RESEARCH ARTICLE

Air quality analysis model and prediction system based on federated learning and mobile swarm sensing

Yuechun Feng*

College of Computer Science and Engineering, Ningxia Institute of Science and Technology, Shizuishan, Ningxia, China.

Received: April 19, 2024; **accepted:** August 24, 2024.

Air quality is regarded as the core index to measure environmental health. Therefore, accurate monitoring and prediction of air quality are very important. Traditional monitoring methods cannot meet the needs of real-time and refined monitoring due to low coverage density and large sampling interval, especially in complex and changeable urban environments. How to use modern information technology to improve monitoring efficiency and accuracy, while protecting user privacy, has become the focus of current research. This study aimed to develop an air quality analysis and prediction system combining mobile swarm intelligence and federated learning technology to improve the coverage, accuracy, and prediction ability of air quality monitoring while ensuring the privacy of users. A federated learning-based air quality prediction model was proposed, which focused on the trade-off relationship between location privacy protection and model performance and detailed the application in single-pollutant prediction and joint multi-pollutant prediction. The results showed that, in the single-pollutant prediction, appropriately lowering the learning rate could improve the prediction accuracy, while reinforcing the privacy protection would bring a decrease in the prediction performance. When comparing federated learning with centralized learning, although federated learning had an advantage in privacy protection, the prediction performance was slightly lower than that of centralized learning. In addition, the temporal and spatial distribution characteristics of pollutant concentration data had a significant impact on model performance, suggesting that the model should adapt to spatial and temporal variations under different environmental conditions. In the joint multi-pollutant prediction, the distributed gradient updating mechanism fused with differential privacy strategy was applied to construct the prediction model. As the privacy budget tightened, the noise intensity increased and the prediction performance subsequently decreased, which visually demonstrated the trade-off between privacy protection and prediction performance. This study laid a foundation for real-time and refined management of environmental monitoring and guaranteed personal privacy, which not only promoted the progress of air quality monitoring technology, solved the limitations of traditional monitoring systems, but also provided advanced tools for environmental protection and public health management. The proposed privacy protection framework also provided a solution for other sensitive data related fields with a wide range of application potential and social impact.

Keywords: federated learning; mobile swarm sensing; air quality analysis; air quality prediction.

***Corresponding author:** Yuechun Feng, College of Computer Science and Engineering, Ningxia Institute of Science and Technology, Shizuishan 753000, Ningxia, China. Email: yuechun_feng@hotmail.com.

Introduction

With global environmental problems becoming increasingly prominent, accurate monitoring and

prediction of air quality, the key indicator of environmental health, has become an important issue [1]. Traditional air quality monitoring relies on the establishment of fixed monitoring stations. However, such methods suffer from low coverage density and large data sampling intervals, making it difficult to meet the needs of modern society for refined and real-time environmental information. In rapidly developing urban environments, air quality is especially characterized by high spatial and temporal variability, which requires more efficient and refined monitoring methods [2].

In recent years, the rapid development of mobile internet and Internet of Things (IoT) technology makes the large-scale application of mobile device sensors possible, which, in turn, has given rise to the emerging concept of "Mobile Cluster Sensing". Mobile swarm sensing utilizes smartphones and smart wearable devices widely distributed in the hands of the publics to sense and upload real-time ambient environmental parameters such as PM2.5, PM10, SO_2, NO_2, and other air quality indicators, thus building a wide-coverage, fine-grained air quality monitoring network. However, the user location information and environmental data involved in this process are extremely sensitive, and once leaked, it will seriously threaten the privacy and security of users [3]. With the convergence of artificial intelligence (AI) and big data technologies, federated learning emerges as a technical framework that enables multi-party collaborative learning while protecting data privacy. Federated learning allows different mobile devices computing model parameters locally and aggregates model updates only in the cloud without directly transmitting the raw data, which provides a strong support for solving the data privacy protection problem in the context of mobile swarm intelligence sensing [4]. The current air quality monitoring system faces serious challenges, which include inadequate coverage and detail of the monitoring network. Although a certain number of fixed air quality monitoring stations have been deployed around the world, these stations are usually unable to cover all geographical areas evenly, especially in the complex spatial structure within cities, as well as in rural areas and remote zones [5, 6]. This limitation prevents comprehensive access to air quality data, thus restricting the ability of environmental management authorities to implement refined management of air quality conditions within localized microenvironments and specific spatial regions. Further, traditional monitoring methods are often difficult to accurately capture transient fluctuations in air quality within a short period of time as well as spatial variability on small scales due to sparse distribution points and limited sampling frequency. Furthermore, the bottleneck in air quality prediction is the lack of data integration and in-depth analysis. Air quality is affected by various complex factors such as meteorological conditions, geographic factors, and human activities. However, the existing data processing technologies and algorithms are not yet able to make full use of the huge amount of data from various sources to build high-precision, long-lasting prediction models. Therefore, how to effectively explore the correlation and regularity between data to improve the predictability and accuracy of air quality prediction is an important issue that needs to be solved urgently [7, 8].

With the popularization of mobile smart terminals and technological advances, mobile group sensing, relying on the huge mobile device user group, builds an unprecedentedly large and vivid environmental information sensing network. Each cell phone, smart wearable device, and even in-vehicle sensors have become small sentinels for real-time monitoring of environmental quality, continuously contributing diversified data covering all kinds of environmental parameters including, but not limited to, the air quality index, temperature, humidity, wind speed, and so on. This bottom-up, large-scale data collection has greatly enriched the data sources for air quality monitoring, allowing us to paint a vivid picture of air quality in cities and around the world with unprecedented detail and immediacy [9]. However, users' location information, life

trajectories, and other related environmental data are highly sensitive. How to ensure the integrity of the data required for air quality management, while strictly adhering to the principle of data minimization and effectively preventing the leakage of users' privacy has become a double test of technological development and legal regulation [10].

Early models such as diffusion models and atmospheric chemical transport models can simulate the behavior of atmospheric pollutants, but they are limited by high data dependence and computational complexity. In contrast, data-driven methods such as deep learning, ensemble learning, time series analysis, support vector machines, and random forests significantly improve prediction accuracy. Mobile Crowd Sensing technology shows great potential in environmental monitoring, building a more comprehensive and immediate air quality monitoring network through real-time environmental data collected by smartphones and other mobile devices [11]. However, issues such as data quality, user engagement, and location privacy remain unresolved. Federated learning is a new step forward in privacy protection with the ability to train models collaboratively without revealing raw data. It has been applied in many fields such as medical care and finance and has begun to dabble in air quality prediction. However, there are still challenges in how to efficiently and securely implement federated learning in environmental monitoring, especially in protecting the location privacy of mobile device users [12]. Therefore, how to effectively integrate a large amount of data generated by mobile group intelligence perception and remove interference while ensuring prediction accuracy, as well as how to apply federated learning among large-scale heterogeneous mobile devices to design efficient communication and collaborative optimization strategies while ensuring location privacy are the current research trends [13, 14]. Federated learning technology with its unique distributed learning concept brings new ideas to solve the above data security and privacy protection challenges. The core concept of federated learning is to let the data stay local. Each participant (i.e., mobile device) trains the model locally and independently, and exchanges only the model parameter updates that are encrypted and differentially privacy-protected instead of sharing the original data directly, which ensures that large-scale data-driven air quality models can be fully learned and optimized, while minimizing the risk of data leakage and ensuring that users' privacy are effectively protected.

This research aimed to find a balance between air quality analysis and prediction and location privacy protection by developing an air quality analysis model and prediction system that integrated federated learning and mobile swarm sensing techniques with a particular focus on solving the following key problems with location privacy-preserving characteristics including respecting and protecting the location privacy of the air quality data providers, adopting an innovative approach to realize the precise localization of pollution sources, thus improving the effectiveness of environmental regulation while ensuring that individual privacy were not violated. A spatial interpolation algorithm for pollutant concentration that considered location privacy protection was designed and implemented to utilize large-scale, real-time, multidimensional air quality data generated by mobile devices to characterize the spatial distribution of pollutants in high resolution, thus accurately reflecting the changes of air quality in different geographic regions.

Materials and methods

Federated learning theory framework and location privacy protection mechanisms

Federated learning is a distributed machine learning paradigm that aims to solve the data silo problem while safeguarding data privacy (Figure 1) [15]. In the air quality analysis and prediction system, a set of ensembles containing multiple mobile devices $D = \{D_1, D_2, ..., D_N\}$ was defined

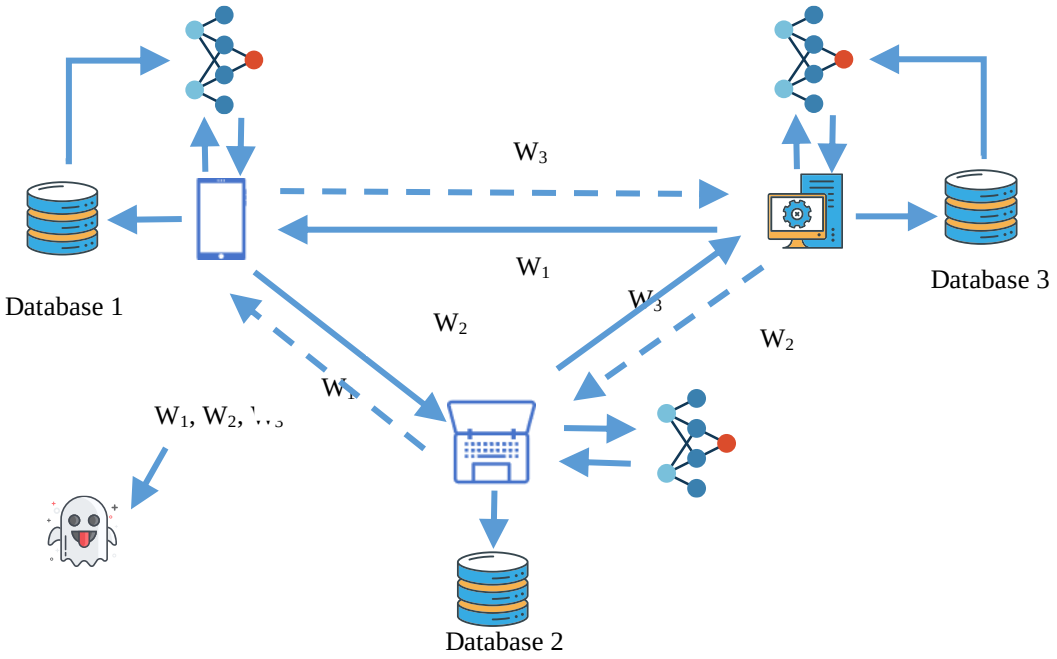

Figure 1. Federal learning framework.

first. Each device had a local dataset D_k and updated the model parameters θ_k through local training. The local training process could be represented as follows.

$$\theta_k^{t+1} = \theta_k^t - \eta \nabla L(\theta_k^t; D_k)$$

where η was the learning rate. $L(\cdot)$ was the loss function. ∇L was the gradient of the loss function with respect to the model parameter θ_k [16]. To protect the location privacy of participating devices, privacy-preserving strategies were implemented during data transmission. For example, a privacy-preserving gradient could be realized by adding noise to the local gradient $\nabla L(\theta_k^{t+1}; D_k)$ through a differential privacy mechanism below.

$$g_k'^{t+1} = \nabla L(\theta_k^{t+1}; D_k) + N(0, \sigma^2 I)$$

where $N(0, \sigma^2 I)$ was a Gaussian noise with zero mean and covariance matrix of $\sigma^2 I$. The central server then performed a global model update by aggregating the encrypted or noisified gradients (or model parameters) of all devices. For model parameter aggregation, the FedAvg (Federated Averaging) algorithm could be used as follows.

$$\theta^{t+1} = \sum_{k=1}^{N} \frac{n_k}{n} \theta_k^{t+1}$$

or in the case of noisified gradients as below.

$$\theta^{t+1} = \theta^t - \eta \sum_{k=1}^{N} \frac{n_k}{n} g_k'^{t+1}$$

where n_k was the number of data samples from device k. $n = \sum_{k=1}^{N} n_k$ was the total number of samples from all devices. In addition to adding noise for differential privacy, homomorphic encryption could be used to further enhance location privacy protection. In the Paillier homomorphic encryption system, a device D_k could encrypt its model updates and send them securely to the server as follows [17].

$$Enc(\theta_k^{t+1}) = Enc(\theta_k^t) \times Enc(1-\eta \frac{\nabla L(\theta_k^t; \mathrm{D}_k)}{\| \nabla L(\theta_k^t; \mathrm{D}_k)\|}) \mod N^2$$

The application of federated learning theoretical framework in air quality analysis and prediction system combined various privacy protection techniques such as local training, differential privacy, homomorphic encryption, *etc.*, aiming to effectively protect the location privacy information of the participating devices without affecting the model performance.

Mobile group intelligence perception data acquisition and preprocessing

Mobile swarm sensing technology utilized a network of sensors distributed across mobile devices to collect air quality related multivariate data in real time. Let the pollutant concentration data collected by device i at a certain moment be x_i and its coordinate position be p_i, the air quality data set collected by the whole mobile swarm sensing network could be expressed as below.

$$\mathrm{X} = \{(x_1, p_1), (x_2, p_2), ..., (x_n, p_n)\}$$

where n was the number of mobile devices participating in data collection. In the data preprocessing, data cleaning was first performed to remove invalid or erroneous data records, and outliers could be excluded by setting a threshold value as follows.

$$\mathrm{X}_{clean} = \{(x_i, p_i) \mid x_i \in [\alpha, \beta], \forall i \in [1, n]\}$$

For subsequent spatial interpolation and predictive analysis, the raw data needed to be transformed to a unified spatial coordinate system and geographic coordinate transformation. If device i used latitude and longitude coordinates, it could be converted to a planar rectangular coordinate system by method of Mercator projection shown below.

$$p_{i'} = f(p_i)$$

where $f(\cdot)$ was the conversion function from geographic to planar coordinates [18]. To further protect location privacy, device coordinates could be fuzzified by introducing noise through a differential privacy mechanism as:

$$p_{i''} = p_{i'} + \grave{o} \cdot \eta .$$

where $\grave{o}$ was the privacy budget parameter. η was the noise that conformed to some probability distribution (e.g., Gaussian or Laplace distribution). In addition, considering the possible differences in measurement accuracy between different devices, data normalization was needed to reduce the bias between the data, usually using min-max normalization:

$$x_i^{norm} = \frac{x_i - \min(\mathrm{X}_{clean})}{\max(\mathrm{X}_{clean}) - \min(\mathrm{X}_{clean})}$$

or z-score normalization:

$$x_i^{norm} = \frac{x_i - \mu}{\sigma}$$

where μ was the mean value of the pollutant concentration in the data set X_{clean}. σ was its standard deviation [19]. These preprocessed data were further used for subsequent spatial interpolation and prediction model construction.

Pollutant space interpolation algorithm for location privacy protection

Privacy-Preserving Adaptive Spatial Interpolation (PPASI) was introduced in this study. A discretized observational dataset of pollutant concentrations was applied as follows.

$$\mathrm{D} = \{(p_i, z_i)\}_{i=1}^n$$

where p_i was the device location information that had been obfuscated. z_i was the pollutant concentration measured at location p_i. To achieve privacy protection, a technique similar to

differential privacy was used to add noise to the original location data p_i as:

$$p_{i'} = p_i + \eta_i$$

where η_i was a random noise that conformed to some probability distribution to ensure that location privacy was not easily resolved. To perform spatial interpolation, an adaptive weight function $w(p_{i'}, p_{j'})$ was first constructed, which automatically adjusted the weights based on the relative position and concentration differences between the location $p_{j'}$ and the point p.

$$w(p_{i'}, p_{j'}) = \exp(-\alpha \cdot \| p_{i'} - p_{j'} \|^2) \cdot \exp(-\beta \cdot | z_i - z_j |)$$

where α and β were positive parameters controlling the influence of spatial distance and concentration difference. $\| p_{i'} - p_{j'} \|$ was the Euclidean distance between two points [20]. The spatial interpolation under location privacy protection could then be realized by the following equation.

$$Z^*(p) = \sum_{i=1}^{n} w(p_{i'}, p) \cdot z_i$$

For any point p to be interpolated, its pollutant concentration estimate $Z^*(p)$ was obtained by weighted summation of the observations z_i at the surrounding fuzzified locations, while the weighting function $w(p_{i'}, p)$ took into account both spatial proximity and concentration similarity and ensured location privacy protection. To optimize the performance of the algorithm, the optimal parameters α and β , as well as the noise intensity, were determined by means of cross-validation to strike a balance between ensuring the interpolation accuracy and location privacy protection.

Federalization of single-pollutant data

Under the framework of federated learning-based air quality prediction, the single-pollutant concentration dataset collected by a single mobile device *k* was denoted as:

$$\mathrm{D}_k = \{(x_i^k, y_i^k)\}_{i=1}^{m_k}$$

where x_i^k was location-independent environmental features (e.g., time of day, meteorological parameters, *etc*.). y_i^k was the corresponding pollutant concentration value. m_k was the number of data samples for device k. The data were collected by a single mobile device, which was a single-pollutant model. Each device used its own data to train a local model independently. Usually, a neural network model was chosen, and the model parameters were denoted as θ_k . The training process of the local model could be expressed by the following gradient descent method.

$$\theta_k^{t+1} = \theta_k^t - \eta \nabla_{\theta_k} L_k(\theta_k^t; \mathrm{D}_k)$$

where t was the current communication round. η was the learning rate. L_k was the loss function of device k. $\nabla_{\theta_k} L_k$ was the gradient of the loss function with respect to the model parameter θ_k [21]. To realize federated learning, all devices interacted through multiple communications to aggregate the model parameter updates obtained from their respective local training to the central node, which updated the global model parameters θ through aggregation as follows.

$$\theta^{t+1} = \sum_{k=1}^{N} w_k \theta_k^{t+1}$$

where N was the total number of devices participating in federated learning. w_k was the weight factor of the data volume share of device k, which ensured that the global model could

reflect the data information of all devices in a balanced way.

Privacy-preserving strategies and performance evaluation of predictive models

The model performance should not be significantly degraded while protecting the user's location privacy. Differential privacy techniques were employed to perturb the local gradient to prevent the actual location information of the user from being inferred through the gradient inverse. Specifically, before uploading the gradient, device k added noise to its gradient that satisfied differential privacy as below.

$$g_k^{DP} = \nabla_{\theta_k} L_k(\theta_k^t; \mathrm{D}_k) + \eta_p$$

where $\eta_p \sim \mathrm{N}(0, \sigma^2 I)$ was a random noise drawn from a Gaussian distribution with zero mean and variance σ^2. σ was associated with a set privacy budget ò, which was selected and parameterized by either the Laplace mechanism or the Gaussian mechanism to ensure that it satisfied the preset ò value [22].

Model performance evaluation

Model performance evaluation involved two dimensions including prediction accuracy and degree of privacy protection. Prediction accuracy was often measured using metrics such as Mean Square Error (MSE) and Mean Absolute Error (MAE) as follows [23].

$$MSE = \frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2$$

$$MAE = \frac{1}{m}\sum_{i=1}^{m}|y_i - \hat{y}_i|$$

The degree of privacy protection was then measured by the privacy budget of differential privacy ò . Additional experiments were performed to determine whether the prediction performance of the model was affected after adding a privacy protection strategy. A series of experiments were conducted by changing the intensity of the noise (i.e., adjusting the values of σ or ò) to explore the relationship between the privacy protection intensity and the model performance. Through the observation of the changes in the prediction performance, the optimal balance between privacy protection and prediction performance was then determined. Four experiments were performed to explore the effects of learning rate, privacy budget, learning method, and data distribution on model performance.

Co-processing and correlation analysis of multi-pollutant data

Assuming that m pollutants were involved, the dataset could be denoted as

$$\mathrm{D} = \{(X_t, Z_t)\}_{t=1}^{T}$$

where X_t was the d dimensional environmental feature matrix at time step t . Z_t was the corresponding m concentration vector of the pollutants. To reveal the intrinsic correlation among pollutants, the covariance matrix Σ_Z and the Pearson correlation coefficient matrix R_Z were calculated to understand the pattern of change of pollutant concentrations and the degree of interaction.

Privacy-preserving framework design for joint multi-pollutant concentration prediction

In constructing a joint multi-pollutant prediction model under the federated learning framework, a distributed gradient updating mechanism was adopted and incorporated a differential privacy preserving strategy. Assuming there were N devices and each device k had a local model parameter of Θ_k , the goal was to co-optimize the global model parameter Θ . Device k computed the gradient locally and incorporated differential privacy noise as follows.

$$\tilde{G}_k = \nabla L_k(\Theta_k; \mathrm{D}_k) + \eta_k$$

Table 1. Co-processing and correlation analysis of multi-pollutant data.

Type of pollutant	Sample size	Average concentration	Maximum concentration	Minimum concentration
PM2.5	10,000	50 μg/m³	200 μg/m³	10 μg/m³
NO_2	10,000	40 ppb	100 ppb	20 ppb
SO_2	10,000	20 ppb	60 ppb	10 ppb
O_3	10,000	50 ppb	150 ppb	20 ppb

η_k was sampling from a Laplace or Gaussian distribution that satisfied the differential privacy constraints. The server side applied a secure aggregation protocol to summarize the encrypted gradients as below.

$$\bar{G} = \sum_{k=1}^{N} Decrypt(\bar{G}_k)$$

The global model parameters were then updated based on the aggregated gradients Θ.

Results and discussion

Learning rate

The learning rate was determined by fixing three different learning rates of 0.01, 0.001, 0.0001 and compared their performance with the same privacy budget. A learning rate of 0.001 resulted in the optimal prediction performance of the model as evidenced by the lowest mean square error (MSE) and mean absolute error (MAE), which suggested that a moderately low learning rate helped the model to converge better during the training process, thus improving the prediction accuracy [24]. The results showed that MSE and MAE of the model decreased and then increased with decreasing learning rate when learning rates were 0.01, 0.001, and 0.0001, respectively, indicating that there was a better learning rate value that could produce the best prediction performance. As the privacy budget decreased and noise intensity increased, MSE and MAE increased accordingly, indicating that model prediction performance declined to some extent as privacy protection levels increased.

Comparison of federated learning-based models and centralized learning models

Under the same dataset and privacy protection level, federated learning model demonstrated superiority in privacy protection. However, the prediction performance was slightly lower than that of centralized learning model. The results reaffirmed the trade-off between privacy protection and model performance and suggested the consideration of the trade-off between privacy protection and model prediction performance in practical applications [25].

The influence of spatial and temporal distribution characteristics of pollutant concentration data on the prediction performance

The effect of the distribution of pollutant concentration data over time and geographic region on model performance was examined. The data used in this study covered peak period and average peak period in two different geographical regions. The results showed that the prediction performance of the model under different scenarios was significantly different, which might be due to the different spatial and temporal characteristics of pollutant concentrations. The results suggested that the spatiotemporal characteristics of practical application scenarios should be fully considered when constructing and optimizing the model to improve the prediction accuracy of the model

Table 2. Comparison of federal learning parameter settings and noise intensity.

	Learning rate	Privacy budget (ò)	Noise intensity (σ)	Predicted percentage of performance degradation
Case 1	0.01	1.0	0.1	5%
Case2	0.01	0.5	0.2	10%
Case3	0.01	0.1	0.5	15%

under various environmental conditions.

Co-processing and correlation analysis of multi-pollutant data

A total of 10,000 samples for all contaminants was involved in this study, which was adequate to represent all pollutants and provided basic data support for studying their distribution characteristics, interactions, and the establishment of prediction models (Table 1). There was a strong positive correlation between PM2.5 and NO_2 with the correlation coefficient as 0.75, which indicated that the concentrations of the two pollutants tended to rise and fall simultaneously. In contrast, PM2.5 and O_3 showed a negative correlation with the correlation coefficient as -0.40, indicating that the concentration trends of these two pollutants were opposite.

Privacy-preserving framework design for joint multi-pollutant concentration prediction

Three different cases that involved a learning rate of 0.01 were tested. As the privacy budget decreased from 1.0 to 0.1, the corresponding noise intensity gradually increased from 0.1 to 0.5 to satisfy stronger privacy protection. The increase in the noise intensity led to a decrease in the prediction performance of the model, which was manifested by the increase in the percentage of the decrease in the prediction performance from 5% (Case 1) to 15% (Case3) gradually (Table 2).

Joint prediction model performance evaluation and case study

To evaluate the performance of federated learning in joint multi-pollutant prediction, the following performance metrics including root mean square error (RMSE), MAE, and coefficient of determination (R^2 score) were employed. The results showed that the federated learning-based joint multi-pollutant prediction model maintained high prediction accuracy while ensuring the privacy of users' locations in real case studies. The federated learning algorithm performed well in the multi-pollutant prediction of a monitoring station In Beijing China, especially in the prediction of O_3, which proved the effectiveness of the algorithm in practical application. The results showed that FedAvg-DP was superior to its variant FedProx in maintaining prediction accuracy and model adaptability. Although FedAvg-DP with differential privacy made a compromise in protecting user privacy and resulted in a slight decline in prediction performance, it still demonstrated the potential and flexibility of federated learning frameworks for multi-pollutant joint prediction tasks (Table 3).

Table 3. Performance assessment of joint federal learning multi-pollutant prediction (case study of a monitoring station in Beijing, China).

	PM2.5	NO_2	SO_2	O_3
RMSE	15.0	10.2	8.5	7.3
MAE	10.5	7.8	6.5	5.8
R^2 score	0.82	0.87	0.90	0.93

Conclusion

An in-depth study on air quality prediction models based on federated learning and their location privacy protection was conducted, and a set of comprehensive solutions combining distributed training, differential privacy protection, and spatial interpolation algorithms was constructed in this study. The results showed that, in single-pollutant prediction models, more accurate prediction results could be achieved while protecting user location privacy by reasonably choosing the learning rate, regulating the privacy budget, and adopting an efficient noise injection strategy. Although enhanced privacy protection might lead to a slight decline in prediction performance in some cases, the results demonstrated that an effective balance between privacy protection and prediction performance could be reached through careful parameter tuning and model design. In multi-pollutant joint prediction models, complex correlations among pollutants were identified, and a privacy-preserving federated learning framework applicable to multi-pollutant joint prediction was innovatively proposed. By comparing different federated learning algorithms such as FedAvg, FedProx, and FedAvg-DP, it was confirmed that the federated learning-based multi-pollutant prediction model could demonstrate high prediction accuracy even under the strict protection of user location privacy. There are still some shortcomings in this research. The current model had not yet fully explored the effects of complex environmental factors such as weather, seasonal changes, *etc*. on the prediction performance. It is necessary to incorporate these factors into the model in the future to improve the prediction accuracy. Although differential privacy and homomorphic encryption had been used to protect user location privacy in this study, how to dynamically adjust the privacy budget to adapt to changing security requirements in practical applications still needs further exploration. In addition, how to optimize communication efficiency and model convergence speed of federated learning in large-scale heterogeneous data environments is also an urgent issue.

Acknowledgments

This study was supported by Ningxia Hui Autonomous Region Natural Science Foundation Project (Grant No. 2022AAC03346).

References

1. Li XH, Zhu Q. 2019. Game based incentive mechanism for cooperative spectrum sensing with mobile crowd sensors. Wirel Netw. 25(4):1855-1866.
2. Gad-ElRab AAA, Alsharkawy AS. 2018. Statistical-based data quality model for mobile crowd sensing systems. Arab J Sci Eng. 43(12):8195-8207.
3. Wang JZ, Li HM, Lu HY. 2018. Application of a novel early warning system based on fuzzy time series in urban air quality forecasting in China. Appl Soft Comput. 71:783-799.
4. Sharma E, Deo RC, Prasad R, Parisi AV, Raj N. 2020. Deep air quality forecasts: suspended particulate matter modeling with convolutional neural and long short-term memory networks. IEEE Access. 8:209503-209516.
5. Spiridonov V, Jakimovski B, Spiridonova I, Pereira G. 2019. Development of air quality forecasting system in Macedonia, based on WRF-Chem model. Air Qual Atmos Health. 12(7):825-836.
6. Casciaro G, Cavaiola M, Mazzino A. 2022. Calibrating the CAMS European multi-model air quality forecasts for regional air pollution monitoring. Atmospheric Environ. 287:119259.
7. Silva N, Marques ERB, Lopes LMB. 2018. FLUX: A platform for dynamically reconfigurable mobile crowd-sensing. ACM Trans Sens Netw. 14(3-4):1-25.
8. Aram SA, Nketiah EA, Saalidong BM, Wang H, Afitiri AR, Akoto AB, *et al*. 2024. Machine learning-based prediction of air quality index and air quality grade: a comparative analysis. Int J Environ Sci Technol. 21(2):1345-1360.
9. Xu Z, Mei L, Choo KKR, Lv ZH, Hu CP, Luo XF, *et al*. 2018. Mobile crowd sensing of human-like intelligence using social sensors: A survey. Neurocomputing. 279:3-10.
10. Middya AI, Roy S. 2022. Pollutant specific optimal deep learning and statistical model building for air quality forecasting. Environ Pollut. 301:118972.
11. Lee S, Park S, Lee MI, Kim G, Im J, Song CK. 2022. Air quality forecasts improved by combining data assimilation and machine learning with satellite AOD. Geophys Res Lett. 49(1):e2021GL096066.
12. Wang ZJ, Huang DJ. 2018. Privacy-preserving mobile crowd sensing in ad hoc networks. Ad Hoc Netw. 73:14-26.
13. Ng MK, Zhu ZC. 2019. Sparse matrix computation for air quality forecast data assimilation. Numer Algorithms. 80(3):687-707.

14. Pisoni E, Albrecht D, Mara TA, Rosati R, Tarantola S, Thunis P. 2018. Application of uncertainty and sensitivity analysis to the air quality SHERPA modelling tool. Atmospheric Environ. 183:84-93.
15. Espinosa F, Bartolomé AB, Hernández PV, Rodriguez-Sanchez MC. 2022. Contribution of singular spectral analysis to forecasting and anomalies detection of indoors air quality. Sensors. 22(8):3054.
16. Guo QGH, He ZF, Li SS, Li XZ, Meng JJ, Hou ZF, *et al*. 2020. Air pollution forecasting using artificial and wavelet neural networks with meteorological conditions. Aerosol Air Qual Res. 20(6):1429-1439.
17. Yu CQ, Tan J, Cheng YH, Mi XW. 2024. Data analysis and preprocessing techniques for air quality prediction: A survey. Stoch Environ Res Risk Assess. 38:2095-2117.
18. Rahman RR, Kabir A. 2023. Spatiotemporal analysis and forecasting of air quality in the greater Dhaka region and assessment of a novel particulate matter filtration unit. Environ Monit Assess. 195(7):824.
19. Wang DY, Wang JZ, Yang YQ, Jia WX, Zhong JT, Wang YQ, *et al*. 2023. Predicting air quality using a quantitative forecasting model of PM2.5 micro-scale variation nested with wave spectrum analysis. Front Environ Sci. 11:1232121.
20. Curceac S, Ternynck C, Ouarda T, Chebana F, Niang SD. 2019. Short-term air temperature forecasting using Nonparametric Functional Data Analysis and SARMA models. Environ Model Softw. 111:394-408.
21. Du RJ, Li JJ, Dong GG, Tian LX, Qing T, Fang GC, *et al*. 2020. Percolation analysis of urban air quality: A case in China. Physica A. 541:123312.
22. Malloy JW, Cerveny RS. 2019. Atmospheric composite sounding analysis as a forecasting tool for ozone accumulation in a semiarid metropolitan area during the North American Monsoon: Phoenix, Arizona. Weather Forecast. 34(4):887-904.
23. Jin XB, Wang ZY, Gong WT, Kong JL, Bai YT, Su TL, *et al*. 2023. Variational Bayesian Network with Information Interpretability Filtering for Air Quality Forecasting. Mathematics. 11(4):837.
24. Li HM, Wang JZ, Yang HF, Wang Y. 2022. Air Quality Deterministic and Probabilistic Forecasting System Based on Hesitant Fuzzy Sets and Nonlinear Robust Outlier Correction. Knowl Based Syst. 237:107789.
25. Nguyen TN, Zeadally S. 2022. Mobile Crowd-Sensing Applications: Data Redundancies, Challenges, and Solutions. ACM Trans Internet Technol. 22(2):1-15.