Natesan Ramamurthy, and K. R. Varshney, "Optimized pre-processing for discrimination prevention," *Advances in neural information processing systems*, vol. 30, 2017.

[19] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, "Certifying and removing disparate impact," in *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, 2015, pp. 259–268.

[20] J. Yang, R. Shi, D. Wei, Z. Liu, L. Zhao, B. Ke, H. Pfister, and B. Ni, "Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification," *Scientific Data*, 2023.

[21] F. Juraev, M. Abuhamad, E. Chan-Tin, G. K. Thiruvathukal, and T. Abuhmed, "Unmasking the vulnerabilities of deep learning models: A multi-dimensional analysis of adversarial attacks and defenses," in *2024 Silicon Valley Cybersecurity Conference (SVCC)*. IEEE, 2024, pp. 1–8.