Federated Learning on Heterogeneous Sensor Data

MALACHI MCKNIGHT*, University of Massachusetts, Amherst
KENNETH LIN, University of Massachusetts, Amherst
OM TANK, University of Massachusetts, Amherst

Federated Learning (FL) is a fundamental methodology in applications ranging from autonomous systems to smart cities and healthcare, as it provides a decentralized method for training machine learning models while protecting data privacy. Adversarial attacks and non-IID data distributions are FL's two main obstacles, though, since both impede global model convergence and limit generalization capabilities. The effects of data heterogeneity and adversarial attacks within FL systems are the primary concerns of this study's investigation of these limitations. Dirichlet alpha values are used to partition data at different levels of heterogeneity based on benchmark datasets (MNIST, SVHN, FashionMNIST, CIFAR100, and Adult); more heterogeneity is indicated by smaller alpha values. To evaluate the relationship between homogeneity, attack resilience, and learning performance, the study employs a full-knowledge trim attack that targets 40% of clients and manipulates 60% of their data. The findings show that although heterogeneity promotes model variety but raises susceptibility and computational cost, homogeneity improves communication efficiency and lessens adversarial impact but runs the risk of overfitting and reduced generalization. This paper examines global model accuracy in both "No Attack" and "Attack" situations, revealing significant compromises while providing suggestions for creating reliable, effective FL systems that can function well in hostile, dynamic, and heterogeneous environments.

CCS Concepts: • **Theory of computation → Multi-agent learning**; *Adversarial learning*.

Additional Key Words and Phrases: Federated learning, Heterogeneous data, Adversarial attacks

1 Introduction

Federated Learning (FL) is a novel machine learning technique that enables several people to work together by enabling them to train a common model in a dispersed fashion. This strategy has proven helpful in a number of industries where data security and laws are essential, including banking, health, and other smart devices. Adversarial assaults and disparities in the data of the participating nodes are still the two primary issues, nevertheless. Malicious actors who purposefully provide non-IID data distributions to create suboptimality, together with differences in data distributions across nodes, or non-IID distributions, restrict the amount of loss that can be achieved and limit how well models can be trained.

As a foundation for an extensive spectrum of applications in smart cities, healthcare, autonomous systems, and industrial automation, sensor data is essential in the modern world. Federated learning is especially appealing in settings where centralized data collection is impracticable or presents privacy concerns since it can use sensor data from diverse and dispersed sources. However, to guarantee the accuracy and reliability of the models learned via federated learning, sensor data robustness in unfavorable environments–such as those with noise, unstable networks, or malevolent interference–is indispensable. Compounding the difficulties of non-IID distributions are ambient conditions, hardware constraints, and inconsistent sensor performance, which can cause significant discrepancies in the data. To fully realize the potential of federated learning for sensor networks, these shortcomings must be resolved. This will make it possible for the creation of robust systems that can function successfully in dynamic and frequently hostile environments. This reinforces the necessity of methods that preserve computational and transmission efficiency while improving the robustness and integrity of sensor data.

This study attempts to investigate the effects of adversaries and data participation distribution in federated learning with the objective to mitigate such problems. We examine the trade-offs between average consensus convergence speed, generalization performance, and computational efficiency using datasets including SVHN, FashionMNIST, CIFAR100, and Adult. For the purpose of identifying the system's weakness, a full-knowledge trim attack is simulated, and the consequences of controlling the homogeneity levels using the Dirichlet distribution are also examined. This will assist us in creating appropriate federated learning strategies that enable us to produce refined, scalable, and confidential computations that are resistant to malevolent or hostile attacks.

1.1 Motivation

Federated learning is becoming more and more popular as a result of the growing need for collaborative machine learning techniques that protect individual privacy. However, real-world implementations reveal crucial trade-offs in a larger system. Even though homogeneous data speeds up convergence and reduces mild adversarial effects, it weakens generalization and tolerance to unknown and highly variable data. Conversely, heterogeneous data improves information customization and diversity, but it also raises transmission costs and makes it more challenging to avoid adversarial attacks.

The imperative to manage these trade-offs rigorously is what motivates us. The influence of data homogeneity on model performance between the adversary's presence and absence is the primary focal point of this study. We examine how attack routes, system stability, and heterogeneity interact with a concentration on a full-knowledge trim attack. Understanding these factors proves essential for developing FL systems that do not significantly compromise the systems' accuracy, resilience, and efficiency or the privacy of their users.

1.2 Problem Statement

Despite the privacy-driven nature of federated learning, the partitioning of data among clients may not always strike a compromise between preserving the unpredictability or untraceability of individual client habits and achieving effective model learning. It could be detrimental to single out a certain group of clients for an adversarial attack if they have an incumbent set of data for the model that

*Code used for this project: https://github.com/malachimcknight/ECE-535

.

differs from the other clients. Conversely, a minimal attack is less successful if all data samples are approximately the same across all clients. Our goal is to show that, depending on the values that partition the data according to a Dirichlet distribution, each dataset that we have chosen is either vulnerable to a significant attack or to deflection and rebounding. We seek to demonstrate the impact of data partitioning on learning rate and attack success by developing an adversary that can choose a subset of clients and launch an attack on their data, repeated with different degrees of heterogeneity.

1.3 System Specifications

We have selected the following datasets: MNIST, SVHN, FashionMNIST, Adult, and CIFAR-100. A binary annual revenue classification of either ≤ \$50k or > \$50k (2 classes) makes up the Adult dataset. There are 100 classes in the object recognition dataset CIFAR-100. FashionMNIST is a ten-class classification system for clothing items. There are real-world images of the house numbers (10 classes) on Street View House Numbers (SVHN). Finally, MNIST is a dataset of handwritten numbers (10 classes). After normalizing the data and aggregating the classes for every worker, we initiate various heterogeneous partitions by varying the Dirichlet alpha levels [0.1, 0.3, 1.0, 5.0, and 10.0]. Data that is highly heterogeneous would have a small alpha value, whereas data that is more homogeneous would have a higher alpha value. We have our adversary choose 40% of the clients and attack 60% of their data with a full-knowledge trim attack. As each model learns based on the Dirichlet alpha values and the presence of an attack, we can then determine the accuracy of each dataset.

2 Literature Review

Our methodology primarily leverages research on communica-tion-efficient learning. In contrast to stochastic gradient descent (SGD), [McMahan et al. 2017] provide FederatedAveraging as an optimization technique for the analysis of unbalanced and non-IID data distributions. In order to ascertain how the partitioning of each dataset would learn with or without some adverse circumstance, we specifically aimed to categorize the effects of heterogeneous data within a federated learning framework and augment an adversarial approach. Furthermore, our findings benefited from the privacy considerations in federated learning, which aggregate gradients to obfuscate information about specific targets.

According to [Kumar et al. 2024], the adversarial attack survey offers a thorough examination of adversarial vulnerabilities in federated learning settings. In addition to classifying attacks according to their source and technique—such as backdoor manipulation, gradient inversion, or targeted poisoning—the study also divides them into utility-centric and privacy-centric categories. The study employs an empirical methodology to assess these attacks in three dimensions: visibility, impact, and resource budget. It does this by exposing trade-offs through the use of visualization tools such as 3D scatterplots. Case examples from the real world show how vulnerable FL systems are to both typical and unique attacks. They used global test accuracy to evaluate utility-centric attacks and mean squared error and structural similarity index to evaluate privacy-centric threats. The survey also examines protection measures, such as aggregation techniques, adversarial training, and differential privacy, and critically evaluates how well they work against different types of attacks.

[Darzi et al. 2024] explore the particular difficulties of using adversarial attacks in FL systems for processing medical images. The work addresses the susceptibility of FL models trained on domain-specific datasets such as brain MRI and histopathologic cancer images using a variety of adversarial techniques, including FGSM, BIM, and PGD, as well as more sophisticated strategies like C&W and Distributionally Adversarial Attack (DAA). In order to assess system resilience, the experimental setup includes a convolutional neural network architecture, a non-IID data distribution across clients, and differential privacy techniques including Gaussian noise addition and gradient clipping. They perform a quantitative analysis of attack success rates (ASR) and transferability parameters, including perturbation degree and norm selection.

In order to counteract adversarial attacks in federated learning, [Queyrut et al. 2023] proposed PELTA, a defense mechanism that makes use of Trusted Execution Environments (TEEs). The authors discuss the difficulties of gradient-based evasion assaults in white-box environments, where attackers create adversarial samples that undermine the FL model by taking advantage of the backpropagation process. By safely keeping important gradient-related variables, including Jacobian matrices, behind TEEs like ARM TrustZone, PELTA partially obscures the gradient flow in a model's first layers. This significantly blocks the attackers' attempts by interfering with their ability to compute gradients for perturbation creation. Because of its lightweight design, PELTA has no secure memory overhead and is compatible with cutting-edge designs such as Vision Transformers (ViTs) and Convolutional Neural Networks (CNNs). The assessment demonstrates notable resilience against several types of adversarial assaults, including the Self-Attention Gradient Attack (SAGA) and Projected Gradient Descent (PGD).

2.1 Design Alternatives

From our study of relevant federated modeling techniques, although not as sophisticated, we took the analysis approach under the condition of specific partitioning through heterogeneous to homogeneous client sampling, where other research did not particularly delve. Our use of a full-knowledge trim attack serves as an example of something easily comprehensible. An attack that would generally be more effective but much less straightforward is data poisoning. We aggregated the data and standardized the datasets we used after the other research; however, to be more thorough, we have selected five datasets. The scope of the data and the outcome of executing an isolated attack on a global client set are much more significant in conveying the adverse effects of heterogeneous data in a federated learning environment, even if our approach was more straightforward.

3 Background

Many clients work together to jointly train a global model using Federated Learning (FL), a decentralized machine learning technique, without exposing their raw data. Rather, clients use private data to train local models and communicate updates, such gradients or

parameters, to a central server. This configuration improves privacy and adherence to regulations such as GDPR and HIPAA.

Data heterogeneity, or non-IID (independent and identically distributed) data among clients due to different user behaviors, devices, or settings, is a troubling obstacle in FL. This hinders the global model's ability to converge and generalize. FL is less successful when models trained on skewed data distributions perform poorly on the entire population.

Threats from adversaries exacerbate these problems. Adversaries can alter datasets or updates in FL in order to interfere with the global model. The full-knowledge trim attack, for example, eliminates particular data classes during training, increasing the risks brought on by heterogeneity and lowering the accuracy of the model. It requires one to carefully weigh the trade-offs between security, performance, and efficiency while trying to address such hazards.

This research investigates these dynamics using benchmark datasets. While the binary Adult dataset investigates income categorization, datasets such as SVHN and FashionMNIST (10-class tasks) and CIFAR100 (100-class task) evaluate FL's performance on visual identification tasks. Through the use of a Dirichlet distribution, these datasets enable us to investigate FL systems at different levels of heterogeneity.

We measure global model accuracy under "No Attack" and "Attack" scenarios to assess the system's resistance to adversarial attacks. This approach draws attention to important trade-offs: homogeneity lowers communication complexity but runs the risk of model redundancy, whereas heterogeneity frequently enhances personalization and generalization but also makes systems more vulnerable to attacks. The research presented here provides perspective on how adversarial robustness and heterogeneity interact to create FL systems that strike a balance between security, performance, and privacy.

4 System Design

A central server processes the model that is trained by several clients in the framework of centralized federated learning. In a comparable way, we set up 100 clients for the datasets, each of which receives distinct data partitioned for training and evaluation. After partitioning the data according to the heterogeneous levels indicated by the Dirichlet values we have chosen, we aggregate the data. This allows us to compare the impact of more homogeneous division with that of heterogeneous data. A new model is trained on its own workers for each Dirichlet value, making it straightforward to evaluate the intended demonstration of several heterogeneous levels. We have an adversary that targets clients directly in addition to our centralized structure. To prevent bias in choosing something that would be deemed essential to the learnt models, each attack is conducted on a controlled percentage that is chosen at random by the algorithm. This makes it clear that while heterogeneity varies, the datasets themselves are vulnerable to their own threats (Fig. 1).

When training their model, individual clients are most vulnerable to attacks of some kind. It would not be very common in our situation for an adversary to be able to select any number of clients from the entire list. To support the impact of heterogeneous data under adverse conditions, we differentiate this study from others. When

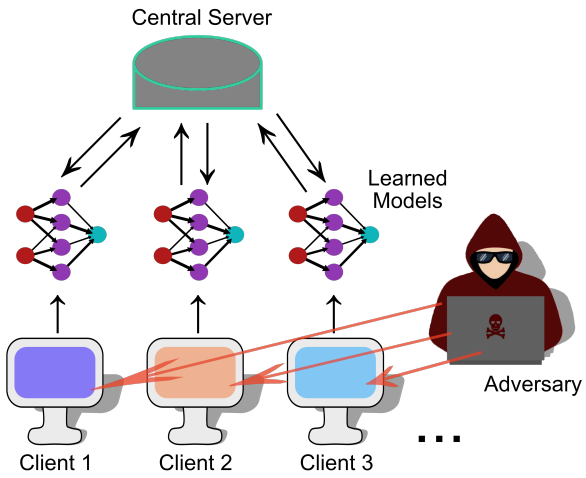

Fig. 1. An adversary targets the individual clients in a federated learning environment.

analyzing the accuracy of each dataset, the aggregated learning models would function as a central server. In general, our approach uses a centralized federated learning model's predetermined structure to exploit different partitioning situations with and without an adversary, rather than trying to alter the model's format.

5 Implementation

This study is entirely software-based and depends on the FedLab datasets and Python. For every dataset, we used a different Jupyter notebook for all of our programming. We would first import the specific dataset, normalize it, and then partition the data according to the Dirichlet values we had chosen. We would train on each partition's workers using a simple CNN. The models would be aggregated and their accuracy assessed after training. As a global accuracy metric, we calculated accuracy using the average of the accuracy and the rate of learning for each epoch. To ascertain how well the datasets would withstand an assault of some kind, we would also have the adversary execute a full-knowledge trim attack on the same data sampling (Fig. 2).

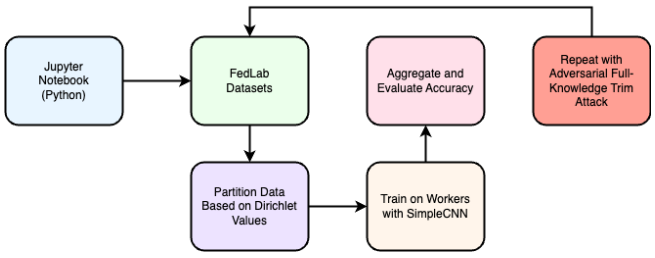

Fig. 2. Block setup of a federated learning environment with the addition of an adversary.

We simplified the process of processing all the partitioning by running for a restricted amount of epochs in order to compare our

five selected datasets. Since demonstrating heterogeneous data partitioning with and without adversarial conditions is more important than spending time training a model without any analysis, overall accuracy is presumed not to be substantially affected. Our expertise with straightforward machine learning techniques led us to use a simple CNN as a way to ensure repeatability across the different tiling factors of the datasets we studied. Furthermore, we were only able to compute the algorithms using our own dedicated GPU, whereas Google Colab merely offered a far less efficient use of computing power.

6 Evaluation

Our selection of the Adult, MNIST, CIFAR-100, FashionMNIST, and SVHN datasets offers a comprehensive perspective on the impact of heterogeneous data by revealing a wide range of insights regarding when specific partitions can be harmful or resilient enough to withstand an attack. Regardless of the heterogeneous partition, datasets with few classes are the most resistant to attack. This is because, compared to the entire model set, each client may only slightly skew the data from a small portion of what they collect. In the opposing situation, datasets with a significant number of classes are challenging to train, making it complicated to conduct an assault successfully. Heterogeneous partitioning and, hence, an adversarial attack can most effectively exploit datasets that fall somewhere in the center, neither large in terms of classes nor arduous to train.

For each dataset, we have the mean aggregated model accuracy for each Dirichlet alpha. Furthermore, we contrast the accuracy with the results of an attack on the same partitioned collection. In order to demonstrate the distinction between heterogeneous levels and other partition levels that approach homogeneous data partitioning, we have used this metric.

To determine how successfully a model can be trained and how introducing an adversarial attack can hinder its learning rate, we also examine the accuracy across each epoch. The individual, aggregated Dirichlet values, both with and without an attack, serve as the basis for each accuracy. We have chosen this to draw attention to the impact of heterogeneous data and better illustrate the resilience of some datasets.

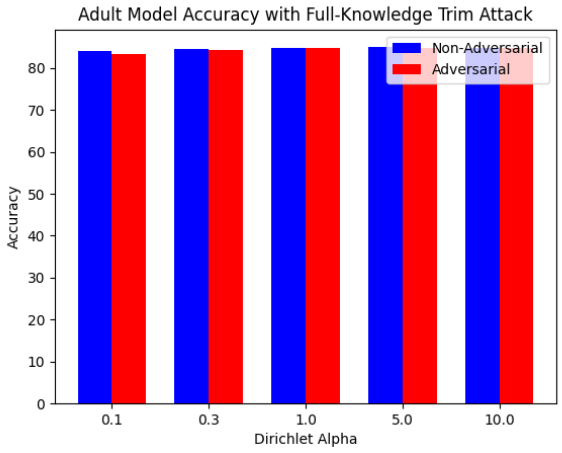

Fig. 3. Aggregated results from Adult dataset with and without an attack.

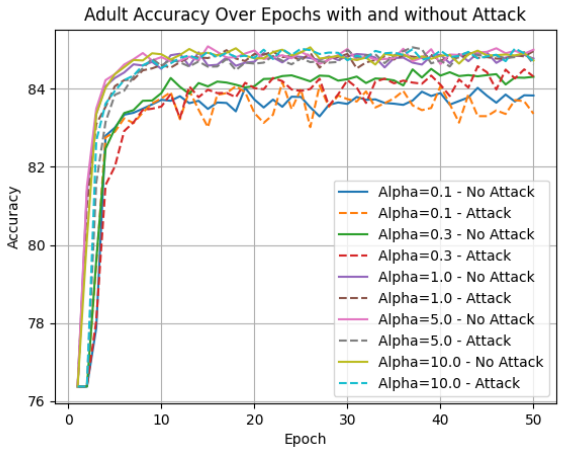

Fig. 4. Learning rate over 50 epochs with and without an attack for Adult dataset.

6.1 Results

Considering there are only two classes in the Adult dataset, it is an effective model to train. Regardless of the Dirichlet alpha, the accuracy is above 80% when using 50 epochs (Fig. 3). Furthermore, the aggregated model to learn appears to be unaffected by the full-knowledge trim attack. It is also demonstrated that the learning accuracy over epochs (Fig. 4) converges in a limited amount of time steps. It would be clear from this that the Adult dataset has little effect on the degree of client heterogeneity because of its binary classification. Similarly, an attack on this dataset would not eliminate any important information that presumably a different client might have.

The MNIST dataset is also a reasonably balanced set that appears to be unaffected by adversarial attacks or heterogeneous data. The accuracy is above 90% for each Dirichlet alpha, according to the mean accuracy (Fig. 5). Once more, even when heterogeneity is at its highest, the adversarial attack has minimal effect. MNIST exhibits relatively slow convergence for the learned accuracy (Fig. 6). But this makes it more evident that high heterogeneity, or an alpha of 0.1, is detrimental to model learning, and that adding an attack further reduces it.

The CIFAR-100 dataset makes it clearer that having an abundance of classes results in poor accuracy and makes learning more challenging. For all Dirichlet alphas, the average accuracy for no attack is nearly below 40% (Fig. 7). A further consequence of the inverse trend of low accuracy with high heterogeneous data is the more apparent nature of the adversarial attack. This dataset's learning rate begins at the lowest accuracy and gradually rises over the course of 30 epochs (Fig. 8). It is evident that the attack at alpha equal to 0.1 differs from the approximate grouping of the remaining conditions. This would suggest that having exceptionally heterogeneous data

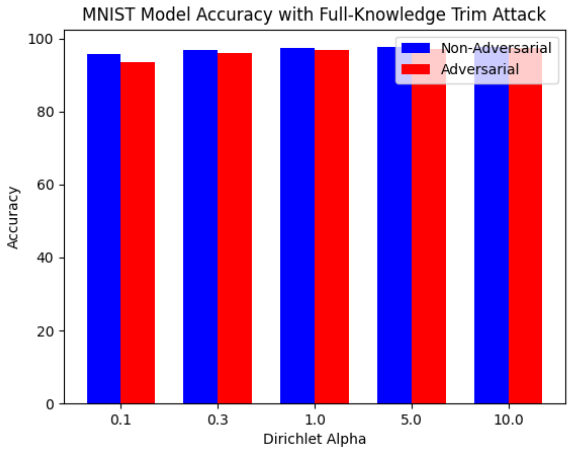

Fig. 5. Aggregated results from MNIST dataset with and without an attack.

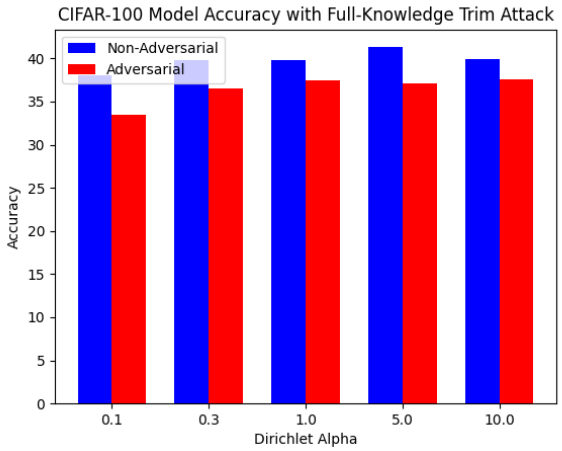

Fig. 7. Aggregated results from CIFAR-100 dataset with and without an attack.

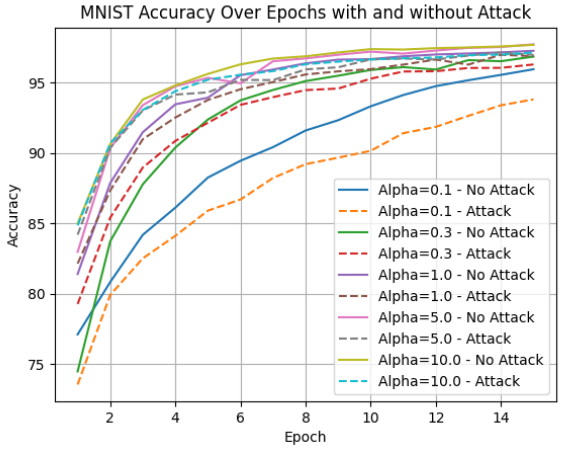

Fig. 6. Learning rate over 15 epochs with and without an attack for MNIST dataset.

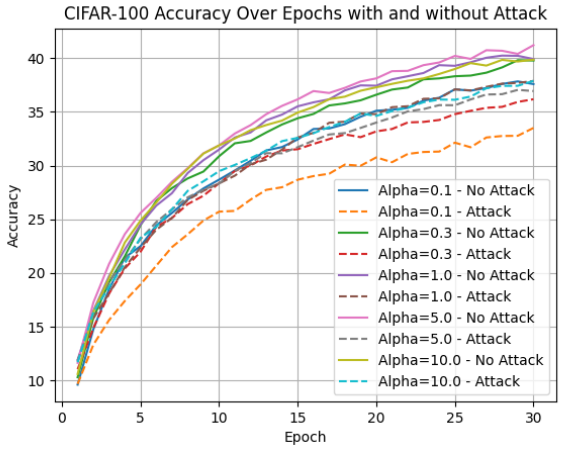

Fig. 8. Learning rate over 30 epochs with and without an attack for CIFAR-100 dataset.

would be most vulnerable to an effective assault, even with poor overall accuracy.

Despite being more balanced overall, FashionMNIST exhibits the effects of heterogeneous data partitioning. As client data becomes more uniform, the mean accuracy exhibits a trend toward increasing accuracy (Fig. 9). This suggests that even at the expense of overfitting, low levels of heterogeneity do, in fact, permit faster learning convergence. Although the attack is extremely minor yet closely resembles normal conditions, high levels of heterogeneity have poorer accuracy performance when the learning rate is within 15 epochs (Fig. 9).

One of the clearest examples of the impact of heterogeneous data partitioning is the Street View House Numbers (SVHN) dataset. As the data becomes more uniform, there is a predicted trend for the mean accuracy to improve (Fig. 11). High heterogeneity makes the performance under adverse circumstances the most revealing. Our chosen dataset is particularly vulnerable to attacks in which client data is not uniform. Accuracy across epochs (Fig. 12) shows how adding an attack to increasing heterogeneity severely hampers the model's ability to learn.

7 Limitations

The expectations made during the design phase, such as the homogeneity of devices, steady environmental conditions, and consistent resource availability, determine how adaptable our strategy is to various scenarios. Various resource limitations, dynamic situations, and heterogeneous devices are common in real-world systems, which may lessen the effectiveness of our approach. Furthermore, our suggested methods' computational complexity might make them

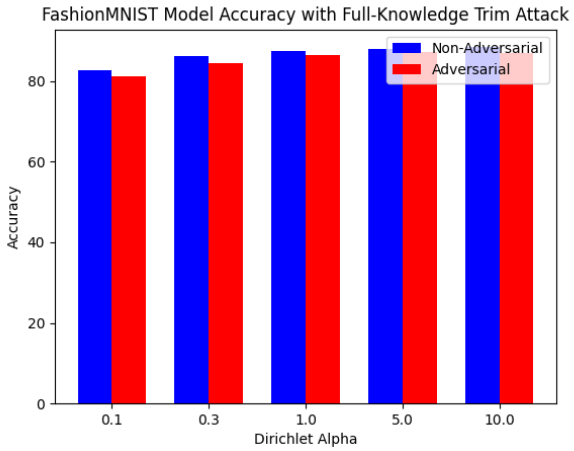

Fig. 9. Aggregated results from FashionMNIST dataset with and without an attack.

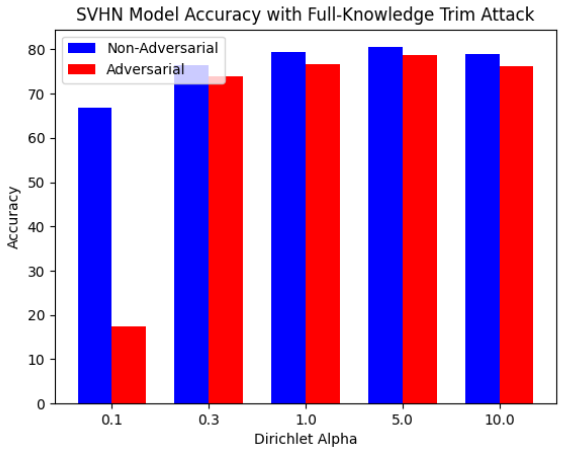

Fig. 11. Aggregated results from SVHN dataset with and without an attack.

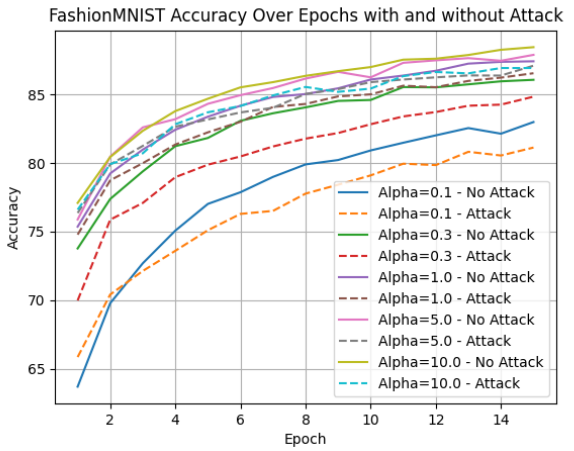

Fig. 10. Learning rate over 15 epochs with and without an attack for FashionMNIST dataset.

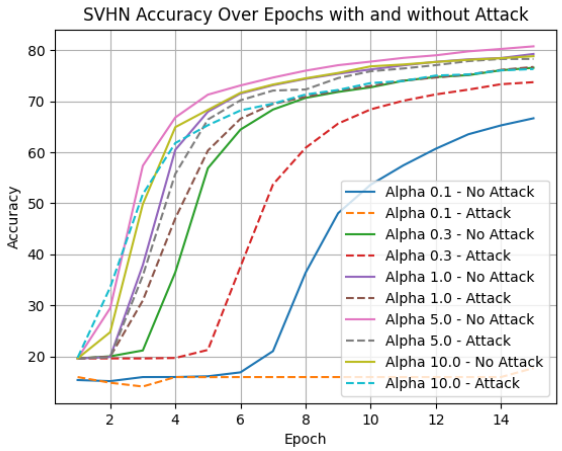

Fig. 12. Learning rate over 15 epochs with and without an attack for SVHN dataset.

unsuitable for contexts with severe resource constraints, such low-power edge devices. Additionally, the results are extensively reliant on the quality of the input data; performance may be greatly impacted by noisy or insufficient observations. In addition, scalability is still an issue since the techniques may not perform well in larger-scale deployments with an abundance of linked devices, requiring more improvement for wider use.

8 Conclusion

The influence of adversarial attacks and data heterogeneity on federated learning systems is explored in this research, which offers significant perspectives into the difficulties encountered while implementing FL in practical settings. We discovered that heterogeneity, which is reflected in different Dirichlet alpha values, affects model performance in two distinct manners. In one way, it permits the model to become more diverse and personalized, which improves its ability to generalize across many data sources. However, it also makes clients more susceptible to hostile attacks that alter data and hamper learning, such as the full-knowledge trim attack.

Our findings show that whereas homogeneity increases communication effectiveness and lowers the chance of adversarial influence, it comes at the expense of generalization, which may result in overfitting and subpar performance on unknown data. This trade-off points out how FL systems need to carefully balance attack resilience with data distribution. Our results additionally demonstrate the necessity of strong federated learning frameworks that can protect data integrity and privacy in dynamic, varied, and hostile contexts.

8.1 Future Directions

Future work should concentrate on creating more robust FL frameworks that can manage significant levels of heterogeneity and adversarial threats at the same time, building on the results of this study. Enhancing model robustness without sacrificing efficiency or privacy can be achieved by integrating refined defense mechanisms like adversarial training, differential privacy, or secure aggregation techniques. Furthermore, more research is required to fine-tune the effects of various adversarial attack types in a variety of application domains, such as autonomous systems, smart cities, and healthcare, where FL has great potential.

A more thorough examination of the interactions between various attack techniques, including model inversion or data poisoning, and varying degrees of data homogeneity may provide guidance for developing more scalable and safe systems. Additionally, investigating hybrid strategies that blend decentralized and centralized training techniques may provide a means of reducing the dangers of adversarial attacks as well as the difficulties brought on by non-IID data distributions. Furthermore, expanding the assessment to larger, more intricate datasets and genuine FL contexts will be essential to confirming the efficacy of these strategies in real-world applications.

References

Erfan Darzi, Florian Dubost, Nanna. M. Sijtsema, and P.M.A van Ooijen. 2024. Exploring Adversarial Attacks in Federated Learning for Medical Imaging. *IEEE Transactions on Industrial Informatics* 20, 12 (2024), 13591–13599. https://doi.org/10.1109/TII.2024.3423457

Kummari Naveen Kumar, Chalavadi Krishna Mohan, and Linga Reddy Cenkeramaddi. 2024. The Impact of Adversarial Attacks on Federated Learning: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46, 5 (2024), 2672–2691. https://doi.org/10.1109/TPAMI.2023.3322785

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. 2017. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (Proceedings of Machine Learning Research, Vol. 54)*, Aarti Singh and Jerry Zhu (Eds.). PMLR, 1273–1282. https://proceedings.mlr.press/v54/mcmahan17a.html

Simon Queyrut, Valerio Schiavoni, and Pascal Felber. 2023. Mitigating Adversarial Attacks in Federated Learning with Trusted Execution Environments. In *2023 IEEE 43rd International Conference on Distributed Computing Systems (ICDCS)*. 626–637. https://doi.org/10.1109/ICDCS57875.2023.00069