On the Impact of Data Heterogeneity in Federated Learning: A Case Study on Air Quality Prediction

Zheyu Qiu

Computer Science and Technology, Nanjing University of Information Science & Technology, Nanjing, China

202383290020@nuist.edu.com

Abstract. Federated learning has become a popular model to apply in privacy-preserving modeling in distributed settings, particularly when the models are applied to data that is distributed among various locations and is often sensitive, such as in air quality prediction. This paper looks into how effective federated learning is for predicting ozone (O_3) concentrations, under both independent and non-independent data distributions. In particular, two representative algorithms Federated Averaging (FedAvg)and Federated Averaging (FedAvg) are experimented on a real-world air quality dataset. An experimental framework was established that was relatively comprehensive and centralized training and local-only models were used as baselines. Mean absolute error (MAE) and root mean squared error (RMSE) are used to measure model performance. The findings suggest that federated learning performs much better than the isolated local models and the performance is similar to that of the centralized training. Having said that, data heterogeneity does present certain issues-it slows down convergence and decreases accuracy in prediction. In such non-IID conditions, FedProx is more stable and less erroneous than FedAvg implying that it is more resistant to client drift.Most importantly, this paper provides empirical data on the impact of data heterogeneity on federated learning, and proves that federated learning can be a viable alternative in privacy-sensitive environmental prediction problems.

Keywords: Federated Learning, Federated Averaging, FedProx, Air Quality Prediction.

1. Introduction

The recent increased development of low-cost sensors and Internet of Things (IoT) technologies has greatly improved the fine-grained environmental monitoring and resulted in big amounts of high-resolution data that are critical to predicting air quality and reducing health risks [1]. Traditional machine learning methods often use centralized data gathering to come up with valid predictive models. Nevertheless, these centralized systems have severe constraints, such as privacy threats, excessive communication, and regulatory limitations, which limit the sharing of data [2].

Federated Learning (FL) has become a promising distributed learning paradigm to overcome these issues. In FL, several clients jointly train a global model via coordination of a central server without exchanging their raw data, thus maintaining privacy and minimizing communication expenses [3]. Although it has its benefits, the application of FL in practice, e.g. in air quality forecasting, is not easy because of statistical heterogeneity. The data produced by various clients are usually non-independent and non-identically distributed (Non-IID) [4].

The differences in sensor calibration, geographic position, and local sources of pollution may cause very different distributions of data among clients. This heterogeneity may worsen the work of the standard FL algorithms like Federated Averaging (FedAvg), leading to slow convergence, unstable training, and lower model accuracy [5]. To overcome this problem, Federated Optimization (FedProx) proposes a proximal term to the local objective function to limit local updates and enhance training stability in heterogeneous environments [6]. Besides, recent works have also delved into more sophisticated aggregation methods and hybrid techniques to further enhance performance in heterogeneous federated learning settings [7,8].

Even though the literature has addressed federated learning in Non-IID settings, much of the research has been on classification or synthetic datasets, with little consideration of real-world environmental regression problems like ozone prediction. Moreover, there has been no systematic

empirical study on the joint effects of important variables like data distribution, choice of algorithms, and local training effort on model performance in the heterogeneous federated learning environment. This weakness limits the application of FL in environmental monitoring systems.

This paper provides an in-depth empirical analysis of federated learning algorithms in ozone (O_3)prediction based on a real air quality dataset. Two of them are FedAvg and FedProx, which are considered representative. The analysis is designed in a series of controlled experiments. To measure the effect of heterogeneity, first, performance is compared in the case of IID and Non-IID data distributions. Second, the strength of FedAvg and FedProx when Non-IID settings are used is tested. Third, the impact of local computation is examined by changing the local training epochs. Lastly, hyperparameter optimization is carried out to determine the best proximal coefficient (μ) of FedProx.

A modular experimental framework is created in Python with NumPy to provide efficient computation to guarantee reproducibility and scalability. Experiments are configured using a centralized configuration file, and allow over 40 experimental setups to be run systematically. To achieve statistical reliability, each configuration is run with several random seeds. The results offer a good understanding of how data distribution, algorithm design, and local computation interact in federated learning, and offer useful recommendations on how to implement FL in real-world settings with heterogeneous environments.

2. Method

2.1. Dataset Preparation

The data applied in this research is found in the UCI Machine Learning Repository, which includes air quality data gathered during March 2004-April 2005. After preprocessing, a total of 8,762 valid samples are retained. The samples are comprised of several environmental characteristics, such as carbon monoxide (CO), temperature, and humidity, which are chosen as the main predictors of the ozone (O_3) concentration. The prediction is developed as a regression problem.

A number of preprocessing operations are carried out to guarantee the quality of data and model performance. To begin with, incomplete records are removed by dismissal of missing values. According to the authors, features are selected based on their relationship with the ozone concentration. Lastly, Z-score standardization technique is employed to normalize all the numerical features that determine the scaling of features for the model to converge faster.

To mimic a federated learning scenario, the data splits into multiple client groups. Data can be distributed either IID or Non-IID, In the IID setting, the data is randomly distributed among the clients that create similar local data distributions. In the opposite direction, Non-IID environment distributes data on the basis of the inherent feature attributes, making the client data distribution non-homogeneous which is similar to real world situations [6]. This set-up is in line with the usual federated learning setups where the data and statistical heterogeneity are decentralised, and it has a significant effect on model training and performance [5,6]. Federated learning systems, particularly IoT and edge intelligence systems, are faced with non-IID data distributions [1,2].

2.2. Federated Learning-Based Prediction

Federated Learning is a distributed learning model in which a number of clients form a global model without exchanging raw data which helps to keep data private while reducing communication costs [3]. The training takes place through a sequence of interactions. The global model is shared with all participating clients who locally train on their own datasets for each round repetition. The central server processes the newly updated model parameters from devices and aggregates them to update the global model.

This paper uses a Multi-Layer Perceptron (MLP) as the foundation prediction model. The network structure is specified as 4-64-32-1, which has an input layer with four features, two hidden layers with 64 and 32 neurons respectively, and an output layer to regression. Rectified Linear Unit(ReLU)

is the activation function applied to the hidden layers and Mean Squared Error (MSE) is the loss function.

Two federated learning algorithms are run: Federated Averaging (FedAvg) [5] and FedProx [6]. FedAvg is the baseline algorithm, which averages local model updates by weighted averaging. However, its performance may degrade under Non-IID conditions due to divergence in local updates. FedProx is an extension of FedAvg that adds a proximal regularization term to the local objective function, which limits local model updates and enhances stability in heterogeneous settings.

Besides this, recent research has suggested advanced aggregation and hybrid schemes to further improve the performance of federated learning in the presence of heterogeneity [7,8].

The training is set up to use the Adam optimizer, a learning rate of 0.002, 5 local training rounds per communication round and 30 global communication rounds in total. The proximal coefficient (μ) is adjusted to achieve a tradeoff between local adaptation and global consistency. Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) are used to measure model performance that complement each other as measures of prediction accuracy and strength.

3. Results and Discussion

3.1. Overall Performance Comparison

The experimental results clearly demonstrate that the prediction accuracy of models based on the federated learning algorithm is improved as compared to similar models built directly by using the local data. The local model has a 308.94 MAE that is quite high. In comparison, the federated approaches result in substantial error reductions: with FedAvg, the MAE drops to 187.03, and with FedProx, to 120.98. The centralized model, as expected, exhibits the best performance with a MAE of 108.25. Thus, this value can be reasonably considered as an upper-bound for the metric.

Federated learning can efficiently leverage a collection of datasets without compromising data privacy. Despite some remaining benefits of centralized training, the difference is not significant, suggesting that federated can achieve centralized performance with ulterior tuning [3,5]. Moreover, FedProx improves the learning of federated systems with heterogeneous data as show later [6,9].

3.2. Impact of Data Distribution (IID vs Non-IID)

In IID and Non-IID comparison, this paper finds out that data distribution plays an important role in the performance of the model. Both FedAvg and FedProx have lower values of MAE and RMSE under IID settings, showing stable and efficient training. However, as the data turns Non-IID, both methods degrade in performance, indicating a strong negative impact of Non-IID data on the convergence and prediction accuracy of federated learning.

Such an observation is consistent with the literature where the local model updates get separated owing to non-homogenous data, and thus the power of global aggregation is reduced [4,5]. Inevitably, there will be such heterogeneity in practical applications, e.g. air quality monitoring; there are variations in locations, types of sensors and environmental conditions. The data collected at each client node has different distributional properties, making the model consistency issue more challenging in federated learning systems. As a result, federated learning must be evaluated in non-IID conditions to test the applicability in real-world and to guarantee the reliability of the model implementation in practice [2,4].

The diversity in the distribution of client-side data is an essential aspect, as shown in Fig. 1 ,which clearly depicts the difference in data structure in the two experimental conditions.

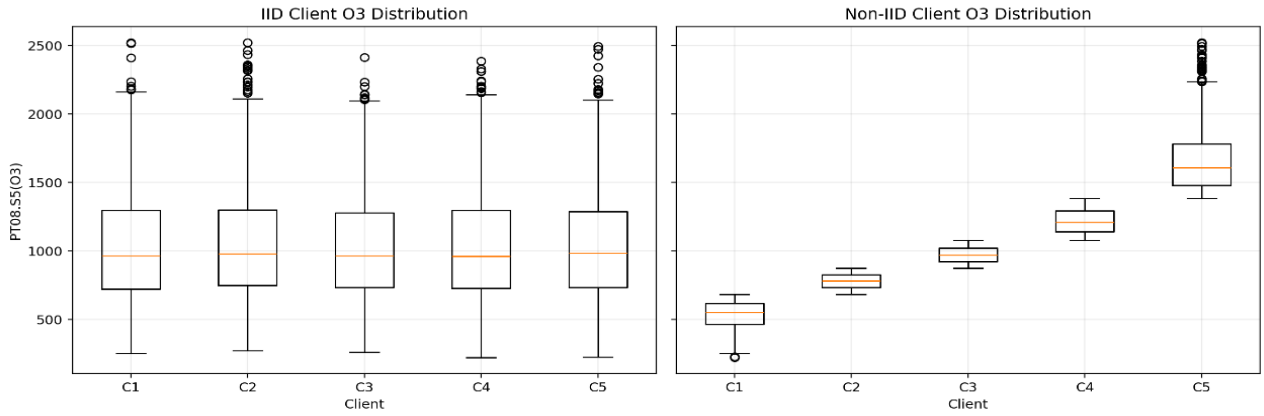

Fig. 1 Distribution of O_3 concentrations across clients under IID and Non-IID settings (Picture credit: Original).

3.3. Comparison between FedAvg and FedProx

In Non-IID settings, FedProx outperforms FedAvg on various metrics during evaluation. This new metric definition is especially useful for volatile time-series problems like stock-price index forecasting.

The proximal term in FedProx reduces the update of a local model far from the global model [6]. This explains the improvement. When mitigating client drift, the process is stabilized and helps in better generalization performance. Other studies also report consistent results, emphasizing the requirement of modified optimization strategies for federated learning with non-IID data [7, 8, 10].

As shown in Fig. 2. This paper assesses the concluding effectiveness of federate learning techniques under IID and Non-IID settings. The bar chart shows the MAE and RMSE values of FedAvg and FedProx.

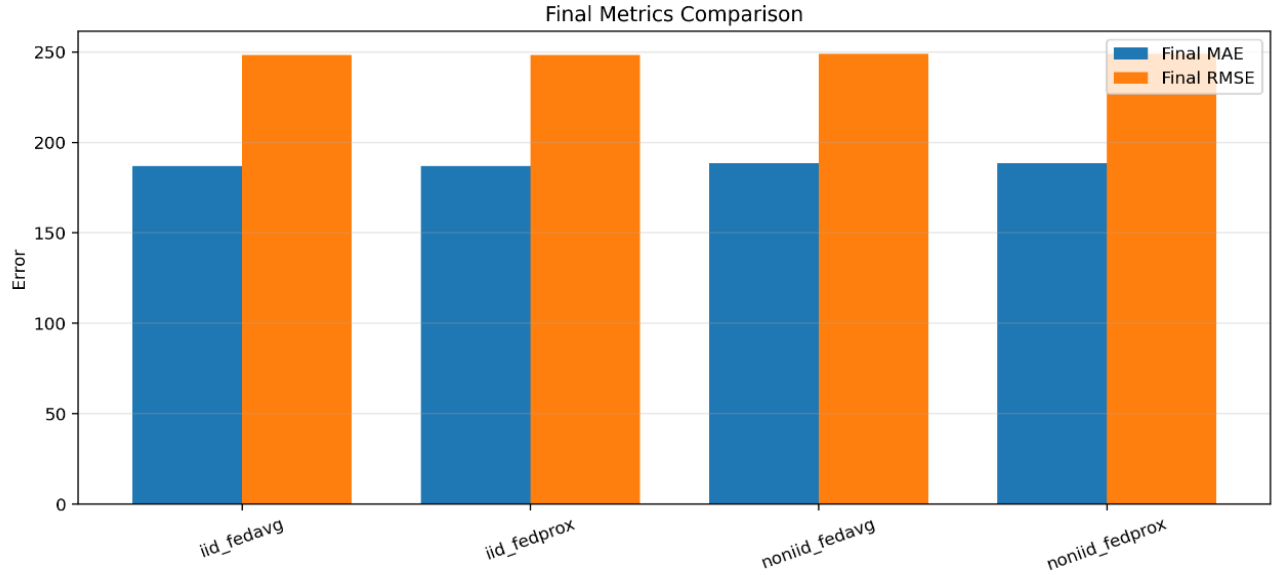

Fig. 2 Final Performance Comparison under IID and Non-IID Settings (Picture credit: Original)

The convergence curves in Fig.3 additional pieces of evidence show that FedProx further stabilizes the reduction in training loss throughout global communication rounds, especially under Non-IID settings, compared to FedAvg.

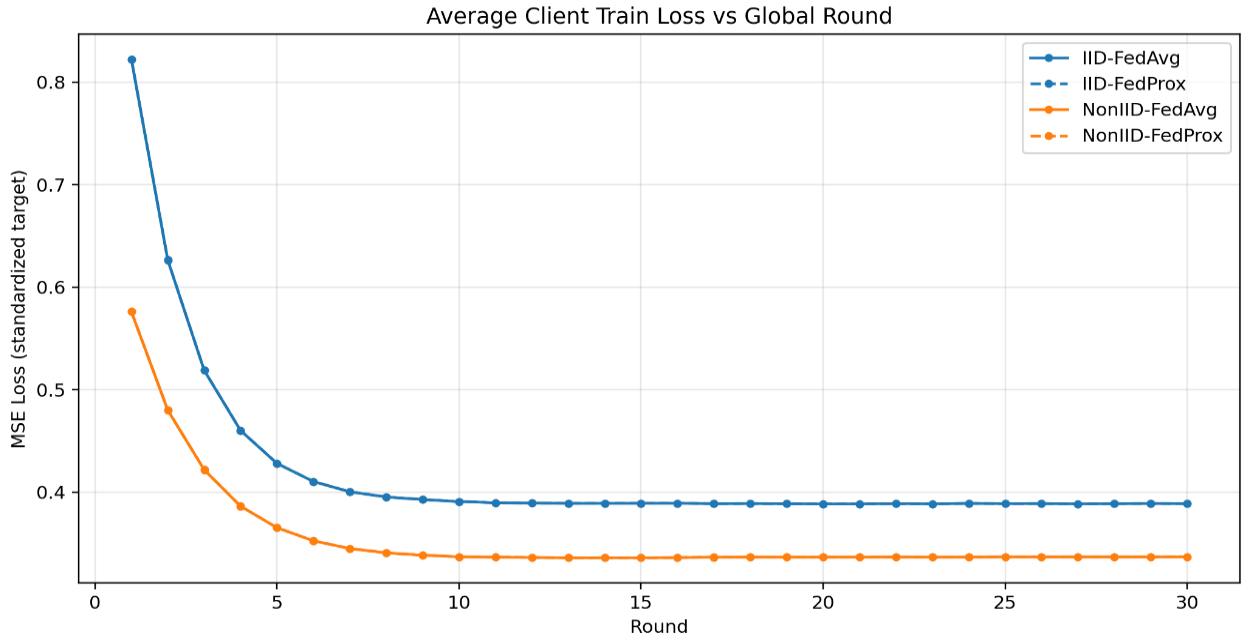

Fig. 3 Convergence of FedAvg and FedProx under IID and Non-IID Settings (Picture credit: Original)

3.4. Effect of Local Training Epochs

The amount of training epoch cycles at the local level (τ) acts as one of the key factors that influence the performance of the algorithm. The appropriate choice would be τ=5 as it provides a good balance between workload and communication. At the same time, increasing τ (for example, to τ=20) will deteriorate the performance especially in case of non-identically and independently distributed data [5,11]. This occurs since more updates made locally result in a greater divergence of the local models, making aggregation less effective.

3.5. Convergence Behavior Analysis

The convergence trends of MAE and RMSE show different behavior depending on the algorithms and distributions of the datasets used. When the datasets follow an IID distribution, it is apparent that the performance of both FedAvg and FedProx converges very quickly with relatively few communication iterations, alongside smooth optimization processes.

However, when the datasets exhibit Non-IID properties, convergence becomes much slower and less stable. The optimization process of FedAvg shows visible oscillations, whereas that of FedProx converges in a more stable manner. It implies that FedProx can better deal with the heterogeneity of the dataset during optimization.

These conclusions are consistent with previous studies [4,6,12], which have shown that training federated models on non-IID data poses certain difficulties. Therefore, FedProx would be a better alternative for practical application owing to its stability.

The convergence trend using MAE is presented in Fig. 4, which contrasts the optimization dynamics of MAE between global communication iterations of FedAvg and FedProx.

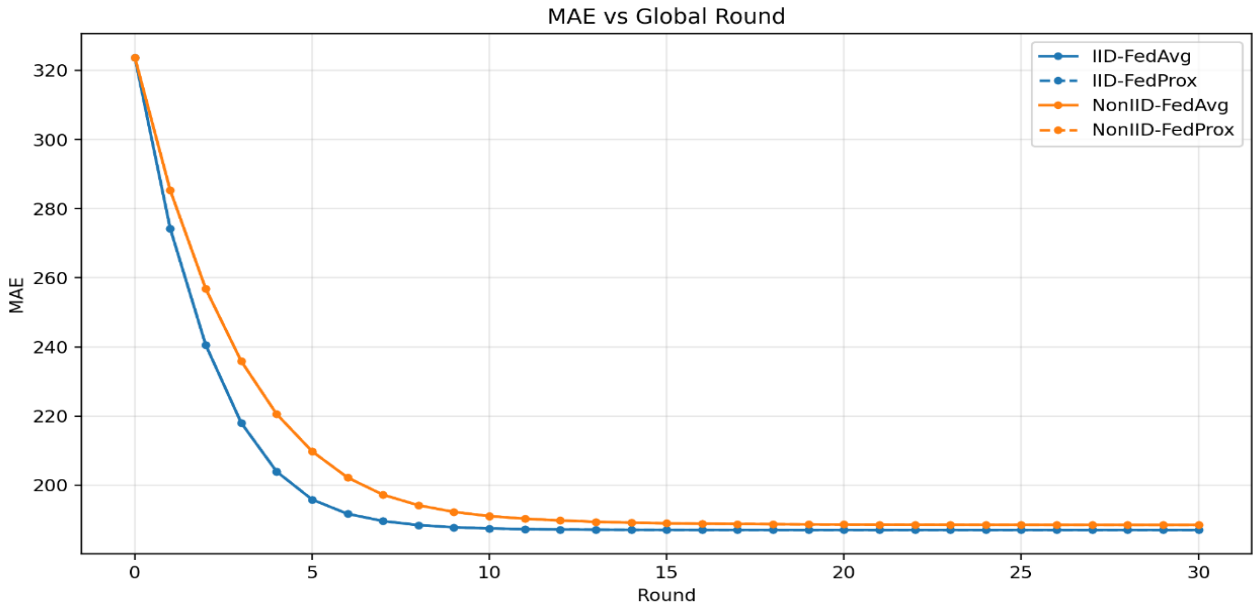

Fig. 4 Convergence of MAE for FedAvg and FedProx (IID vs Non-IID) (Picture credit: Original)

Similarly, Fig. 5 illustrates the evolution of RMSE over global communication rounds under IID and Non-IID settings, where the curves clearly show the differences in model stability and convergence speed between the two federated learning methods.

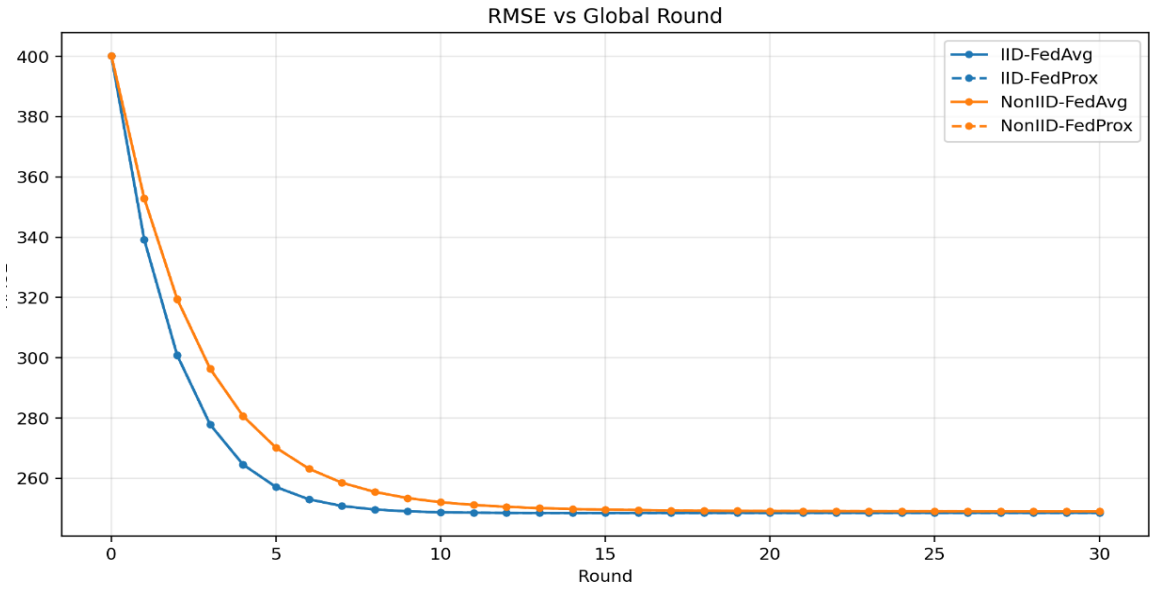

Fig. 5 RMSE Convergence and Stability under IID and Non-IID Settings3.6 Discussion and Insights (Picture credit: Original)

On the whole, there are a few important lessons that can be learned based on the findings. To begin with, federated learning evidently outperforms solitary local training by facilitating collaborative learning without exchanging raw data [1,5]. Second, the heterogeneity of data contributes greatly to the performance and should be handled with care [4,12]. Third, the adverse effects of Non-IID data can be mitigated effectively by enhancing the design of algorithms, e.g., FedProx [6,8].

Although these benefits exist, federated learning is yet to achieve full parity to centralized performance implying that it can be improved. This gap may be bridged in future studies by adaptive aggregation techniques, personalized federated learning, or hybrid systems [7,10,11].

Overall, this paper supports the claim that federated learning is an effective approach to privacy-preserving environmental prediction. Nonetheless, it is strongly dependent on the ability to control data heterogeneity and select appropriate optimization strategies.

4. Conclusion

This article has provided an empirical investigation of federated learning in predicting ozone (O_3) concentration in heterogeneous data distributions. Based on a real-world air quality dataset, this paper systematically compared the performance of FedAvg and FedProx in IID and Non-IID conditions, and further examined the local training intensity and baseline comparisons with centralized and local-only models.

The experimental findings show that federated learning significantly improves over isolated local training, indicating the usefulness of collaborative model optimization without sharing raw data. Although centralized training yields the highest predictive accuracy, the difference in performance between federated and centralized approaches is not very high, which means that federated learning can be used in practice in privacy-sensitive environmental monitoring tasks.

More importantly, the present study proves that statistical heterogeneity is a major factor that influences convergence behavior and ultimate model performance. FedProx is always more stable and with lower prediction error than FedAvg in the Non-IID setting, confirming the usefulness of proximal regularization in reducing client drift. The convergence analysis also indicates that to balance the efficiency of communication and the stability of optimization, it is crucial that the local training epochs are properly controlled.

Most importantly, this work offers empirical data that algorithmic adaptations are important in implementing federated learning in real-world environmental systems with heterogeneous data sources. Future studies can consider adaptive aggregation schemes, customized federated learning plans, and adaptive client engagement schemes to reduce the performance gap between centralized learning and improve robustness in large-scale applications.

References

[1] Li J, Wang S. A comprehensive survey on energy-efficient and privacy-preserving federated learning for edge intelligence and IoT. Results in Engineering. 2025;28:107849.

[2] Gad E, et al. A robust federated learning approach for combating attacks against IoT systems under non-IID challenges. In: Proceedings of the International Conference on Smart Applications, Communications and Networking (SmartNets); 2024 May; Harrisonburg, VA, USA. p. 1–6.

[3] Singh A, Gupta R. Privacy-preserving and computationally efficient federated learning for household load estimation. Sustainable Computing: Informatics and Systems. 2025;42:101234.

[4] Sun Z, et al. A survey on federated recommendation systems. IEEE Transactions on Neural Networks and Learning Systems. 2025;36(1):6–20.

[5] McMahan B, et al. Communication-efficient learning of deep networks from decentralized data. In: Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS); 2017. p. 1273–1282.

[6] Li S, Giannakis GB, Sun Y. Federated optimization in heterogeneous networks. In: Proceedings of Machine Learning and Systems (MLSys); 2020.

[7] Zeng D, et al. On the power of adaptive weighted aggregation in heterogeneous federated learning. In: Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS). 2025;258:1081–1089.

[8] Keerthika P, et al. Proximal guided hybrid federated learning approach with parameter-efficient adaptive intelligence for pneumonia diagnosis. Scientific Reports. 2025;16(1):2412.

[9] Karimireddy SP, et al. SCAFFOLD: Stochastic controlled averaging for federated learning. In: Proceedings of the International Conference on Machine Learning (ICML); 2020. p. 5132–5143.

[10] Li Q, He B, Song D. Model-contrastive federated learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2021. p. 10713–10722.

[11] Wang H, Yurochkin M, Sun Y, Papailiopoulos D, Khazaeni Y. Federated learning with matched averaging. In: Proceedings of the International Conference on Learning Representations (ICLR); 2020.

[12] Zhao Y, et al. Federated learning with non-IID data. arXiv preprint. 2018;arXiv:1806.00582.