Performance Analysis of Aggregation Algorithms in Cross-Silo Federated Learning for Non-IID Data

Mathis Delehouzée
Faculty of Engineering—ILIA/Infortech
University of Mons
Rue de Houdain, 9
Mons, 7000, Belgium
Email: mathis.delehouzee@umons.ac.be

Xavier Lessage
CETIC - Applied Research Center
Av. Jean Mermoz, 28
Charleroi, 6041, Belgium
Email: xavier.lessage@gmail.com

Théo Reginster
Faculty of Engineering—ILIA/Infortech
University of Mons
Rue de Houdain, 9
Mons, 7000, Belgium
Email: theo.reginster@student.umons.ac.be

Saïd Mahmoudi
Faculty of Engineering—ILIA/Infortech
University of Mons
Rue de Houdain, 9
Mons, 7000, Belgium
Email: said.mahmoudi@umons.ac.be

***Abstract*—Federated learning is a powerful machine learning paradigm that enables a large number of machine learning applications that must comply with strict and complex data privacy regulations. In this paper, we focus on one of the fundamental components of Federated Learning: federated aggregation algorithms. These algorithms play a pivotal role in consolidating insights and the model updates from various clients while preserving data privacy and security. We compare their efficacy across different types of dataset configurations, including balanced IID (Independent and identically distributed) data, unbalanced IID data, and non-IID data characterized by label distribution skew and feature distribution skew. For our specific use case, the experiments presented in this work show that FedProx is the overall best performing state-of-the-art algorithm for binary classification tasks of medical X-rays distributed in datasets of a small number of hospitals.**

***Index Terms*—Federated Learning, Aggregation Algorithms, Non-IID Data, Machine Learning**

I. Introduction

Federated Learning [1] is a pioneering framework in machine learning that enables collaborative model training across multiple entities while maintaining data privacy. Unlike traditional centralized learning approaches that store and process data on a central server, FL operates on a decentralized model. In this paradigm, training data remains distributed across various participants, known as federated clients, which helps to comply with stringent data privacy regulations. However, this decentralized approach introduces several challenges, including handling non-Independent and non-Identically Distributed (non-IID) data, ensuring efficient communication, and maintaining the privacy and security of the data.

One of the critical components of FL is the federated aggregation algorithm, which consolidates model updates from different clients to form a global model. The choice of aggregation algorithm can significantly impact the performance and stability of the federated learning process. This study focuses on evaluating various federated aggregation algorithms, such as FedAvg, FedProx, and others, which were chosen for their prominence and effectiveness as reported in existing literature. These algorithms represent a spectrum of approaches to addressing the unique challenges of federated learning, including handling non-IID data distributions and ensuring robust convergence.

FL has applications in various domains, including healthcare, finance, and mobile applications. For instance, in healthcare, FL enables collaborative training of models across hospitals while ensuring patient data privacy. In the finance sector, FL allows institutions to collaboratively detect fraud patterns without sharing sensitive customer data. The mobile industry leverages FL to improve user experiences by training models on-device, thus preserving user privacy [2].

Non-IID data poses specific challenges in federated learning because the data held by different clients can vary significantly in distribution. This heterogeneity can lead to instability and poor performance if not properly managed by the aggregation algorithm. For instance, clients may have different label distributions or feature distributions, which can cause the global model to converge more slowly or inaccurately. Addressing these challenges requires careful selection and tuning of aggregation algorithms to ensure they can effectively manage the variability and complexity of real-world data distributions.

This study aims to conduct a comparative analysis of these aggregation algorithms to determine their efficacy across different dataset configurations. By understanding how these algorithms perform in various scenarios, particularly in medical imaging applications where data privacy is paramount, we can identify the most suitable approaches for specific use cases.

979-8-3315-1987-2/24/$31.00 ©2024 European Union

II. METHOD

This section presents the methods and experiments conducted to determine the most suitable federated learning aggregation algorithm for our use case, specifically a binary classification problem of X-ray images distributed across different databases from a small number of hospitals. A simple Convolutional Neural Network (CNN) was employed as the local learning algorithm on each client, as it is well-suited for image classification tasks and provides a solid baseline for evaluating aggregation methods.

We first introduce the experimental dataset, with several artificially generated configurations created from it to simulate different distribution scenarios. Next, we select and evaluate several state-of-the-art aggregation algorithms. Specifically, we tested FedAvg, FedProx, qFedAvg, and Scaffold on seven artificially generated dataset configurations. For FedProx and qFedAvg, experiments were conducted using three different values of their respective hyperparameters (μ for FedProx and q for qFedAvg), resulting in a total of 56 federated learning simulations. The candidate set $\{0.1, 0.5, 1.0\}$ was chosen for μ in FedProx based on previous studies [3], while the set $\{1\times10^{-4}, 1\times10^{-5}, 1\times10^{-6}\}$ was selected for q in qFedAvg after preliminary experiments [4]. Five federated clients were simulated, each training a CNN model for 5 local epochs per communication round using a stochastic gradient descent (SGD) optimizer with no momentum. Each simulation ran for 100 communication rounds, and the best hyperparameter values were selected based on performance.

A. Selection of the Experimental Dataset

For the purpose of this project, a publicly available chest X-ray image dataset was selected to conduct experiments and evaluate results. The dataset, composed of healthy chest X-rays and X-rays of patients with pneumonia, was chosen because it is suitable for binary image classification tasks. The data used in both cases is of similar nature, generated using the same technology and in the same hospital environment with strict procedures.The dataset consists of X-rays used to classify pneumonia, as described in [5].

To ensure the dataset's cleanliness and relevance for this study, all X-rays were selected based on quality requirements, and low-quality or unreadable images were removed. The diagnoses for each X-ray were evaluated by two experts, with these evaluations checked by a third expert to ensure no grading errors were made.

B. Federated aggregation algorithms

This section provides an overview of the current landscape of federated learning aggregation algorithms, beginning with the inaugural federated learning algorithm developed by Google in 2016 [6], and then presenting the most cutting-edge algorithms. The key characteristics of each algorithm are summarized in Table I for easier comparison.

1) FedAvg: Federated Averaging [6] is one of the earliest federated learning algorithms. It involves a central federated server performing model averaging based on the resulting models from multiple clients, which perform local SGD on their local private datasets. The objective function is:

$$\min_w f(w) = \sum_{k=1}^{K} \frac{n_k}{n} F_k(w) \tag{1}$$

where w is the global model parameter, $F(w)$ is the global objective function, K is the total number of participating clients, n_k is the number of samples held by the k-th client, n is the total number of samples held by all clients, and $F_k(w)$ is the local objective function of the k-th client.

2) FedProx: FedProx [7] introduces a proximal term to the local objective function of each client to limit the distance between the local model of each federated client and the global model. This helps improve performance and stability, particularly with non-IID data. The proximal term ensures each client converges towards the global objective rather than their local optima.

3) qFedAvg: q-FedAvg [8] addresses fairness by giving more weight to poorly performing clients. The objective function is:

$$\min_w f_q(w) = \sum_{k=1}^{K} \frac{\frac{n_k}{n}}{q+1} F_k^{q+1}(w) \tag{2}$$

Where F_k^{q+1} denotes the local cost function to the power of (q+1) of the client *k*. And *q* is the new introduced parameter that adjusts the level of fairness we want to impose. The higher *q* is, the higher the contribution to the global model of federated clients with a high local loss will be, and the higher the level of fairness of the training accuracy distribution will be.

4) FedNova: FedNova [9] normalizes local updates by scaling them according to the number of local steps taken by each client, addressing the issue of objective inconsistency in heterogeneous networks.However, we do not use FedNova in our study as it is designed for scenarios with heterogeneous clients having varying computation power, which is not applicable to our use case where the clients (hospitals) have relatively similar computational capacities.

5) Scaffold: Scaffold [10] estimates and corrects the local drift of clients to ensure their updates move towards the global optimum rather than local optima, improving performance in non-IID scenarios.

C. Artificial generation of dataset configurations

To compare different aggregation algorithms, several dataset configurations were artificially generated based on the chest X-ray dataset. These configurations simulate various types of data distributions and statistical heterogeneity across federated clients, representing real-life scenarios. The final federated learning configuration proposed in this work is used in a

TABLE I
SUMMARY OF FEDERATED AGGREGATION ALGORITHMS

Method	Key Characteristics	Practical Implications
FedAvg	Weighted average of local models	Simple and effective in IID data scenarios but struggles with non-IID data.
FedProx	Proximal term in the local objective function	Improves stability and performance with non-IID data, useful in heterogeneous client environments.
qFedAvg	Weighted average with fairness adjustment	Aims to provide fairer performance across clients, beneficial when some clients have significantly worse performance.
FedNova	Normalizes local updates by local steps	Addresses objective inconsistency, ensuring faster error convergence in heterogeneous networks.
Scaffold	Corrects local client drift	Ensures updates move towards the global optimum, improving performance in non-IID scenarios.

context of five federated clients (i.e., hospitals). Fig. 1 provides an overview of all the dataset configurations proposed in this work.

1) Configuration 1 - Balanced datasets & IID data: The first artificially generated dataset configuration represents a case with no statistical heterogeneity. In this configuration, the local datasets of all federated clients contain the exact same number of X-rays, indicating no quantity skew. Additionally, each client has a 25-75 label distribution (25% healthy), signifying no label distribution skew. Moreover, there is no feature distribution skew between the datasets. While this configuration is unlikely to occur in real-life scenarios, it will be used as a starting point for the comparison.

2) Configuration 2-3 - Unbalanced datasets & IID data : Two different configurations with different levels of unbalanced datasets are generated. Configuration #2 is less unbalanced and still relatively homogeneous with 827 data samples in the smallest dataset and 1246 in the largest one. Configuration #3 is more extreme with clients 1, 4, and 5 having disproportionately small datasets compared to clients 2 and 3. These two configurations still represent cases of IID data as all the federated clients have the same label distribution (25-75), regardless of the total size of their local dataset. There is no feature distribution skew between the datasets.

3) Configuration 4-5 - Label distribution skew (non-IID): Configurations #4 and #5 represent cases which combine both unbalanced datasets and non-IID data. These configurations specifically involve label distribution skew, meaning that the distribution of labels (healthy and pneumonia X-rays) is different for each local dataset. The configurations are only slightly unbalanced as the main goal is to analyze the effects of label distribution skew. Configuration #4 remains relatively homogeneous with the highest proportion of healthy X-rays at 39% (client 1) and the lowest at 17% (client 2). Configuration #5, on the other hand, represents a case where one of the hospitals (client 2) has a completely different label distribution than the others.

4) Configuration 6-7 -Feature distribution skew (non-IID): Configurations #6 and #7 introduce non-IID distributions among clients, specifically feature distribution skew. As explained previously, the distribution of features may be skewed by the fact that different hospitals may use different types of imaging equipment or different image acquisition parameters. This can result in significant differences in the X-rays generated between different hospitals.

To simulate feature distribution skew, the local datasets of a portion of the five clients were artificially degraded with varying levels of Gaussian noise or Gaussian blur. For Configuration #6, small levels of Gaussian noise were introduced to the datasets of clients 1 and 2. For Configuration #7, higher levels of Gaussian noise were applied to the datasets of clients 1 and 2, and Gaussian blur was added to the dataset of client 3. This process involved adding Gaussian noise with a specified mean and variance to the pixel values of the images and applying a Gaussian blur filter with a specified kernel size to create blurred images. These modifications were intended to mimic the variations in image quality that might result from different imaging equipment or settings used in different hospitals.

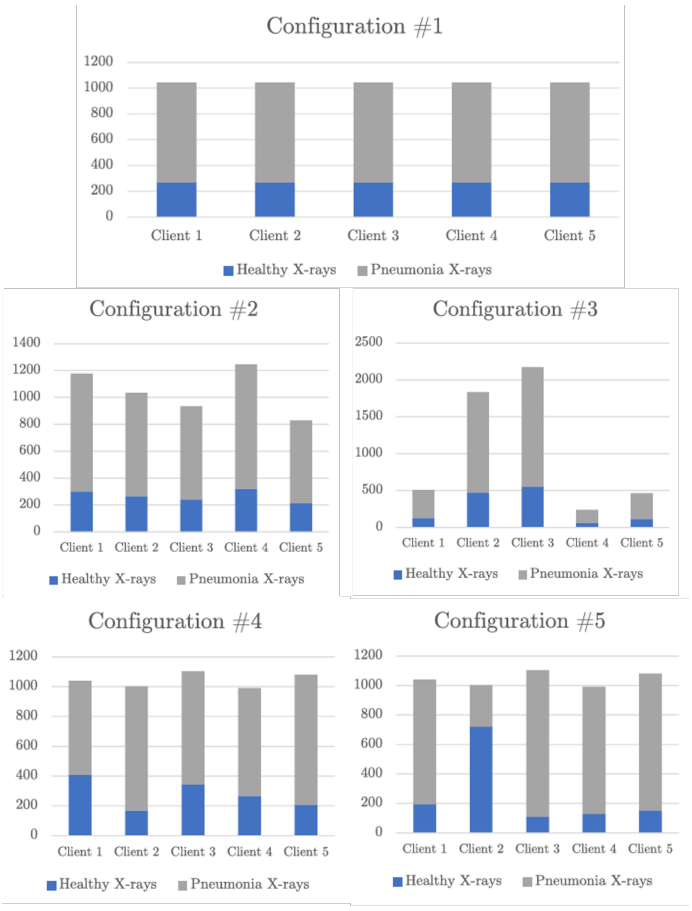

Fig. 1. Illustration of different dataset configurations

TABLE II
DETAILS OF FEATURE DISTRIBUTION SKEW CONFIGURATIONS

Client	Configuration #6	Configuration #7
C1	Gaussian noise (μ=0, σ^2=0.002)	Gaussian noise (μ=0, σ^2=0.005)
C2	Gaussian noise (μ=0, σ^2=0.005)	Gaussian noise (μ=0, σ^2=0.01)
C3	No modification	Gaussian blur (kernel size=5x5)
C4	No modification	No modification
C5	No modification	No modification

The application of Gaussian noise and blur is illustrated in Table II, which shows examples of modified X-rays. This approach allows us to evaluate the robustness of the aggregation algorithms under conditions that simulate real-world variability in data quality across different clients.

D. Selection of appropriate aggregation algorithms

The appropriate aggregation algorithms are selected based on six criteria: ability to perform in cross-device and cross silo settings, ability to handle IID and non-IID data, ability to provide privacy-friendly solution, and others. FedAvg, FedProx, qFedAvg, and Scaffold were selected as they meet the criteria. FedAvg may decrease in performance with non-IID data, but it is included as the level of heterogeneity in the use-case is small. FedNova was not selected as it is designed for heterogeneous clients with different computation power, which does not apply to the use-case.

E. Federated Learning benchmarking framework

The selected aggregation algorithms are benchmarked using the easyFL open-source framework developed in [11]. This framework is a research-oriented experimental platform that offers multiple reusable and adaptable modules, enabling researchers to conduct various experiments on both existing and new federated learning algorithms.

III. EXPERIMENTAL RESULTS

This section presents the results of the simulations for the different dataset configurations.

A. Balanced datasets & IID data

Fig.2 shows the results of the simulations for Configuration #1 for the last 40 communication rounds. Both FedAvg and FedProx quickly reached over 80% accuracy but experienced some stability issues around round 20, ultimately stabilizing around round 100. Scaffold and qFedAvg were slower to reach 80% accuracy and ultimately achieved lower final accuracies.

The fact that FedAvg and FedProx show relatively similar behavior can be explained by the fact that their algorithms are very similar. As explained in Section II.B, FedProx is a re-parametrization and generalization of FedAvg and only introduces a proximal term to the local objective function of each client. For a simple balanced and IID configuration, FedAvg provides the highest final accuracy.

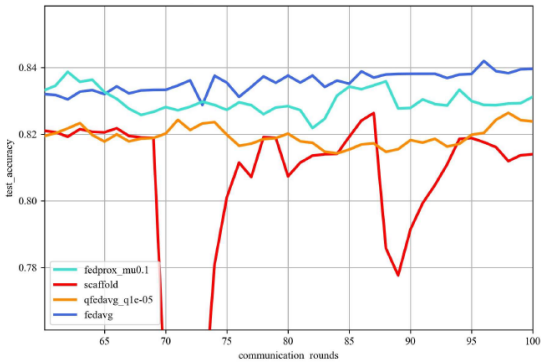

Fig. 2. Conf #1 - Testing accuracy for the last 40 communication rounds.

B. Unbalanced datasets & IID data

For Configurations #2, as illustrated in Fig.3, FedAvg and FedProx again showed higher accuracies compared to other algorithms. In Configuration #2, FedAvg performed better, while in Configuration #3, both algorithms showed nearly identical accuracies around 84%.

However, in Configuration #3, as the quantity bias increases (i.e., becomes more unbalanced), both algorithms exhibit nearly similar accuracies, with 84.13% and 84.29% respectively.

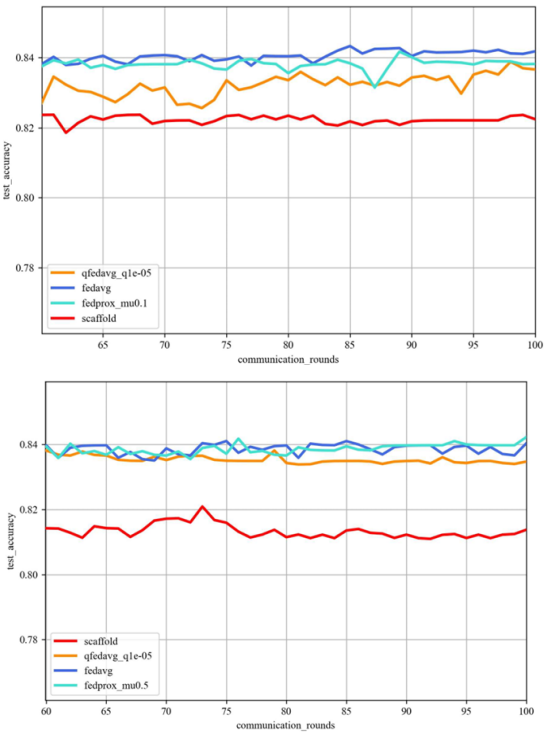

Fig. 3. Conf #2 & #3 Testing accuracy for the last 40 communication rounds.

C. Label distribution skew (non-IID)

Configurations #4 and #5 introduce statistical data heterogeneity, specifically label distribution skew (i.e., the distribution of labels varies across each local dataset). In both configurations, the test accuracies are significantly less stable (primarily during the first 50 communication rounds) compared to the first three IID configurations. This outcome is expected as aggregation algorithms must handle non-IID data distributions.

Another interesting result is that the FedAvg algorithm, which has consistently provided the highest (or second highest) accuracies, now shows the lowest accuracies at round 100 for both Configurations #4 and #5 (82.37% and 81.73% respectively), as shown in Fig.4. On the other hand, Scaffold, which previously yielded relatively poor results, has improved significantly and now provides accuracies nearly as good as FedProx.

D. Feature distribution skew (non-IID)

Configurations #6 and #7 also introduce non-IID distributions among clients, specifically feature distribution skew. As explained in Section II.C, these configurations are characterized by degrading a portion of the local datasets of the five clients with varying levels of Gaussian noise or Gaussian blur. Similar to the previous two configurations, and as shown in Fig.5, the test accuracies for these configurations are significantly less stable compared to the first three IID configurations. The stability of the accuracies in Configuration #7 is the worst because the datasets of three out of five clients are degraded, compared to only two clients, and with lower levels of degradation in Configuration #6, as shown in Table III in Section 4.

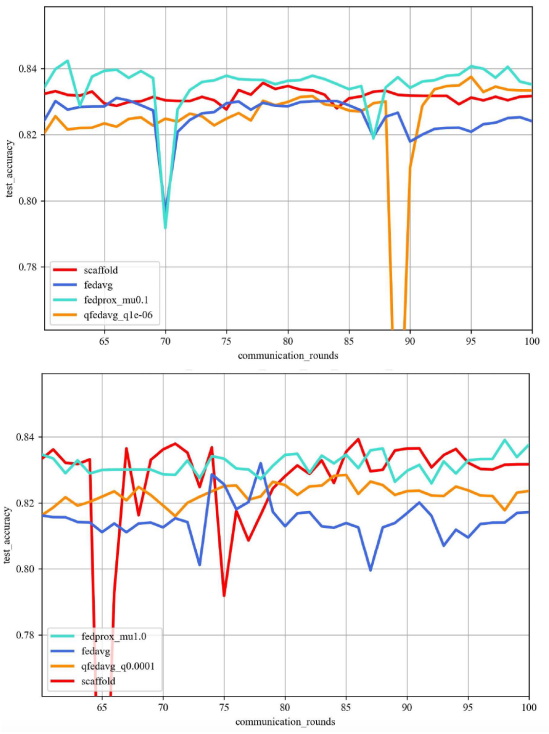

Fig. 4. Conf #4 & #5 - Testing accuracy for the last 40 communication rounds.

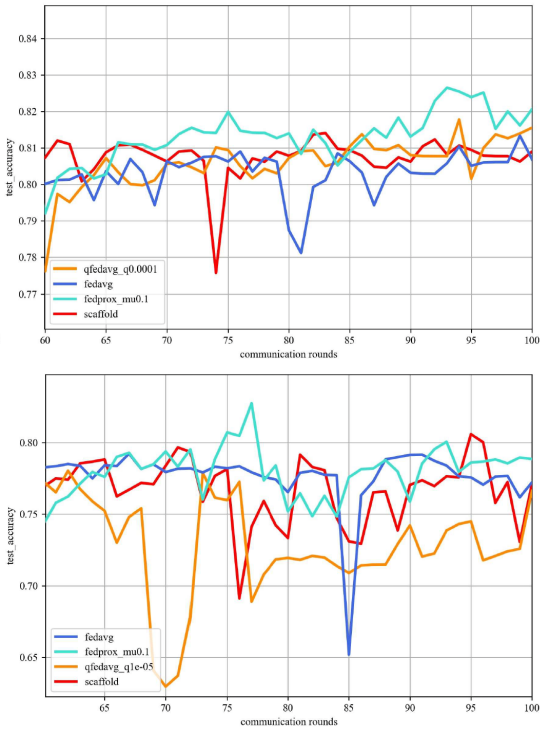

Fig. 5. Conf #6 & #7 - Testing accuracy for the last 40 communication rounds.

The qFedAvg's accuracy remains consistent for the first 40 consecutive communication rounds before stabilizing and reaching over 80% accuracy by round 48. For the more extreme Configuration #7, Fig.5 shows that qFedAvg and Scaffold are extremely unstable during the first 40 communication rounds and remain relatively unstable compared to FedAvg and FedProx in the last 50 rounds.

The final accuracies of all algorithms are logically lower than those of the previous configurations, as part of the datasets are degraded.

IV. Discussion

The experimental results demonstrate that the performance of federated aggregation algorithms varies significantly depending on the distribution of data among federated clients. For example, the FedAvg algorithm shows excellent performance in contexts where data is IID among clients, but its performance decreases significantly when data is non-IID, such as in cases of label distribution skew or feature distribution skew. Conversely, algorithms like FedProx and Scaffold are more robust in non-IID scenarios but underperform compared to FedAvg in IID contexts.

Configuration	FedAvg	qFedAvg	FedProx	Scaffold
Conf #1 IID Balanced	**83.96%**	82.27%	83.17%	81.41%
Conf #2 IID UnBalanced	**84.20%**	83.65%	83.82%	82.21%
Conf #3 IID UnBalanced +	84.13%	83.49%	**84.29%**	81.41%
Conf #4 Non-IID Label skew	82.37%	83.33%	**83.48%**	83.17%
Conf #5 Non-IID Label skew +	81.73%	82.37%	**83.81%**	83.17%
Conf #6 Non-IID Feature skew	80.61%	81.57%	**82.11%**	80.92%
Conf #7 Non-IID Feature skew +	77.34%	77.08%	**78.84%**	77.56%

TABLE III
ACCURACIES ACHIEVED AT COMMUNICATION ROUND 100 BY ALL AGGREGATION ALGORITHMS FOR EACH CONFIGURATION.

Table III summarizes the final accuracies achieved at communication round 100 by all aggregation algorithms for the seven artificially generated dataset configurations.

According to the results presented above, FedAvg provides the best performance in cases without statistical heterogeneity. It shows significantly higher test accuracies than the other algorithms for the first two IID configurations and nearly matches FedProx for the third IID configuration.

In non-IID scenarios, performance dynamics shift notably. FedAvg records the lowest accuracies in three of the four non-IID configurations and ranks second-worst in the final one. This degradation under non-IID conditions is expected and aligns with theoretical expectations.

FedProx consistently achieves the best performance in non-IID configurations, with more stable accuracies than qFedAvg and Scaffold at higher non-IID levels, as shown in Fig.4 and Fig.5. Although qFedAvg is slow to reach good accuracies, it performs well in most cases except the last, where high feature distribution skew leads to instability. Scaffold performs worst in the first three IID setups but improves notably in the final four non-IID configurations.

V. CONCLUSION

Aggregation algorithms play a crucial role in federated learning. Our study shows that their performance is closely tied to the data distribution among clients. FedAvg excels in IID scenarios but underperforms in non-IID settings with label or feature distribution skew. In contrast, FedProx and Scaffold perform better in non-IID environments but are less effective in IID contexts. This highlights the need for careful selection of aggregation algorithms based on data characteristics.

Practitioners should consider data distribution when implementing federated learning. For IID data, simpler algorithms like FedAvg may suffice, while non-IID scenarios require more robust options such as FedProx or Scaffold.

Future work could explore ensemble learning [12] and personalized federated learning (PFL) to address non-IID challenges. Ensemble learning, by combining multiple local models, could enhance robustness in heterogeneous settings. Similarly, PFL techniques like model clustering [13] or meta-learning [14] can better adapt models to specific client data, reducing the effects of statistical heterogeneity.

REFERENCES

[1] A. Yang, Z. Ma, C. Zhang, Y. Han, Z. Hu, W. Zhang, X. Huang, and Y. Wu, "Review on application progress of federated learning model and security hazard protection," *Digital Communications and Networks*, 2022.

[2] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji *et al.*, "Advances and open problems in federated learning," *arXiv preprint arXiv:1912.04977*, 2019.

[3] A. K. Sahu, T. Li, M. Sanjabi, M. Zaheer, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," *arXiv: Learning*, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:59316566

[4] T. Li, M. Sanjabi, A. Beirami, and V. Smith, "Fair resource allocation in federated learning," in *International Conference on Learning Representations*, 2020. [Online]. Available: https://openreview.net/forum?id=ByexElSYDr

[5] D. Kermany, M. Goldbaum, W. Cai, C. Valentim, H. Liang, S. Baxter, A. McKeown, G. Yang, X. Wu, F. Yan, J. Dong, M. Prasadha, J. Pei, M. Ting, J. Zhu, C. Li, S. Hewett, J. Dong, I. Ziyar, A. Shi, R. Zhang, L. Zheng, R. Hou, W. Shi, X. Fu, Y. Duan, V. Huu, C. Wen, E. Zhang, C. Zhang, . Li, X. Wang, M. Singer, X. Sun, J. Xu, A. Tafreshi, M. Lewis, H. Xia, and K. Zhang, "Identifying medical diagnoses and treatable diseases by image-based deep learning," *Cell*, vol. 172, pp. 1122–1131.e9, 2018.

[6] 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. [Online]. Available: https://proceedings.mlr.press/v54/mcmahan17a.html

[7] M. Sahu, Sanjabi, Zaheer, Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," *arXiv: Learning*, 2020.

[8] T. Li, M. Sanjabi, and V. Smith, "Fair resource allocation in federated learning," *ArXiv*, vol. abs/1905.10497, 2020.

[9] J. Wang, Q. Liu, H. Liang, G. Joshi, and H. V. Poor, "Tackling the objective inconsistency problem in heterogeneous federated optimization," in *Neural Information Processing Systems (NeurIPS)*, 2020.

[10] S. P. Karimireddy, S. Kale, M. Mohri, S. Reddi, S. Stich, and A. T. Suresh, "Scaffold: Stochastic controlled averaging for federated learning," in *Proceedings of the 37th International Conference on Machine Learning*, 2020.

[11] Z. Wang, X. Fan, J. Qi, C. Wen, C. Wang, and R. Yu, "Federated learning with fair averaging," 2021.

[12] X. Wu, J. Pei, X.-H. Han, Y.-W. Chen, J. Yao, Y. Liu, Q. Qian, and Y. Guo, "Fedel: Federated ensemble learning for non-iid data," *Expert Systems with Applications*, vol. 237, p. 121390, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417423018924

[13] J. H. Yoo, H. M. Son, H. Jeong, E.-H. Jang, A. Y. Kim, H. Y. Yu, H. J. Jeon, and T.-M. Chung, "Personalized federated learning with clustering: Non-iid heart rate variability data application," in *2021 International Conference on Information and Communication Technology Convergence (ICTC)*, 2021, pp. 1046–1051.

[14] A. Fallah, A. Mokhtari, and A. Ozdaglar, "Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 3557–3568. [Online]. Available: https://proceedings.neurips.cc/paper$_files/paper/2020/file/24389bfe4fe2eba8bf9$
$Paper.pdf$