Proceedings of the 4th International Conference on Mechatronics and Smart Systems
DOI: 10.54254/2755-2721/2026.34043

Impact of IID and Non-IID Data Distributions on Federated Learning for Air Quality Prediction

Kehao Bao

School of Computer Science, Sun Yat-sen University, Guangzhou, China
baokh3@mail2.sysu.edu.cn

Abstract. Air quality prediction plays an important role in environmental management and daily health protection. However, when air quality data are collected by different monitoring stations, centralized modeling may face privacy risks and practical difficulties in data sharing. To explore a more suitable distributed setting, this paper studies air quality prediction under federated learning and focuses on the effect of data distribution on model performance. The PRSA multi-station dataset is used, and PM2.5 is selected as the prediction target. After missing-value processing, feature construction, standardization, and sliding-window sample generation, a long short-term memory model is trained under two settings. In the IID setting, training samples are randomly shuffled and evenly assigned to clients. In the Non-IID setting, K-means clustering is applied and different clusters are distributed to different clients to simulate heterogeneous local data. The experimental results show that the IID setting achieves lower error, faster convergence, and better stability, while the Non-IID setting converges more slowly and shows larger fluctuations during training. These findings indicate that data heterogeneity has a clear influence on federated air quality prediction and should be carefully considered in future model design and practical deployment.

Keywords: Federated learning, air quality prediction, IID, Non-IID, data heterogeneity

1. Introduction

Air quality problems are closely related to the natural environment and the health of the human body [1]. With the development of industrialization and modernization, the influence of air pollution has gradually stood out in the public eye, which causes many problems in public health and social outdoor activity. Therefore, it is meaningful to make a precise prediction of the air quality. Air quality prediction can not only issue a warning of the coming air pollution and affect people's daily routine, but it can also become a basis for the environmental institution to take action to reduce the emission of gases [2]. Currently, how to use the historical data to predict the future air quality has become an important issue in intelligent computing and environmental science [3].

At present, machine learning and deep learning are widely used in air quality prediction [3, 4]. However, most research is based on modeling in a concentrated server, which collects all the data from the original station and models in the central server [5]. In this method, although it is convenient to model, the data are dispersed in the stations that are located in different areas where

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.34043

the temperature, humidity, weather, etc., are not the same; the effectiveness of modeling might be affected by the network, and in the process of exchanging the data, the privacy of the data might not be guaranteed. Under this background, Federated Learning (FL) as a method of distributed modeling is used in this scenario, which allows people not to share the original data and only model on the local computer [6]. Some recent studies have also applied federated learning to air quality prediction tasks [7, 8].

Although FL has the advantage in this distributed scenario, its performance will be influenced by the heterogeneity of data among different clients [9]. The data are divided into two categories, which are independent and identically distributed (IID) and non-independent and identically distributed (Non-IID) [9]. In IID, the samples are independent, and they would not influence each other; Non-IID is the opposite. Previous studies have shown that when the data of different clients is Non-IID, the performance of prediction and the speed of convergence will decrease [9, 10]. And it is a common situation that each station has heterogeneity in the source of the pollution, the condition of the weather, and the features of the environment [11]. Therefore, Non-IID is a realistic situation and a representative case. Meanwhile, although there are many studies that focus on the FL prediction, the research focusing on how IID and Non-IID will influence the performance of the air quality prediction in different heterogeneous data is still not sufficient [8].

Given the above, this paper takes air prediction as the task and compares the performance of the model in the conditions of IID and Non-IID based on FL. Among the two scenarios, the experiment will randomly divide the data into several clients in the IID scenario and divide the data into several clusters by the K-means method in the Non-IID scenario, and distribute different clusters to different clients [12]. Through the comparison of the performance and process of the two methods, this paper aims to analyse the influence of the feature of the data distribution in these two methods and give some reference for future work.

2. Methodology

2.1. Dataset

In this paper, the PRSA air quality dataset is used [13]. This dataset contains air pollution data and meteorological data collected from different monitoring stations in Beijing. The data used in this study cover the period from March 2013 to February 2017.

Each record represents the air condition of one station at one hour. The dataset includes several variables, such as PM2.5, PM10, SO_2, NO_2, CO, O_3, temperature, pressure, dew point, rainfall, wind direction and wind speed. In this paper, PM2.5 is selected as the prediction target.

2.2. Data preprocessing

Before training, the raw data needs to be processed. First, there are some missing values in the dataset. Linear interpolation is first applied to fill the missing values in numerical features. After that, forward filling is used, and finally backward filling is applied to handle the remaining missing parts. In this way, the missing values can be filled as much as possible while keeping the time order.

Then, some additional features are constructed. Since air quality data have periodic patterns, cyclic features are generated from hour and month. Specifically, hour is transformed into hour_sin and hour_cos, and month is transformed into month_sin and month_cos. This is because time features such as hour and month are periodic, and using sine and cosine functions can better represent this periodicity compared to using raw values directly.

DOI: 10.54254/2755-2721/2026.34043

Also, categorical features such as wind direction and station are transformed by one-hot encoding. After that, the data are grouped by station and sorted by time. Each station is divided into training, validation and test sets according to time order. Standardization is applied using StandardScaler, which is fitted on the training set and then applied to the validation and test sets.

Finally, a sliding window method is used to construct samples. For each station, the past 24 hours are used as input, and the PM2.5 value at the next hour is used as the prediction target. Then the window moves forward by one time step, which means the earliest hour is removed and a new hour is added. In this way, a sequence of samples can be generated continuously. After constructing samples for each station, all samples are combined together.

2.3. Data distribution settings

In this paper, two data distribution settings are considered, which are IID and Non-IID. For the IID setting, all training samples are mixed together and randomly shuffled. Then the samples are evenly divided into different clients, so that each client has similar data distribution.

For the Non-IID setting, K-means clustering is applied to the training samples. Each sample is first represented by statistical information such as mean and standard deviation, and then clustering is performed. After that, different clusters are assigned to different clients. In this way, the data distribution among clients becomes different, which is used to simulate the Non-IID situation.

2.4. Federated learning framework

In this paper, a basic federated learning framework is used.

Each client trains the model locally using its own data. After local training, the model parameters are sent to the central server. Then the server uses the Federated Averaging (FedAvg) method to aggregate the local models and obtain a global model. After that, the global model is sent back to each client for the next round of training.

2.5. Model and training settings

In this study, an LSTM model is used for air quality prediction. The model contains one LSTM layer with a hidden size of 64, followed by a fully connected layer.

The batch size is set to 128 and the learning rate is set to 0.001. The Adam optimizer is used.

In the federated learning setting, the number of clients is set to 4. Each client trains the model locally for 1 epoch in each round, and the total number of communication rounds is 5.

3. Results and discussion

3.1. Overall performance comparison

The performance of IID and Non-IID settings is compared based on the experimental results.

From the final results, it can be observed that the IID setting achieves better performance than the Non-IID setting. For example, the RMSE of IID is around 0.23–0.24 (about 0.236 after convergence), while the RMSE of Non-IID is higher, around 0.26–0.27 after convergence. A similar trend can also be observed in MAE.

DOI: 10.54254/2755-2721/2026.34043

Table 1. Final test results under IID and Non-IID settings

Setting	MAE	RMSE	R^2
IID	0.1268	0.2451	0.9573
Non-IID	0.1515	0.2892	0.9406

The curves in Figure 1 and Figure 2 are evaluated on the validation set during training, while the results in Table 1 are obtained on the test set after training.

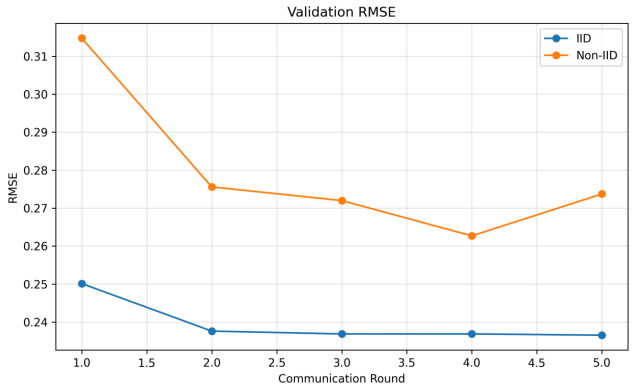

Figure 1. RMSE of IID and Non-IID settings over communication rounds (picture credit: original)

From the RMSE curve, it can be seen that both IID and Non-IID settings show a sharp decrease in the first round, which means the model can quickly learn useful information at the beginning.

However, after that, the behaviour becomes different. In the IID setting, the RMSE decreases once and then quickly becomes stable around 0.23–0.24. There is almost no obvious change in the following rounds, which shows that the model converges fast and remains stable.

In contrast, in the Non-IID setting, the RMSE continues to decrease slowly after the first round. For example, it decreases from around 0.31 to 0.28 and then to about 0.26. This shows that the model still needs more rounds to adjust. After several rounds, the RMSE increases again to around 0.27, which indicates that the training becomes unstable.

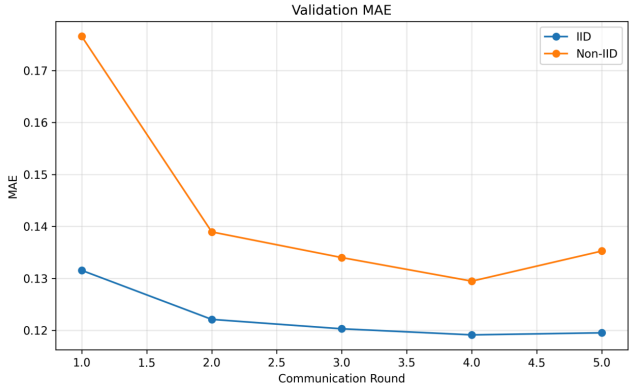

Figure 2. MAE of IID and Non-IID settings over communication rounds (picture credit: original)

From the MAE curve, both IID and Non-IID settings show a decreasing trend. The IID starts from around 0.13 and decreases to about 0.12, and then becomes stable. The Non-IID starts from a

DOI: 10.54254/2755-2721/2026.34043

higher value (around 0.18), decreases quickly to around 0.14, and then continues to decrease to around 0.13.

However, in the last round, both settings show a slight increase, but the increase in Non-IID is more obvious, while the change in IID is very small. This shows that Non-IID is more sensitive and less stable. Overall, IID shows faster convergence and better stability, while Non-IID shows slower convergence and larger fluctuation.

3.2. Training behaviour

From the training curves, the difference between IID and Non-IID settings can be clearly observed.

The IID setting converges quickly and remains stable after a few rounds. In contrast, the Non-IID setting shows a slower convergence process and larger fluctuations. It needs more rounds to approach a stable state, and even shows performance degradation in later rounds.

This indicates that the Non-IID setting is more difficult to train compared to the IID setting.

3.3. Discussion

The difference between IID and Non-IID settings can be explained by the data distribution.

In the IID setting, the data on different clients are similar. Therefore, the local models trained on different clients have similar update directions. When these updates are aggregated, the global model can converge easily and remain stable.

However, in the Non-IID setting, the data on different clients are different. Each client may learn different patterns, and the update directions are not consistent. When these updates are aggregated, it becomes difficult to find a model that fits all clients well. As a result, the training becomes slower and less stable.

In this study, the Non-IID setting is constructed using K-means clustering. This method groups similar samples together, which increases the similarity within each client but enlarges the difference between clients. As a result, the gap between clients becomes larger, which further increases the difficulty of model aggregation.

4. Conclusion

Air quality prediction is important for environmental management and daily life. Based on this background, this paper studied how different data distributions affect federated learning in an air quality prediction task.

A federated learning framework was built with a long short-term memory model. The PRSA dataset was processed through missing-value filling, feature construction, standardization, and sliding-window sample generation. Two settings were then compared: IID, where data were randomly shuffled and evenly divided, and Non-IID, where K-means clustering was used to assign different data clusters to different clients.

The results showed that the IID setting achieved better overall performance, with lower RMSE and MAE, faster convergence, and more stable training. In contrast, the Non-IID setting showed slower convergence and larger fluctuations, which suggests that heterogeneous client data make global aggregation more difficult.

This study still has some limitations, such as the simple model setting and the limited number of communication rounds. In future work, more federated learning methods, stronger prediction

DOI: 10.54254/2755-2721/2026.34043

models, and more realistic Non-IID scenarios can be tested to further improve performance and analysis.

References

[1] World Health Organization. (2004). Health aspects of air pollution: results from the WHO project" Systematic review of health aspects of air pollution in Europe".
[2] Bai, L., Wang, J., Ma, X., & Lu, H. (2018). Air pollution forecasts: An overview. International journal of environmental research and public health, 15(4), 780.
[3] Zhang, Z., Zhang, S., Chen, C., & Yuan, J. (2024). A systematic survey of air quality prediction based on deep learning. Alexandria Engineering Journal, 93, 128-141.
[4] Zhang, B., Rong, Y., Yong, R., Qin, D., Li, M., Zou, G., & Pan, J. (2022). Deep learning for air pollutant concentration prediction: A review. Atmospheric Environment, 290, 119347.
[5] Alwabli, A. (2024). Federated learning for privacy-preserving air quality forecasting using iot sensors. Engineering, Technology & Applied Science Research, 14(4), 16069-16076.
[6] McMahan, B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. In Artificial intelligence and statistics (pp. 1273-1282). Pmlr.
[7] Dey, S., & Pal, S. (2022). Federated learning-based air quality prediction for smart cities using BGRU model. In Proceedings of the 28th annual international conference on mobile computing and networking (pp. 871-873).
[8] Abimannan, S., El-Alfy, E. S. M., Hussain, S., Chang, Y. S., Shukla, S., Satheesh, D., & Breslin, J. G. (2023). Towards federated learning and multi-access edge computing for air quality monitoring: Literature review and assessment. Sustainability, 15(18), 13951.
[9] Li, T., Sahu, A. K., Zaheer, M., Sanjabi, M., Talwalkar, A., & Smith, V. (2020). Federated optimization in heterogeneous networks. Proceedings of Machine learning and systems, 2, 429-450.
[10] Zhao, Y., Li, M., Lai, L., Suda, N., Civin, D., & Chandra, V. (2018). Federated learning with non-iid data. arXiv preprint arXiv: 1806.00582.
[11] Agbehadji, I. E., & Obagbuwa, I. C. (2024). Systematic review of machine learning and deep learning techniques for spatiotemporal air quality prediction. Atmosphere, 15(11), 1352.
[12] Govender, P., & Sivakumar, V. (2020). Application of k-means and hierarchical clustering techniques for analysis of air pollution: A review (1980–2019). Atmospheric pollution research, 11(1), 40-56.
[13] Chen, S. (2019). BeijingMulti-Site Air Quality. UCI Machine Learning Repository. https: //archive.ics.uci.edu/dataset/501/beijing+multi+site+air+quality+data