RESEARCH & INNOVATION DIVISION
IJIST

A Federated Framework for Air Quality Prediction in Smart Cities

Shahan Uddin[1], Talha Waheed[2], Hamid Raza Malik[1], Abdul Basit Dogar[1], Naeem A. Nawaz[1], Kashif Ishaq[1]

[1]School of Systems and Technology, University of Management and Technology, Lahore, Pakistan.

[2]Department of Computer Science, University of Engineering and Technology, Lahore, Pakistan.

*Correspondence: kashif.ishaq@umt.edu.pk

Citation | Uddin. S, Waheed. T, Malik. H. R, Dogar. A. B, Nawaz. N. A, Ishaq. K, "A Federated Framework for Air Quality Prediction in Smart Cities", IJIST, Vol. 07, Issue. 03 pp 1971-1987, August 2025

Received| August 03, 2025 **Revised|** August 15, 2025 **Accepted|** August 17, 2025 **Published|** August 18, 2025.

Over the last couple of decades, due to the constant increase in urbanization and industrialization, the concern in terms of air pollution has become a serious issue. In most cities, the pollution in the air is mostly comprised of Nitrogen Dioxide (NO_2), Ozone (O_3), Carbon Monoxide, and Particulate Matter, all of which can cause serious health issues. There is an emergent need for a system to detect air pollution. This research presents a framework that uses Federated Learning to lessen the communication overhead during the prediction process and ensure data privacy. The research also uses different Machine Learning algorithms, such as Random Forest, Support Vector Machine (SVM), and Logistic Regression, to train and evaluate the research.

Keywords: Air Pollution Detection, Federated Learning, Machine Learning Algorithms, Urbanization and Industrialization Impact, Health Risks

Introduction:

In the context of smart cities, managing air pollution has emerged as a critical challenge with the rapid pace of industrialization and urbanization. These cities host a large number of factories and vehicles that release hazardous pollutants such as CO, NO_2, O_3, and PM2.5 [1]. Writing such regulations is the responsibility of the government, which bases them on the proven harm of the pollutants. Given the urgency of the problem, it is necessary to establish a viable system for air pollution control [2].

Many smart cities have deployed sensors in well- thought- out locations to resolve this issue. These sensors monitor the pollution levels in the air and generate large amounts of real-time data [3]. This data contributes to the calculation of the Air Quality Index (AQI), a key indicator for the state of the air. The Air Quality Index (AQI) offers the public a standardized and easily comprehensible measure for comparing air pollution levels across different cities and locations. This enables individuals to make informed decisions, such as whether to go for walks or avoid high-risk areas, as the information is integrated into various applications for further processing and dissemination [4].

The application of cutting-edge technology, such as big data and machine learning, is one of the key contributions to the development of air pollution control. Extensive research has utilized big data analytics to analyze large datasets, uncover patterns and trends, and provide deeper insights into pollution dynamics [5]. Moreover, machine learning methods have been used to enhance the prediction of air pollution-related events. These models can forecast potential pollution spikes by analyzing traffic patterns, weather conditions, and historical data, thereby enabling pre-emptive interventions [6].

The integration of sensor networks, data analytics, and machine learning has transformed the way air pollution is managed in smart cities. This comprehensive approach not only locates the causes of pollution but also provides the public and government with the means to confront and lessen the harmful impacts of air pollution on the environment and human health [7]. This integrated approach serves as a guiding framework for fostering environmentally conscious and sustainable urban living as smart cities continue to evolve [8].

Problem Statement:

Despite the availability of different methods for the prediction and identification of a number of air pollutants, some issues still need to be addressed. One of the major issues this research focuses on is the communication overheads, such as data volume transmission, frequent updates, model complexity, and scalability, when the data sensed by different sensors is to be communicated to the central server. Even with different ML algorithms, after the training and evaluation phase, the results and the data are communicated to the central server, increasing the communication overhead. A secondary issue is focused on is the security of the data being compiled at the central server, which, being the only collection point, is susceptible to a breach that could result in the loss of important data. These issues can lead to increased latency and operational costs, particularly in areas with limited infrastructure.

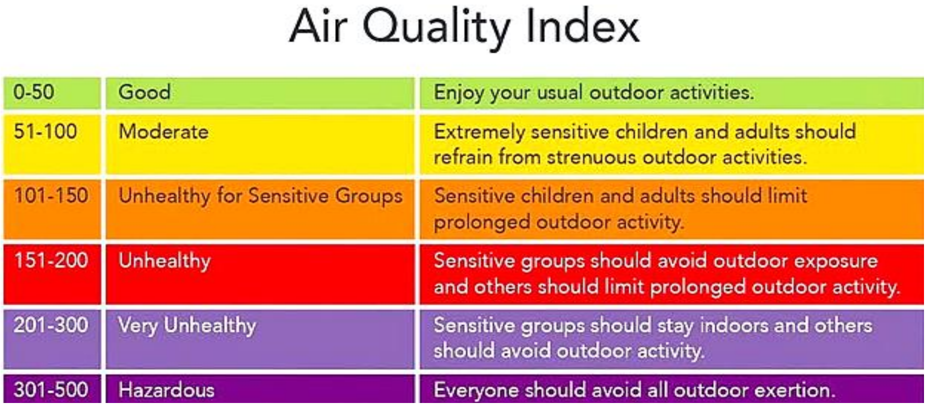
Background:

Air Pollution:

Over the past four decades, rising global temperatures, population growth, and rapid urbanization have collectively contributed to a steady decline in air quality. In short, air pollution refers to the alteration of the natural characteristics of the atmosphere, caused by various physical, chemical, and biological pollutants in both indoor and outdoor environments. Constituents of polluted air, including nitrogen dioxide (NO2), ozone (O3), carbon monoxide (CO), and particulate matter (PM2. 5), exist due to this alteration. Monitoring and assessing air quality is crucial, as these pollutants pose significant risks to both human health and the environment.

In reaction to this environmental issue, various governments have established systems to monitor levels of air pollution, including the Air Quality Index. (AQI). The Air Quality Index (AQI) is A useful summary measure used to provide simple indicators of air quality by outlining the levels of several air pollutants. The AQI serves to categorize air quality into six levels, which correspond to different amounts of pollution and the resulting potential health effects.

These six layers are visualized in Figure 1 below, which shows the different air quality classes. The categories, which typically range from "Good" to "Hazardous," are designed to make it easy for the public and relevant authorities to interpret the potential health hazards of the present conditions of the air quality. The Air Quality Index (AQI) becomes an indispensable tool when making decisions, influencing laws, and igniting public awareness campaigns.



Air Quality Index

0-50	Good	Enjoy your usual outdoor activities.
51-100	Moderate	Extremely sensitive children and adults should refrain from strenuous outdoor activities.
101-150	Unhealthy for Sensitive Groups	Sensitive children and adults should limit prolonged outdoor activity.
151-200	Unhealthy	Sensitive groups should avoid outdoor exposure and others should limit prolonged outdoor activity.
201-300	Very Unhealthy	Sensitive groups should stay indoors and others should avoid outdoor activity.
301-500	Hazardous	Everyone should avoid all outdoor exertion.

Figure 1. Air Quality Index (AQI) Levels [9]

Additionally, the AQI serves as a vital communication tool, an instrument, making it easier for the public to receive information on air quality in real time. This equips individuals with the necessary information to make informed outdoor activity decisions while also providing a foundation for implementing effective pollution control measures. The government's dedication to preserving environmental health and public health in the face of growing pollution challenges is demonstrated by creating and applying AQI systems.

Federated Learning:

Federated Learning (FL) is a branch of Machine Learning (ML) that differs from conventional ML algorithms by operating in a decentralized rather than a centralized manner. Authors in [10] introduced the concept of FL, and the idea behind its creation was to ensure that the data of the local client was not transmitted to the central server after the training and evaluation using ML algorithms.

If explained in a simplified manner, the data used by ML algorithms for training is collected from local clients, such as mobile phones, vehicles, sensors, etc. After the training and evaluation, the data, along with the results from the evaluation, are transmitted from the local client to the central server for aggregation. Though usually effective, this process presents several issues, such as a high communication overhead due to data transmission and results. Another issue is that as the data is transmitted to the central server, it poses a severe risk to privacy. Compared to this, FL solves both issues by performing evaluation and aggregating results in a decentralized manner. This decentralized approach reduces communication overhead by minimizing data transmission and simultaneously enhances privacy by keeping raw data local while only sharing model updates. The decentralized nature of federated learning makes it more resilient to attacks.

To better understand the functioning of Federated Learning (FL), it operates iteratively through multiple rounds of communication between a central server and local clients. In this terminology, the exchanges are more commonly called Federated Learning Rounds [11]. In any scenario, the process of FL starts with the central server first sharing the global update model with all the local clients. Upon receiving the global model, the local clients would then use their data to train and evaluate it using different ML algorithms. Once the evaluation is complete, only the results are forwarded to the central server for aggregation. After aggregating the results into an updated global model, the central server would then transmit the global model to all the clients. This process continues for a defined period, after which it terminates, and the final global model is regarded as the robust model [12], as illustrated in Figure 2.

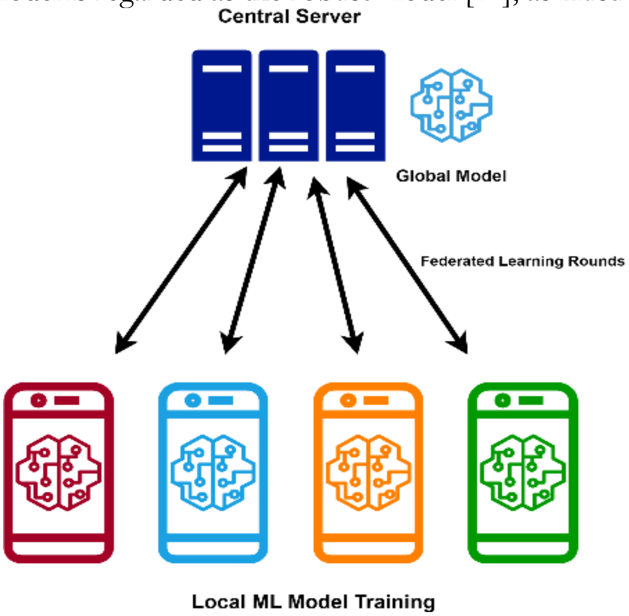


Figure 2. Federated Learning Process

To ensure the working of the FL and the federated learning rounds, the process uses the Federated Averaging Algorithm (FedAvg). The FedAvg algorithm was created by Google, and it was considered the first vanilla FL algorithm for the distributed training of different local clients.

In summary, FL provides a more resilient and efficient training paradigm compared to traditional centralized ML, addressing the dual challenge of efficiency and security. This makes FL particularly suitable for sensitive domains such as healthcare, finance, and environmental monitoring.

Literature Work:

For smart cities, air pollution prediction has amassed several research works over the past few years. These research works range from different types of surveys to experimental papers. After surveying different literature works after the year 2019, it was discovered that most research focused on using different ML algorithms. Some of such research works are briefly discussed below. Authors in [13] presented a research work that details air quality analysis and smog detection using ML regression models such as the Polynomial regression model, Random Forest regression model, Decision tree regression model, and Support Vector Regression model. Using a dataset from the Open Government Data (OGD) Platform India [14], the authors evaluated various regression models and concluded that the Random Forest Regression model outperformed the others. Authors in [15] also used regression algorithms and feature selection techniques to predict PM2.5 in smart cities. In terms of feature selection, the authors used five different techniques: Analysis of Variance, Recursive feature elimination, Variance threshold, random forest, and light gradient boosting. As for the ML algorithms, the authors used six regression and ensemble models: Decision Tree, Extra Tree, Random Forest, XGBoost, AdaBoost, and Light GBM. Using the dataset from five cities in China, the authors concluded that the AdaBoost algorithm and the Light GBM feature selection technique provided the best performance.

Some researchers have focused their research on case studies for specific cities in [16] and [15] and implemented ML algorithms such as Multi-layer Perception (MLP) and Random Forest. These ML algorithms were compared with each other using the Malaysia Air Pollution Dataset for the prediction of PM2.5. Based on their research, the authors concluded that Random Forest performed better than MLP in predicting PM2.5. Similarly, some research works have been done for different cities of India, such as author[15], which tested different ML algorithms to predict the air quality in the capital city of Maharashtra, Nagpur, from which the authors concluded that Boosted Random Forest was the best ML algorithm. While author[16] have tried to analyze the trend in the temporal variations of AQI levels for Pune. The authors have also tried pinpointing the locations in Pune with the most different air pollutants. They processed and used 1 year of data from the smart city office in Pune. Then they used Supervised ML algorithms, Random Forest, and Time-series forecasts to predict air pollution levels.

Beyond India, air pollution prediction research is also being conducted in other regions, such as Australia, where author[17] use a real-world dataset from New South Wales to develop a hybrid deep learning framework for predicting the AQI in smart cities. The authors employ a deep learning forecasting model that integrates 1D-CNN with Bi-GRU. Similarly, authors in. [18] used a 10-year air quality dataset of California to explore a pipeline that stores, processes, and makes predictions using Logistic Regression and Random Forest Classification ML models to predict the AQI values of California.

Aside from the experimental research discussed above, different reviews have also been conducted on air pollution or AQI value prediction in smart cities using ML. Author[19] have extensively reviewed different computing applications in urban environments for air quality predictions using the Internet of Things (IoT), cloud computing, satellites, and different AI/ML methods. Authors[20] [21] reviewed different studies on air pollution prediction using ML algorithms and monitoring based on IoT sensors in the context of different smart cities. They used deep learning techniques, specifically Long Short-Term Memory (LSTM) networks with attention mechanisms, to predict urban air quality. Authors in [22] used different sources when used as monitoring stations and meteorology. With attention mechanisms, the model can learn to assign varying levels of importance to different input features, which helps increase the prediction accuracy. In the work for air-quality prediction in smart cities [23], ensemble-learning algorithms, which include Random Forest, Gradient Boosting Machines, and Support Vector Regression, are utilized to combine multiple base learners, achieving a stronger model for prediction. In most cases, ensemble methods perform better than single models due to the diversity of the base learners. This paper studied the performance of ensemble learning for air quality prediction. Transfer Learning Based Air Quality Prediction in Smart City" presented a transfer learning-based approach to predict air quality in smart cities effectively when the data is scarce or different stationary states between the source and target. The proposed approach utilizes transfer learning by first training models on data-rich cities and then fine-tuning them on target cities with scarce data, thereby improving prediction accuracy through knowledge transfer from source to target domains.

In [24], the problem of generalization and scalability of air quality prediction models across heterogeneous smart cities was considered. This research explored the use of ground and satellite measurements to improve the prediction of urban air quality. To synthesize information from multiple sources, including satellite remote sensing for atmospheric conditions and monitoring stations on the ground for pollutants, they used deep learning techniques to capture the complex spatial and temporal patterns efficiently. The combination of heterogeneous data types enhances urban air quality forecasts by providing more precise and complete predictions. Multi-objective airborne pollutant prediction in smart cities using evolutionary algorithms to jointly optimize prediction accuracy, computational cost, and model interpretability. By taking into account a set of competing objectives, i.e., minimizing the prediction errors and maximizing the diversity of solutions, a pair of sets of trade-offs between different objectives [25].

Table 1. Table describing Air Pollution Prediction Studies

Study	Methodology	Key Findings	Limitations
[13]	ML regression models: Polynomial, Random Forest, Decision Tree, Support Vector Regression	Random Forest regression model performed best for air quality analysis and smog detection.	Lack of consideration for spatial or temporal dynamics and potential overfitting of models due to limited data representation.
[15]	Regression algorithms, Feature selection techniques: ANOVA, Recursive Feature Elimination, Variance Threshold, Random Forest, Light Gradient Boosting	The AdaBoost algorithm with Light GBM feature selection provided the best performance for PM2.5 prediction.	Limited generalizability to other regions, potential biases in data from specific cities.
[16]	ML algorithms: Multi-layer Perceptron (MLP), Random Forest	Random Forest outperformed MLP in predicting PM2.5.	Possible data inconsistencies and a lack of comprehensive evaluation of other ML models.
[17]	ML algorithms: Boosted Random Forest	Boosted Random Forest was Nagpur's best ML algorithm for air quality prediction.	Limited applicability to other cities, potential biases in data from Nagpur.
[18]	Supervised ML algorithms: Random Forest, Time Series Forecast	Analyzed temporal variations of AQI levels in Pune and identified locations with high air pollutant concentrations.	Reliance on data from a single source, potential data quality, or representativeness limitations.
[19]	Hybrid deep learning framework combining 1D-CNN and Bi-GRU	Introduced a hybrid DL framework for AQI prediction in smart cities.	The complexity of DL models may hinder interpretability, pose potential challenges in model deployment, and limit scalability.
[20]	Logistic Regression, Random Forest Classification	Developed a pipeline for AQI prediction in California using ML models.	Limited evaluation of other ML algorithms, and potential biases in data from California.
[21]	Review of computing applications in urban environments	Reviewed various computing applications, including IoT, cloud computing, and ML methods for air quality prediction.	Lack of empirical validation, potential bias in the selection and interpretation of reviewed studies.

[22][23]	Review of studies on air pollution prediction using ML algorithms	Reviewed studies on air pollution prediction using ML algorithms and IoT sensors in smart cities.	Lack of original research, potential biases in the selection and interpretation of reviewed studies.
[24]	Ensemble methods combining Random Forest, Gradient Boosting Machines, and Support Vector Regression	Investigated the effectiveness of ensemble learning for air quality prediction.	The complexity of ensemble models may increase computational costs and potential challenges in model explanation and interpretation.
[25]	LSTM networks with attention mechanisms	Improved air quality prediction accuracy by weighing the importance of input features.	Potential challenges in model interpretability and sensitivity to hyperparameters.

The table summarizes various studies on air pollution prediction in smart cities, detailing their methodologies, data sources, key findings, and associated limitations. Authors in [11] employed ML regression models to analyze air quality and detect smog, finding the Random Forest model to perform best, albeit with potential spatial or temporal dynamics limitations. Author [13] utilized regression algorithms and feature selection techniques to predict PM2.5 levels, highlighting the effectiveness of the AdaBoost algorithm with Light GBM feature selection while acknowledging limited generalizability and potential biases in city-specific data. Other studies, such as authors in [15] and [16], focused on comparing ML algorithms for air quality prediction. However, they may face challenges like data inconsistencies and limited applicability to other cities. Additionally, review studies by authors [21][22] and [23] provided comprehensive overviews of the field but are constrained by a lack of empirical validation and potential biases in the selection and interpretation of reviewed studies. Overall, while these studies contribute valuable insights, they also highlight the need for careful consideration of limitations in air pollution prediction research for smart cities.

Methodology:

In this study, we proposed a novel federated learning (FL) enabled framework to address the critical challenge of air pollution in smart cities. The objective is to enhance the forecasting accuracy of air pollution levels and AQI value at any instant by harnessing the power of a number of machine learning (ML) algorithms, such as Random Forest, Support Vector Machine, and Logistic Regression. This is because Federated Learning offers inherent advantages for addressing the problem by incorporating key aspects such as optimal model aggregation frequency, efficient compression methods, adaptive learning rates, and context-aware participant selection. These methods balance the trade-off between modeling accuracy and communication costs while providing resilience against attacks such as data forging, eavesdropping, device masquerading, and denial-of-service.

Unlike the traditional centralized approach, FL allows the model training to take place on nearby devices or sensors scattered throughout the city. Upon processing by the ML algorithms, just the merged and anonymized results are sent to the central host. This federated training scheme addresses the data privacy and privacy-preserving problem at the same time, it can reduce the communication cost.

This research work proposes an FL-based framework for the prediction of air quality in a smart city. The framework would consist of 3 layers: (i) Application Layer, (ii) Cognitive Layer, and (iii) Data Collection Layer. The proposed framework operates using a bottom-up approach, as illustrated in Figure 3.

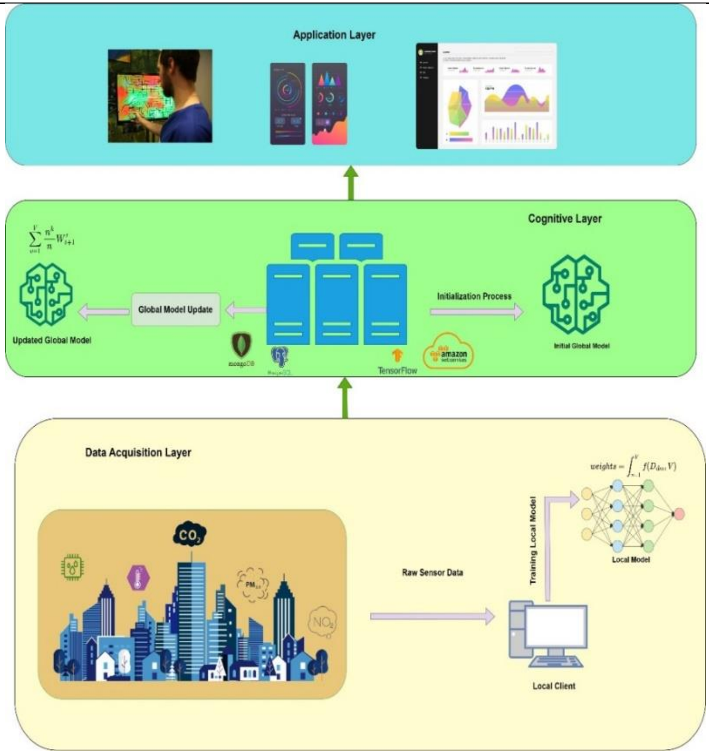


Figure 3. Proposed Federated-Based Framework for Air Quality Prediction in Smart City

As discussed above, the proposed framework operates in a bottom-up manner. Initially, data are gathered from different sensors in a smart city, which range from humidity sensors, temperature sensors, CO_2 sensors, PM2.5 sensors, and NO_2 sensors. After the accumulation of the sensor data, it is transmitted to a local client in a specific grid of the smart city where the sensors are located. Once there, the local client uses the acquired data to train the global model, received from the central server, through the use of different ML algorithms. Once the training is completed, the client communicates the results of the training to the central server present in the cognitive layer. Once there, the results are aggregated and used to update the global model. This process continues for a defined period of iterations after which it concludes. Using the final updated global model, predictions are generated regarding the different air pollutants and transmitted to different air quality monitoring dashboards as represented in the Application layer. The convergence of the global model within a federated learning framework relies on factors such as the number of participating devices, communication frequency, and data heterogeneity. Stopping criteria typically involve maximum training rounds, model stability, and performance metrics on a validation set. By thoughtfully considering these factors, it is possible to effectively determine when the global model has converged.

To better understand the proposed architecture, each layer has been discussed in detail as follows:

Data Acquisition Layer:

In this layer, it is assumed that there are several sensors located around a smart city that are used to gather data in regard to CO_2, NO_2, CO, and PM2.5. To simplify the architecture, a small grid of the smart city is considered as a sample, where all collected data are transmitted to a local client situated at a weather station. In conclusion, the multi-box of the SSD retains the top K predictions, which minimize both location and confidence losses. This is elaborated through Equation 1.

$$S' = \{Sen_1 + Sen_2 + Sen_3 \ldots Sen_n\} \quad (1)$$

Where, S'Represents all of the sensors located in a specific grid. From the equations, it can also be considered that there are n sensors overall present in a smart city gathering data. Once there, the gathered raw data will be trained using different ML models. In place of the gathered data, this research makes use of an air pollution dataset, referred to as D_{AP}. Using this dataset, the local client (*Cl*) would initiate the training process. This is further shown in Equation 2.

$$\int_{\mathrm{Cl}=1}^{\mathrm{Cl}} \mathrm{f}\,(\mathrm{D_{AP}}\,,\mathrm{Cl}) \quad (2)$$

Where the function $f\ (D_{AP}\ ,Cl)$ represents the process of training using the dataset by each local client for a specific grid. Once trained, the trained model results in the form of weights (w) will be communicated to a central server located in the cognitive layer.

Cognitive Layer:

In this layer, the results from the trained local model are aggregated with those from other grids, as shown in Equation 3.

$$w' = w_1, w_2, w_3 \dots w_n (3)$$

Through the accumulation of all the weights, the global model can be generated as represented in Equation 4.

$$C_s = \sum_{Cl_{k=1}}^{Cl_k} \frac{n^k}{n}\, w'_{t+1} \quad (4)$$

Where, C_s represents the central server while Cl_k Represents the serial number of the local client. While n^k Represents the total size of the dataset being used, with n representing the sample size of local clients.

Once this is completed, the central server would transmit the updated global model to a selected number of local clients or weather station terminals, where the global model would be used to train the next batch of data gathered. The accumulated results in the form of a global model would also be shared with cloud storage.

Application Layer:

This layer is associated with an application interface showing a detailed description of the air pollution level and the AQI index level. This application is used by regularly updated users to the different AQI values for their respective smart cities.

Objectives:

The main objectives of this study are:

To develop a federated learning (FL)–enabled framework for accurate air quality prediction in smart cities.

To reduce communication overhead by shifting model training to client devices while maintaining robust performance.

To preserve data privacy and security by preventing raw data transmission to a central server.

To evaluate the performance of different regression-based ML algorithms (Random Forest, Decision Tree, Linear Regression, Support Vector Regression) in both FL and non-FL environments.

To provide a scalable and adaptable solution that can be deployed across heterogeneous smart city environments for real-time air quality forecasting.

Novelty Statement:

This study is novel in that it integrates federated learning with regression-based machine learning models for air quality prediction; an approach not widely explored in the literature. Unlike traditional centralized architectures, the proposed framework offers a privacy-preserving, communication-efficient, and scalable solution. It also provides a direct performance comparison between FL-based and non-FL-based scenarios, demonstrating that even simple models such as Linear Regression benefit significantly from federated training.

Furthermore, the use of the Flower FL framework shows the practicality of adapting open-source federated platforms to real-world environmental monitoring tasks.

Material and Methods:

To evaluate the proposed framework, this study used Flower [25][26], a tool designed for analyzing and assessing federated learning applications. The reasoning behind the use of Flower was that it performs better in terms of system heterogeneity and scalability. Another good feature of Flower is that it has a strong community and incorporation both TensorFlow [27] and PyTorch [28]. The dataset used for this research was obtained from Kaggle (https://www.kaggle.com/datasets/hasibalmuzdadid/global-air-pollution-dataset) to help train and evaluate ML algorithms. It contained global geolocated information regarding NO_2, O_3, CO, and PM2.5 pollutants [29]. The obtained dataset included data about different cities in different countries across the globe and had values recorded for the last decade.

After acquiring the dataset, different pre-processing mechanisms were applied to clean and filter the data from empty and null values and those irrelevant values. Besides containing the different values of air pollutants, the acquired dataset also included the category of each air pollutant. In light of this, the pre-processing also determined each pollutant category and assigned a unique value to each category through One Hot Encoder [30]. One-Hot Encoding was applied because the dataset included categorical attributes for pollutant categories, which cannot be directly processed by regression models. This transformation converted categorical labels into numerical binary vectors, ensuring that all models could interpret and utilize the categorical information without introducing ordinal bias. After this, the final data was converted to NumPy [31] Arrays to be processed by the Flower framework.

In this study, the features correspond to meteorological and pollutant-related attributes (NO_2, O_3, CO, and PM2.5 concentrations along with associated contextual variables), while the target variable is the Air Quality Index (AQI) or pollutant concentration levels to be predicted. This explicit separation ensures the models learn patterns from input pollutant data to estimate the output air quality measure. Once the pre-processing was completed, the final data was partitioned into a 35% split between a set number of clients. The data was also shuffled before the partition so that the same or sequenced data would not be provided to different clients. After this, the FL process was initiated using FL while using the FedAvg algorithm with some specific strategies.

In the initial stage of training, different regression-based machine learning models were considered to analyze the predictive performance of the proposed framework. The selected algorithms included Random Forest Regression (RF), Decision Tree (DT), Linear Regression (LR), and Support Vector Regression (SVR), each offering distinct capabilities for modeling continuous target variables within the air quality dataset. The overall workflow of the proposed framework is illustrated in Figure 4, which summarizes the dataset preprocessing, federated training, and evaluation process.

Results and Discussion:

Results:

The rationale for using regression-based ML algorithms is that the collected data consist of values suitable for classification-based machine learning tasks. Another reason behind this choice was that the gathered data contained values that represented real-time data. To test the efficiency of the results, all the above regression ML models were compared with each other in terms of an FL-based environment and a non-FL-based scenario. The metrics used for this comparison include Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R^2 Score. Table 2 presents in detail the different results obtained for both scenarios.

Evaluating overall performance improvement, the FL-based scenario demonstrates much better performance across most ML algorithms compared to non-FL approaches, with particularly notable improvements in Linear Regression, as shown in Figuress 5 and 6. First, assessing the algorithm-specific performance, Random Forest achieved the best overall performance in both scenarios, with an MAE of 2.13 (Non-FL) vs 2.14 (FL), showing consistent reliability as shown in Figures 5 and 6. Linear Regression showed the most significant improvement in the FL scenario, with R² Score improving from 0.106 to 0.416, indicating better model fit, as illustrated by Table 2 and Figure 5. Decision Tree and SVR demonstrated moderate improvements in the FL environment, as shown in Figure 5.

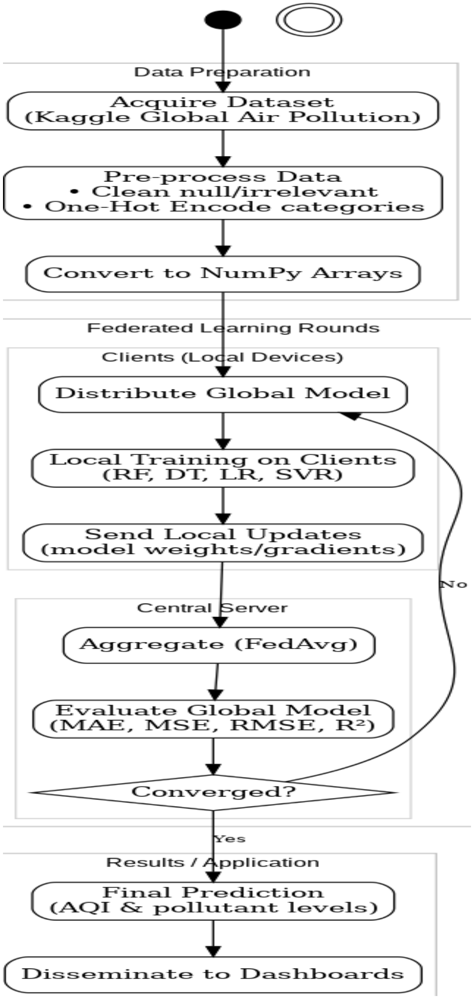

Data Preparation
Acquire Dataset (Kaggle Global Air Pollution)
Pre-process Data • Clean null/irrelevant • One-Hot Encode categories
Convert to NumPy Arrays
Federated Learning Rounds
Clients (Local Devices)
Distribute Global Model
Local Training on Clients (RF, DT, LR, SVR)
Send Local Updates (model weights/gradients)
No
Central Server
Aggregate (FedAvg)
Evaluate Global Model (MAE, MSE, RMSE, R²)
Converged?
Yes
Results / Application
Final Prediction (AQI & pollutant levels)
Disseminate to Dashboards

Figure 4. Workflow Diagram of Proposed Framework

Table 2. Comparison of Non-FL and FL-based Scenarios

	Non-FL				**FL-Based**			
	Mean Absolute Error (MAE)	**Mean Squared Error (MSE)**	**RMSE**	**R^2 Score**	**Mean Absolute Error (MAE)**	**Mean Squared Error (MSE)**	**RMSE**	**R^2 Score**
Random Forest	2.13	12.500	3.53	0.534	2.14	12.84	3.58	0.511
Decision Tree	2.356	20.19	4.49	0.247	2.43	21.34	4.61	0.189
Linear Regression	2.52	23.92	4.89	0.106	2.45	15.35	3.91	0.416
Support Vector Regression	2.54	21.93	4.68	0.182	2.54	22.46	4.73	0.146

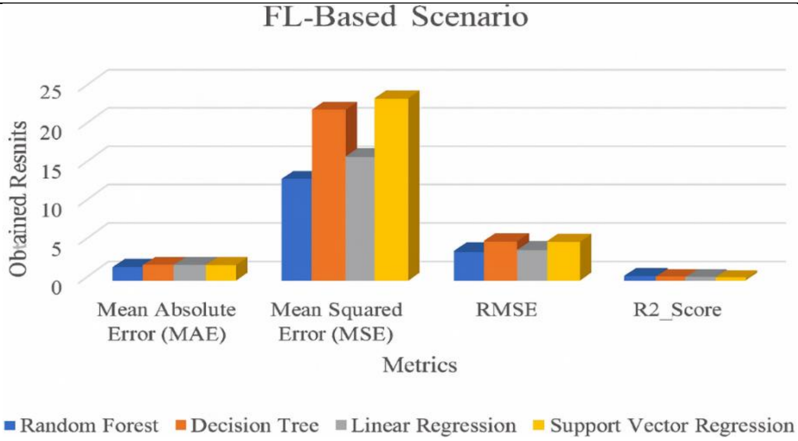


Figure 5. Proposed FL-Based Framework for Air Quality Prediction in Smart City

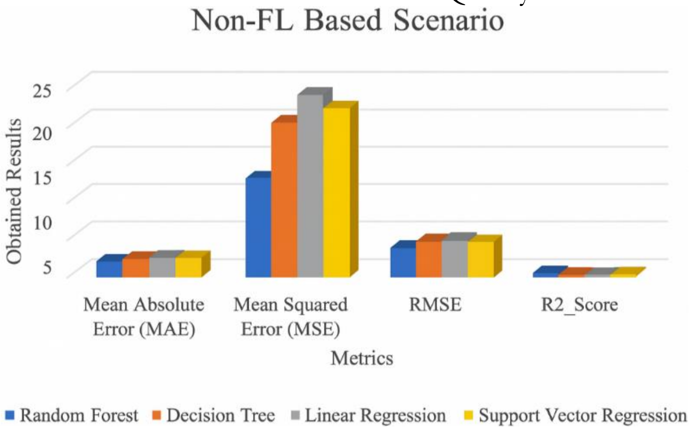


Figure 6. Non-FL-based Scenario

Furthermore, analyzing the performance metrics, about MAE, FL-based models showed competitive or slightly improved absolute error rates. Whereas MSE and RMSE, most FL models achieved lower squared errors, indicating better prediction accuracy. In the R^2 Score, it showed significant improvements in FL scenarios, particularly for Linear Regression (0.106 to 0.416) as illustrated in Table 2 and Figure 5.

Additionally, the FL approach achieved these enhanced results while reducing communication overhead by maintaining data locally and only sharing model parameters.

Discussion:

The findings of this study indicate that a federated learning (FL) paradigm can provide competitive predictive performance for air quality forecasting while preserving data privacy and reducing centralized data transfer. Across the evaluated regression models, Linear Regression (LR) showed a marked improvement under the FL setup relative to the non-FL baseline, suggesting that federated aggregation can enhance generalization even for simple models when data are distributed across heterogeneous clients. Random Forest (RF) remained consistently strong in both settings, reflecting its robustness to data variability.

The framework also demonstrates a practical balance between privacy and utility: model parameters, not raw data, are shared with the central server, which limits exposure of sensitive local records while still enabling global model refinement. This is particularly relevant for smart-city deployments where sensor networks and municipal datasets are fragmented across organizations and jurisdictions. The use of Flower as the orchestration layer further supports scalability to varying numbers of clients and heterogeneous compute environments.

From an operational perspective, the training pipeline (pre-processing, client-side learning, and FedAvg aggregation) proved effective without requiring complex architectures, making the approach computationally tractable. While Decision Tree (DT) and Support

Vector Regression (SVR) were comparatively less performant than RF and LR, they still benefitted from the federated setting, indicating that FL can offer gains even for models that traditionally underperform in centralized scenarios.

Comparison with Existing Studies:

Prior research commonly reports strong performance from tree-based ensembles in centralized settings. Studies such as [11] and [14] found that Random Forest outperforms alternative regressors (and even MLP in some cases). Our results are consistent with this trend; RF remains a robust choice in our experiments, while also finding that LR benefits substantially from the FL setup, narrowing the gap to ensemble methods in certain cases.

Work such as [13] highlights that boosting-based ensembles (e.g., AdaBoost/gradient boosting families) with feature selection can achieve state-of-the-art accuracy for PM forecasting under centralized training. Although our study did not evaluate boosting algorithms, our FL results suggest a complementary path: privacy-preserving improvements via decentralized training, even without specialized feature selection and boosting.

Deep learning approaches (e.g., hybrid CNN/RNN architectures) reported in [17] demonstrate strong accuracy but at higher computational and deployment costs, and with reduced interpretability. In contrast, our FL framework shows that classical regression models can be made competitive and scalable in distributed, privacy-sensitive environments, and an attractive property for resource-constrained smart-city deployments.

Our study aligns with prior findings that RF is a strong baseline, as shown in Table 3, compared with existing studies. It extends the literature by showing that LR can gain notably under FL, improving generalization when data are isolated within specific constraints, which are difficult to access or share across different parts or areas. It also contributes a privacy-preserving, communication-efficient training architecture using Flower that is readily adaptable to heterogeneous clients.

Table 3. Comparison of Proposed Framework with Existing Studies

Study	Learning Paradigm	Models Evaluated	Key Findings Reported	Relation to Our Results
[11]	Centralized	RF, DT, SVR, others	RF reported as best performer	Consistent: RF strong; our FL also boosts simple LR notably
[13]	Centralized (+ feature selection)	Tree ensembles / boosting	Boosting + FS achieves top accuracy	Complementary: FL gives privacy + generalization gains without boosting
[14]	Centralized	RF vs MLP	RF > MLP	Consistent: RF robust; FL adds privacy and retains competitiveness
[17]	Centralized deep learning	Hybrid CNN/RNN	High accuracy, higher complexity	Our classical ML + FL is lighter, interpretable, and privacy-preserving
Proposed Framework	**Federated vs Centralized**	**RF, DT, LR, SVR**	**LR improves under FL; RF remains strong**	**Shows FL can enhance generalization and privacy with modest complexity**

Conclusion:

As the issues associated with urbanization and population expansion increase, our research aims to address the pressing need for accurate systems that can forecast and control air pollution levels, which substantially influence public health. Using the idea of Federated Learning (FL) in the context of smart cities, the research presents a carefully thought-out architecture to precisely calculate the Air Quality Index (AQI) for particular cities. Using the Flower tool for FL analysis renowned for its skillful management of system heterogeneity and scalability, the framework is guaranteed to function reliably in a constantly changing and dynamic setting. The study uses a large dataset from Kaggle that includes geolocated data on NO_2, O_3, CO, and PM2.5 to support the efficacy of the suggested design of five contaminants worldwide within the last ten years. A series of pre-processing steps, such as grouping and conversion to NumPy arrays, prepares the data to be analyzed by FL. During experimentation, a number of regression machine learning models are utilized, such as Support Vector Regression, Decision Trees, Random Forest Regression, and Linear Regression. The FL-based architecture always achieves better results as compared to using a non-FL architecture, which verifies the flexibility to decentralize and process real-time air pollution data. Given the urbanization trend, this study underscores the crucial requirement for accurate air pollution forecasting systems and presents a robust solution, such as the proposed FL-based model. Our findings demonstrate that FL-based methods perform well on a large number of possibly decentralized datasets and enable the development of accurate and reliable models for air quality prediction.

References:

[1] J. R. B. Michael Guarnieri, "Outdoor air pollution and asthma," *Lancet*, vol. 383, no. 9928, pp. 1581–1592, 2014, doi: 10.1016/S0140-6736(14)60617-6.

[2] A. R. Honarvar and A. Sami, "Towards Sustainable Smart City by Particulate Matter Prediction Using Urban Big Data, Excluding Expensive Air Pollution Infrastructures," *Big Data Res.*, vol. 17, pp. 56–65, 2019, doi: https://doi.org/10.1016/j.bdr.2018.05.006.

[3] H. Gupta, D. Bhardwaj, H. Agrawal, V. A. Tikkiwal, and A. Kumar, "An IoT Based Air Pollution Monitoring System for Smart Cities," *1st IEEE Int. Conf. Sustain. Energy Technol. Syst. ICSETS 2019*, pp. 173–177, Feb. 2019, doi: 10.1109/ICSETS.2019.8744949.

[4] L. Z. Zeba Idrees, "Low cost air pollution monitoring systems: A review of protocols and enabling technologies," *J. Ind. Inf. Integr.*, vol. 17, p. 100123, 2020, doi: https://doi.org/10.1016/j.jii.2019.100123.

[5] F. R. Ditsuhi Iskandaryan, "Air Quality Prediction in Smart Cities Using Machine Learning Technologies Based on Sensor Data: A Review," *Appl. Sci*, vol. 10, no. 7, p. 2401, 2020, doi: https://doi.org/10.3390/app10072401.

[6] U. Mahalingam, K. Elangovan, H. Dobhal, C. Valliappa, S. Shrestha, and G. Kedam, "A machine learning model for air quality prediction for smart cities," *2019 Int. Conf. Wirel. Commun. Signal Process. Networking, WiSPNET 2019*, pp. 452–457, Mar. 2019, doi: 10.1109/WISPNET45539.2019.9032734.

[7] H. Brendan McMahan, E. Moore, D. Ramage, S. Hampson, and B. Agüera y Arcas, "Communication-Efficient Learning of Deep Networks from Decentralized Data," *Proc. 20th Int. Conf. Artif. Intell. Stat. AISTATS 2017*, Feb. 2016, Accessed: Jun. 17, 2025. [Online]. Available: https://arxiv.org/pdf/1602.05629

[8] J. R. Keith Bonawitz, Hubert Eichner, Wolfgang Grieskamp, Dzmitry Huba, Alex Ingerman, Vladimir Ivanov, Chloe Kiddon, Jakub Konečný, Stefano Mazzocchi, H. Brendan McMahan, Timon Van Overveldt, David Petrou, Daniel Ramage, "Towards Federated Learning at Scale: System Design," *Proc. Mach. Learn. Syst.*, 2019, doi: https://doi.org/10.48550/arXiv.1902.01046.

[9] Greener Ideal, "The Air Quality Index Explained." Accessed: Aug. 27, 2025. [Online]. Available: https://www.airnow.gov/aqi/aqi-basics/

[10] M. S. F. Momina Shaheen, "Applications of Federated Learning; Taxonomy, Challenges, and Research Trends," *Electronics*, vol. 11, no. 4, p. 670, 2022, doi: https://doi.org/10.3390/electronics11040670.

[11] A. Mishra, Z. M. Jalaluddin, and C. V. Mahamuni, "Air Quality Analysis and Smog

Detection in Smart Cities for Safer Transport using Machine Learning (ML) Regression Models," *Proc. - 2022 IEEE 11th Int. Conf. Commun. Syst. Netw. Technol. CSNT 2022*, pp. 200–206, 2022, doi: 10.1109/CSNT54456.2022.9787618.

[12] "National Informatics Centre - Homepage | India." Accessed: Aug. 27, 2025. [Online]. Available: https://www.nic.gov.in/

[13] A. Banga, R. Ahuja, and S. C. Sharma, "Performance analysis of regression algorithms and feature selection techniques to predict PM2.5 in smart cities," *Int. J. Syst. Assur. Eng. Manag.*, vol. 14, no. 3, pp. 732–745, Jul. 2023, doi: 10.1007/S13198-020-01049-9/METRICS.

[14] R. M. and N. Palanichamy, "Smart City Air Quality Prediction using Machine Learning," *5th Int. Conf. Intell. Comput. Control Syst. (ICICCS), Madurai, India*, pp. 1048–1054, 2021.

[15] G. Sakarkar, S. Pillai, C. V. Rao, A. Peshkar, and S. Malewar, "Comparative Study of Ambient Air Quality Prediction System Using Machine Learning to Predict Air Quality in Smart City," *Lect. Notes Networks Syst.*, vol. 116, pp. 175–182, 2020, doi: 10.1007/978-981-15-3020-3_16.

[16] P. Sonawane, S. Dhanawade, V. Barangule, A. Kulkarni, and P. Mahalle, "Air Quality Analysis & Prediction Using Machine Learning: Pune Smart City Case Study," *2023 IEEE 8th Int. Conf. Converg. Technol. I2CT 2023*, 2023, doi: 10.1109/I2CT57861.2023.10126304.

[17] Z. R. Banani Ghose, "A Deep Learning based Air Quality Prediction Technique Using Influencing Pollutants of Neighboring Locations in Smart City," *JUCS - J. Univers. Comput. Sci.*, vol. 28, no. 8, pp. 799–826, 2022, doi: 10.3897/jucs.78884.

[18] L. Li, Z. Li, L. Reichmann, and D. Woodbridge, "A scalable and reliable model for real-time air quality prediction," *Proc. - 2019 IEEE SmartWorld, Ubiquitous Intell. Comput. Adv. Trust. Comput. Scalable Comput. Commun. Internet People Smart City Innov. SmartWorld/UIC/ATC/SCALCOM/IOP/SCI 2019*, pp. 51–57, Aug. 2019, doi: 10.1109/SMARTWORLD-UIC-ATC-SCALCOM-IOP-SCI.2019.00053.

[19] D. N. Akshara Kaginalkar, Shamita Kumar, Prashant Gargava, "Review of urban computing in air quality management as smart city service: An integrated IoT, AI, and cloud technology perspective," *Urban Clim.*, vol. 39, p. 100972, 2021, doi: https://doi.org/10.1016/j.uclim.2021.100972.

[20] P. K. Bai, "Air Quality Monitoring System Using Machine Learning and IOT," *Int. J. Innov. Res. Inf. Secur.*, vol. 10, no. 03, pp. 369–379, Apr. 2024, doi: 10.26562/IJIRIS.2024.V1003.40.

[21] J. A. D. and R. S. S. Veera Manikandan, Y. Abilash, S. Hari Prasanth, "Internet of Things Enabled ML for Air Quality Assessment: Systematic Review," *7th Int. Conf. Intell. Comput. Control Syst. (ICICCS), Madurai, India*, pp. 1509–1514, 2023.

[22] Y. M. Yun Chia Liang, "Machine Learning-Based Prediction of Air Quality," *Appl. Sci*, vol. 10, no. 24, p. 9151, 2020, doi: https://doi.org/10.3390/app10249151.

[23] K. P. Shilpa Sonawani, "Predicting air quality in smart city using novel transfer learning based framework," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 32, no. 2, p. 11, 2023, [Online]. Available: https://ijeecs.iaescore.com/index.php/IJEECS/article/view/30865

[24] X. Y. Jia Liu, Tianrui Li, Peng Xie, Shengdong Du, Fei Teng, "Urban big data fusion based on deep learning: An overview," *Inf. Fusion*, vol. 53, pp. 123–133, 2020, doi: https://doi.org/10.1016/j.inffus.2019.06.016.

[25] N. D. L. Daniel J. Beutel, Taner Topal, Akhil Mathur, Xinchi Qiu, Javier Fernandez-Marques, Yan Gao, Lorenzo Sani, Kwing Hei Li, Titouan Parcollet, Pedro Porto Buarque de Gusmão, "Flower: A Friendly Federated Learning Research Framework,"

arXiv:2007.14390, 2020, doi: https://doi.org/10.48550/arXiv.2007.14390.

[26] R. Espinosa, F. Jiménez, and José Palma, “Multi-objective evolutionary spatio-temporal forecasting of air pollution,” *Futur. Gener. Comput. Syst.*, vol. 136, pp. 15–33, 2022, doi: https://doi.org/10.1016/j.future.2022.05.020.

[27] B. Pang, E. Nijkamp, and Y. N. Wu, “Deep Learning With TensorFlow: A Review,” *J. Educ. Behav. Stat.*, vol. 45, no. 2, pp. 227–248, Apr. 2020, doi: 10.3102/1076998619872761;CTYPE:STRING:JOURNAL.

[28] K. B. P. S. Imambi, “PyTorch,” Programming with TensorFlow: Solution for Edge Computing Applications. Accessed: Aug. 27, 2025. [Online]. Available: https://pytorch.org/

[29] Kaggle, “Global Air Pollution Dataset.” Accessed: Aug. 27, 2025. [Online]. Available: https://www.kaggle.com/datasets/hasibalmuzdadid/global-air-pollution-dataset

[30] Scikit Learn, “OneHotEncoder .” Accessed: Aug. 27, 2025. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html

[31] “NumPy.” Accessed: Aug. 27, 2025. [Online]. Available: https://numpy.org/