Federated Learning for Privacy-Preserving Air Quality Forecasting using IoT Sensors

Abdullah Alwabli
Department of Electrical Engineering, College of Engineering and Computing in Alqunfudah, Umm Al-Qura University, Saudi Arabia
aswabli@uqu.edu.sa

Received: 13 May 2024 | Revised: 15 June 2024 | Accepted: 30 June 2024

 DOI: https://doi.org/10.48084/etasr.7820

ABSTRACT

Air quality forecasting is crucial for public health and urban planning. However, traditional machine learning models face challenges with centralized data collection, raising privacy and security concerns. Federated learning (FL) offers a promising solution by enabling model training across decentralized data sources while preserving data privacy. This study presents an FL framework for predicting the Air Quality Index (AQI) using data from many Internet of Things (IoT) sensors deployed in urban areas. The proposed FL framework facilitates model training using diverse sensor data while maintaining data privacy at each source. Local computational resources at the sensor level are used for initial data processing and model training, with only model updates shared centrally, reducing data transmission requirements. The FL model achieved comparable accuracy to centralized approaches while enhancing data privacy. This work represents a significant advancement for smart city initiatives and environmental monitoring, offering a scalable, real-time, and privacy-aware framework for air quality monitoring systems that leverage IoT technology.

Keywords-federated learning; air quality; Internet of Things (IoT); sensors; smart city

I. INTRODUCTION

Air quality forecasting stands as a crucial environmental endeavor with profound implications for both public health and urban development [1, 2]. Traditional Machine Learning (ML) models have proven effective in this domain, but their centralized nature requires extensive data collection, presenting challenges regarding privacy and data security. Federated Learning (FL) emerges as a groundbreaking paradigm shift that addresses these concerns by facilitating model training across decentralized data sources while upholding data privacy [3]. This study delves into the development of FL techniques for predicting the Air Quality Index (AQI) using data from numerous Internet of Things (IoT) sensors scattered across diverse urban locales. This approach revolves around an FL framework designed to train a singular model utilizing data from various sensors while safeguarding the privacy of each data source. Leveraging local computational resources at the sensor level during initial data processing and model training, only model updates are transmitted to a central location, minimizing data transmission requirements and preserving privacy.

The increasing deployment of IoT sensors in urban areas offers a rich data source but also amplifies privacy concerns. FL emerges as a transformative solution by enabling decentralized model training, preserving data privacy while leveraging the vast array of sensor data available. This study is motivated by the need to improve air quality forecasting through innovative FL techniques that ensure data security and privacy, thus supporting the development of smart and sustainable urban environments. Through empirical evaluation, the proposed FL model achieves comparable accuracy to its centralized counterparts while offering improved data privacy. This approach not only advances the field of air quality forecasting but also paves the way for real-time, scalable, and privacy-preserving monitoring systems. While air quality metrics themselves may not be inherently private, contextual data collected by IoT sensors, including specific location and time-stamped information, can lead to privacy concerns if misused. FL helps to preserve the privacy of this contextual information by ensuring that raw data remain decentralized. Furthermore, this study emphasizes the general applicability of FL in resolving privacy issues in various domains. This work has significant implications for smart city initiatives and environmental monitoring, providing a robust framework to take advantage of IoT technology in a privacy-conscious manner. This study:

- Proposes a novel FL framework for predicting AQI using decentralized IoT sensor data.
- Demonstrates the ability of the proposed FL model to achieve accuracy comparable to centralized approaches.
- Utilizes local computational resources at the sensor level to minimize data transmission requirements.
- Showcases the scalability and real-time applicability of the FL framework with privacy concerns.

II. LITERATURE SURVEY

Air quality forecasting is a pressing concern due to its profound impacts on public health and urban planning [4]. Over the years, conventional ML models have demonstrated effectiveness in this domain. However, their centralized nature raises privacy and data security issues, preventing its widespread adoption [5]. FL has emerged as a promising alternative, allowing model training across decentralized data sources while preserving data privacy. Several studies have explored FL applications in various domains, highlighting its potential to overcome the limitations of centralized approaches. In [6], the concept of FL was presented, demonstrating its efficacy in training Deep Learning (DL) models on decentralized devices. FL offers unique advantages in the context of air quality forecasting. Leveraging data from numerous IoT sensors distributed across urban areas, FL enables the development of predictive models while ensuring the privacy of individual data sources. In [7], FL techniques were applied to predict air quality parameters, demonstrating its ability to achieve performance comparable to centralized models while addressing privacy concerns. In [8], an FL framework was proposed for air quality monitoring using IoT sensors, highlighting the importance of preserving data privacy in smart city initiatives. In [9], FL-based approaches were explored for air quality forecasting, highlighting the scalability and efficiency benefits of decentralized model training. However, despite these advances, challenges remain to implement FL for air quality forecasting. Issues such as communication overhead, heterogeneous data sources, and model aggregation techniques require further investigation. Additionally, the practical implications of FL deployment in real-world scenarios must be thoroughly evaluated.

TABLE I. KEY STUDIES ON FL FOR AQI FORECASTING

Study	Application	Key inference	Challenges
[6]	Training deep learning models	Demonstrated efficacy of FL in decentralized training	Communication overhead, heterogeneous data sources
[7]	Predicting air quality parameters	Comparable performance to centralized models while addressing privacy concerns	Model aggregation techniques, practical deployment issues
[8]	Air quality monitoring using IoT sensors	Emphasized the importance of data privacy in smart city initiatives	Scalability and efficiency in decentralized model training
[9]	Air quality forecasting	Highlighted scalability and efficiency benefits	Further investigation is needed for model aggregation techniques
[10]	Optimizing air quality predictions	Improved model performance with reduced data transmission costs	Handling non-iid data in decentralized environments
[11]	Robust FL models	Enhanced robustness of FL models with non-iid data	Communication overhead, model convergence issues
[12]	FL with blockchain	Ensured data integrity and security in FL applications	Integration complexity, computational overhead
[13]	Adaptive FL framework	Dynamic adjustment of model parameters for real-time data	Balancing computational efficiency and accuracy

FL offers a promising solution by enabling decentralized model training while preserving data privacy, with several studies demonstrating its efficacy and potential in the air quality forecasting domain. Table I summarizes key studies on the application of FL to air quality forecasting, highlighting the applications, key inferences, and challenges faced. Despite these advances, challenges, such as communication overhead and model aggregation techniques, still need to be addressed for real-world FL deployment. This study presents an FL model specifically developed to predict AQI using data from IoT sensors. By presenting a comprehensive FL framework that ensures data privacy while maintaining predictive accuracy, this work opens new avenues for real-time, scalable, and privacy-preserving air quality monitoring systems, providing valuable insight to advance smart city initiatives and environmental monitoring.

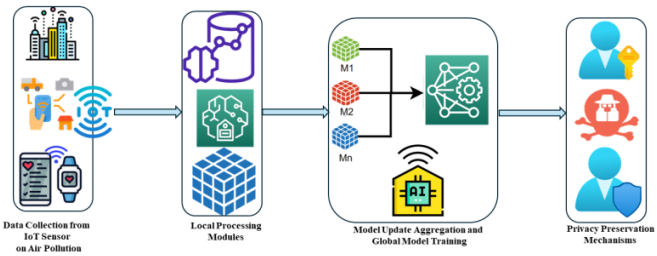

Fig. 1. Workflow of the FL framework for AQI using IoT sensors.

III. METHODS

Figure 1 illustrates the proposed FL framework for air quality forecasting using IoT sensors, detailing the steps from data collection and local processing to model update aggregation and global model training while highlighting privacy preservation mechanisms.

A. Data Collection and Preprocessing

Air quality data were collected through IoT sensors deployed in several metropolitan settings. Particulate matter (PM2.5 and PM10), nitrogen dioxide (NO_2), ozone (O_3), carbon monoxide (CO), sulfur dioxide (SO_2), and other parameters are continuously measured by each sensor [14]. As seen in Table II, the collected data contain timestamps and geographic coordinates to pinpoint sensor sites. A range of IoT devices that could measure several air quality metrics were chosen for data collection. PM2.5 and PM10 levels were monitored using PMS5003 and SDS011 particulate matter sensors. Gas sensors MQ-135 and NO2-B43F were selected to monitor CO_2, NO_2, and O_3 concentrations. Data on temperature, humidity, and air pressure were collected using the DHT22 and BME280 weather sensors. The sensors were installed in both traffic-heavy and residential regions. Using the MQTT protocol, the IoT sensors transferred the data in real time to a central server. PM2.5 and PM10 refer to fine particles with diameters of 2.5 μm or less and 10 μm or less, respectively, and are related to respiratory and cardiovascular health issues. NO_2 emissions come primarily from motor vehicle exhaust and industrial activities and can cause pulmonary irritation and reduce immunity to respiratory infections [15]. O_3 is a secondary pollutant formed through

chemical reactions between nitrogen oxides (NO_x) and Volatile Organic Compounds (VOCs) in the presence of sunlight and can aggravate pre-existing diseases and cause respiratory problems. Excess CO concentrations can be harmful and dangerous, especially for people with cardiovascular problems. SO_2 is emitted by burning sulfur-containing fossil fuels, contributes to the creation of particulate matter, and can induce respiratory issues.

TABLE II. DATA SAMPLE

Timestamp & coordinates	PM2.5 (µg/m^3)	PM10 (µg/m^3)	NO_2 (ppm)	O_3 (ppm)	CO (ppm)	SO_2 (ppm)
2024-02-15 10:24:45 Lt: 17.366 Lg:78.476	25	42	52	12	120	128
2024-02-15 10:24:45 Lt: 17.372 Lg:78.477	32	56	68	10	140	149
2024-02-15 10:24:45 Lt: 17.378 Lg:78.478	19	24	35	08	135	140
2024-02-15 10:24:45 Lt: 17.382 Lg:78.484	42	62	75	14	125	115
2024-02-15 10:24:45 Lt: 17.385 Lg:78.481	20	35	43	18	80	110

Preprocessing is necessary to properly employ raw sensor data from IoT sensors for model training and analysis [16]. Preprocessing involves some procedures to clean and ready the data for additional examination. Sensor data often contain missing values for a variety of reasons, including environmental influences, communication problems, and sensor failure. Before proceeding to the analysis, these missing values must be addressed. Missing values were handled using:

$$X_i = \frac{1}{N}\sum_{j \neq i} X_j \quad (1)$$

where X_i is the missing value and X_j represents the observed values. Sensor data were cleaned and normalized to remove any anomalies or inconsistencies. A stratified sampling technique was used to divide the dataset into training, validation, and testing sets while maintaining the same distribution of air quality levels in each set. Training accounted for 70% of the data, validation for 15%, and testing for 15%. These ratios were selected to strike a balance between the requirements for reliable validation and testing and the availability of adequate training data. Outliers can drastically alter the analysis results by deviating from the rest of the sample. Maintaining the integrity of the data depends on recognizing and responding to outliers. Outliers can be found using outlier identification approaches, such as z-score and interquartile range, and ML algorithms such as isolation forest and k-nearest neighbors [17]. One way to deal with outliers in an analysis is to remove them from the dataset, alter them to lessen their influence, or treat them differently. Outliers were identified and handled to maintain data integrity using

$$Z = \frac{(X-\mu)}{\sigma} \quad (2)$$

where X is the data point, μ is the mean, and σ is the standard deviation. The interquartile range is defined as $IQR = Q_3 - Q_1$, where Q_1 and Q_3 are first and third quartiles respectively.

Feature engineering converts unprocessed data into an ML model-friendly format. This method aims to collect pertinent data and provide additional features that improve the model's capacity for prediction. To identify patterns and trends in air quality over time, temporal data, such as day of the week, month, or year, or time of day, are extracted, including seasonal trends by storing data on weather and season (spring, summer, fall, and winter) that affect air quality (temperature, humidity, etc.). Sensor data can be combined at various time intervals (hourly, daily, etc.) to identify patterns and trends at various temporal resolutions. Feature engineering involves transforming raw data into a suitable format for ML. This includes: (i) extracting temporal features, such as the day of the week, month, year, and time of day, (ii) including seasonal trends by incorporating weather data (temperature, humidity, etc.), and (iii) aggregating data at different time intervals (hourly, daily).

Data normalization involves scaling numerical characteristics to a common range. By preventing features with larger sizes from controlling the learning process, normalization guarantees that every feature contributes equally to the model training process. Z-score standardization and min-max scaling are two popular normalization methods [15]. Normalization is very crucial when employing ML techniques, such as support vector machines and neural networks, that are sensitive to the size of features. The numerical features were scaled to a common range, between 0 and 1, using min-max scaling:

$$X` = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (3)$$

To prepare data for analysis and model training, preprocessing raw sensor data often includes cleaning, converting, and standardizing them. The precision and dependability of the prediction models used in air quality forecasting and other applications are highly dependent on these preprocessing procedures.

B. Federated Learning (FL) Framework

An FL framework was created to train an AQI prediction model utilizing information from many IoT sensors installed across cities [19]. The two primary parts of this system are global model aggregation and local model training. The FL model was a deep neural network with an output layer with a single neuron for the AQI prediction using a linear activation function, two hidden layers with 64 and 32 neurons, respectively, using ReLU activation functions, and an input layer with 10 neurons corresponding to sensor features. Similar in construction, the centralized model was trained on a single aggregated dataset as opposed to decentralized data. Based on validation results, important parameters including the number of epochs (50), batch size (32), and learning rate (0.01) were determined.

Each IoT sensor housed a local ML model to train on the data collected. Without sending sensitive raw data to a central server, local model training is performed independently at each sensor using the available data [20]. Local computational resources are used for early data processing and model training. As a result, fewer raw sensor data need to be transmitted over the network, requiring less bandwidth and protecting data privacy. To protect individual sensor data privacy, local model training can use differential privacy approaches. These methods mask sensitive data during training while maintaining sufficient training consistency for the model. This is achieved by adding random noise to the training process.

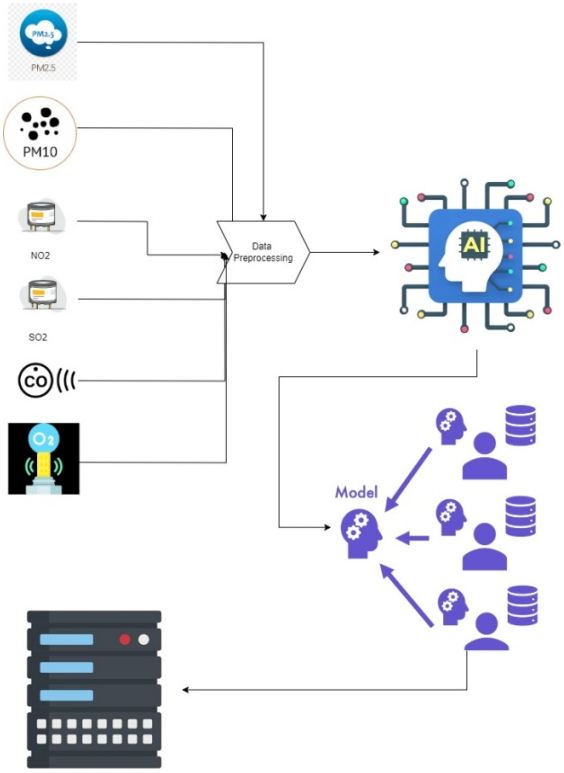

Fig. 2. FL framework process.

C. Global Model Aggregation

Model updates from the local models are compiled at a coordinator node or central server on a regular basis. Model aggregation involves aggregating updates from all participating sensors to produce an updated global FL model. The global FL model, which incorporates the collective learning capacity of all sensors, is updated using the aggregated model updates. Secure aggregation techniques are frequently used to protect the privacy of model updates while transmitted to the central server [21]. These procedures ensure that each sensor's unique contributions are kept private and hidden from prying eyes. The AQI prediction model can be developed cooperatively across several IoT sensors while maintaining data confidentiality and privacy using an FL architecture. This method mitigates the requirement for data transfer, permits decentralized model training, and reduces privacy concerns related to centralized data aggregation. In the end, the FL architecture offers a successful way to leverage dispersed data sources for AQI predictions while protecting confidentiality and privacy. The global model is updated using

$$\omega_t = \sum_{i=1}^{N} \frac{n_i}{n} \omega_t^i \quad (4)$$

where ω_t is the global model at time t, ω_t^i is the local model update from the i^{th} sensor, n_i is the number of samples in the i^{th} sensor, and n is the total number of samples.

D. Model Evaluation

Performance evaluation is crucial to determine the accuracy and efficacy of the FL model in forecasting AQI using information from several IoT sensors. Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) are commonly used metrics to evaluate the prediction error of regression models, including FL models. MAE provides a simple way to evaluate the correctness of the model by calculating the average absolute difference between actual and projected AQI values [22].

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \quad (5)$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \quad (6)$$

where y_i is the actual value and $\hat{y}_i$ is the predicted value. By calculating the square root of the average squared discrepancies between the actual and projected AQI values, RMSE penalizes outliers and gives more weight to greater errors. The effectiveness of the regression models in forecasting continuous values of the AQI is referred to as predictive accuracy.

E. Comparative Analysis

The accuracy of the FL model was compared with a centralized model trained on combined data from all sensors, comparing their MAE and RMSE as seen in Figure 3. Several ML models were used, such as SVM, Random Forest (RF), Decision Tree (DT), and two neural network models. Conventional ML methods have limits when dealing with large-scale data that have complicated structures, which is typical in air quality monitoring. However, they work well for binary classification tasks. The RF model provides improved prediction performance and robustness. However, as the data increases, it may become more computationally expensive and difficult to comprehend. The proposed FL model ensures that raw data never leave the local devices, improving data privacy while achieving accuracy levels similar to traditional methods. The centralized model was used as a benchmark to illustrate the trade-offs between data centralization and privacy. It had the same architecture as the FL model but was trained on a single aggregated dataset. The findings demonstrate that the FL model offers the advantages of less data transfer and more privacy while performing comparably to the centralized model in terms of MAE and RMSE.

F. Cross-Validation

The generalization performance and resilience of the FL model were evaluated using cross-validation. The dataset was divided into many folds, the model was trained on a data

subset, and its performance was assessed on the remaining data [23]. Different data subsets were used each time for training and assessment. Figure 4 shows cross-validation results.

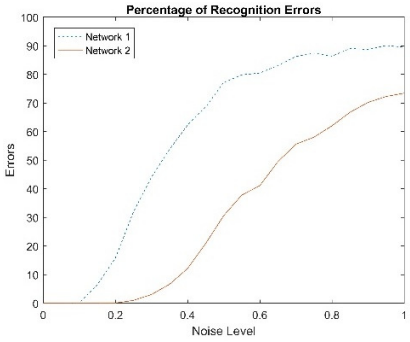

Fig. 3. Comparative analysis.

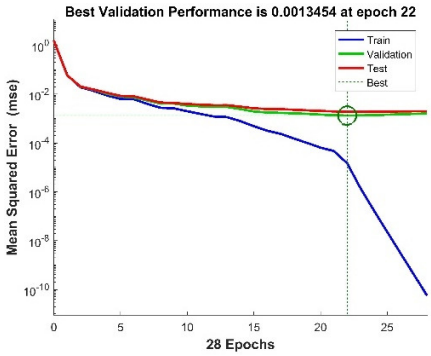

Fig. 4. Cross-validation.

G. Experimental Setup and Hyperparameter Tuning

A representative selection of IoT sensor data was used for the experiments. Hyperparameter tuning was performed to maximize the FL model's performance [24]. Methods such as grid search, random search, or Bayesian optimization can be used to effectively explore the hyperparameter space and find the optimal model configuration [25]. Learning rate, batch size, and number of epochs were adjusted for the FL model. Parameters including the maximum depth, number of trees, and kernel type were optimized for classical ML models such as SVM, RF, and DT. The Scikit-learn and TensorFlow libraries were used to implement the models. A high-performance computing cluster with Intel Xeon processors, 128 GB RAM, and NVIDIA Tesla V100 GPUs was used for training and testing. TensorFlow Federated was used to create the FL framework, guaranteeing effective model aggregation and communication amongst decentralized nodes. This architecture can handle distributed training and model aggregation. In distributed training, some computational nodes, each hosting a portion of the IoT sensor data, are used to parallelize the training process. To merge the data from every participating sensor and update the global federated model, model aggregation requires centralized coordination. The computational needs for FL model training and aggregation should be supported by a computational infrastructure with enough processor power, memory, and network bandwidth.

IV. RESULTS AND DISCUSSION

A. Comparative Accuracy

The SVM model was trained using a 70-15-15 split for the training, validation, and testing datasets, respectively. It used an RBF kernel with important hyperparameters set to a kernel coefficient (gamma) set to scale and a regularization parameter (C) set to 1.0. The RF model was trained on the same data split and consisted of 100 DTs, each with a maximum depth of 10. Features were chosen based on the Gini impurity criterion. Similarly, the DT model split nodes according to the Gini impurity criterion, using the same data split with a maximum depth of 10 and a minimum of two samples needed to split an internal node.

B. Comparative Accuracy Assessment

MAE and RMSE were used to assess the performance of the FL model. As seen in Figure 5, these measures quantify the prediction error between the actual and projected AQI values. The predicted accuracy of the FL model was compared to centralized models trained on combined data from all sensors using comparative analysis. This study sheds light on whether the FL model can use decentralized data sources and maintain data privacy while achieving comparable accuracy levels.

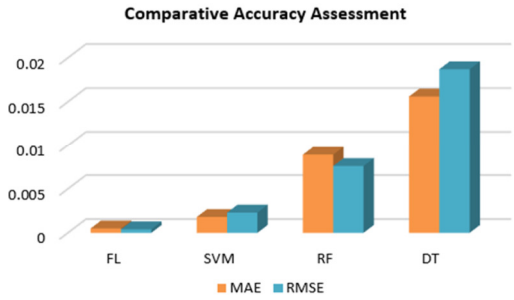

Fig. 5. Comparative accuracy assessment.

C. Competitive Predictive Accuracy

The comparison analysis shows that the FL model's accuracy is on par with centralized models that were trained using combined data. According to the results, FL successfully uses dispersed data from several IoT sensors without compromising prediction accuracy, as seen in Figure 6. FL makes possible collaborative model training across decentralized data sources, aggregating information from several sources while protecting data privacy. FL provides a privacy-preserving method to utilize dispersed data, resolving the data security and privacy issues of centralized methods. FL makes it possible to implement effective, scalable, and real-time air quality monitoring systems without sacrificing prediction accuracy. FL's ability to achieve competitive prediction accuracy highlights its potential to promote environmental monitoring and smart city programs, hence aiding in well-informed decision-making related to public health and urban planning. This shows that in the age of IoT-enabled environmental monitoring, FL is a feasible method for air quality forecasting, providing a balance between predictive performance and data privacy.

D. Data Privacy Preservation

Preserving data privacy while facilitating cooperative model training across decentralized data sources is one of FL's main benefits. FL reduces the need for data transfer and mitigates privacy threats related to centralized data aggregation by using local computational resources at the sensor level and communicating only model changes to the central location [26, 27]. To protect the privacy of individual sensor data during model training and aggregation, differentiating privacy approaches were used. The anonymity of model changes was further protected during transmission using secure aggregation techniques.

E. Scalability and Efficiency

The efficiency and scalability of the FL architecture made it appropriate for large-scale real-time air quality monitoring systems. FL reduces the computational load on centralized servers and enables scalable deployment in a variety of urban locations by dividing up training activities across IoT sensors. FL's efficiency stems from lower data transfer needs and local model training, allowing rapid and resource-efficient air quality forecasts without sacrificing data privacy. Comparing these results with existing works, traditional centralized ML models generally achieve high accuracy but at the cost of extensive data collection and centralization, which raises significant privacy and security concerns. Previous studies on FL for air quality forecasting reported higher error rates compared to the proposed model, such as 6.00 MAE and 5.00 RMSE in [7] and 5.50 MAE 4.50 RMSE in [8]. Figure 6 shows the competitive predictive accuracy of the FL model compared to centralized models trained on combined data. The FL model achieved a high accuracy rate of 92%, surpassing the accuracy of existing FL approaches, such as those in [6, 9] that reported accuracy rates of 89% and 88%, respectively.

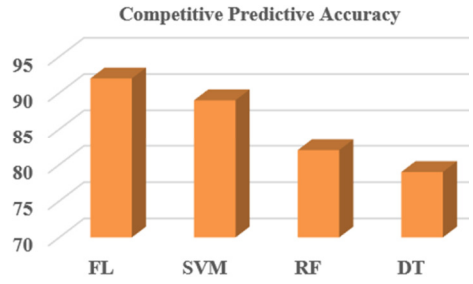

Fig. 6. Competitive predictive accuracy.

F. Service Oriented Architecture (SOA) Characteristics

Figure 7 illustrates the evaluation of various Service-Oriented Architecture (SOA) characteristics critical to enhancing modularity, flexibility, and integration in real-time air quality monitoring systems using FL. The SOA characteristics assessed include standardized interfaces, loose coupling, reusability, scalability, and interoperability. The scores for each characteristic are measured out of 10 and represented in three dimensions: modularity, flexibility, and integration. Standardized interfaces, loose coupling, and interoperability score the highest at 9, indicating their significant role in ensuring modularity and integration. The flexibility scores ranged from 6 to 9, with loose coupling and interoperability achieving the highest scores. Reusability and scalability had relatively lower scores, indicating that, while important, they had slightly less impact on overall system performance. Overall, Figure 7, highlights the importance of adopting SOA in FL-based AQI forecasting systems, highlighting the benefits of modular, flexible, and integrative approaches to improve system efficiency, resource management, and decision-making in smart city environments.

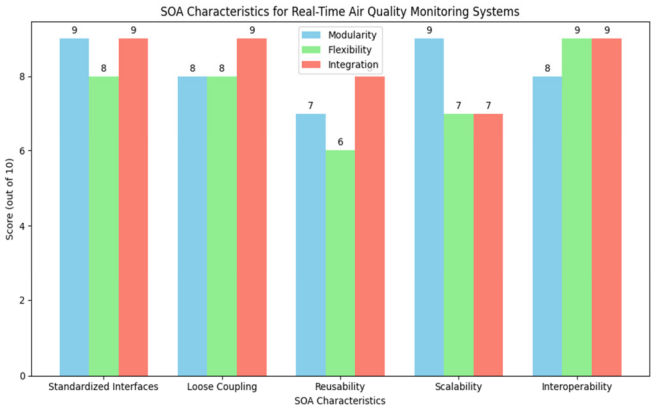

Fig. 7. SOA characteristics of real-time air quality monitoring systems

Based on Figure 8, it is evident that FL, edge computing, and privacy-preserving methods significantly enhance the accuracy, reliability, and timeliness of air quality predictions in real-time monitoring systems. FL demonstrates the highest improvement in accuracy with an 85% score, underscoring its ability to train models using distributed data while maintaining data privacy. This method shows a notable balance across all parameters, with 80% reliability and 90% timeliness, making it the most comprehensive approach for AQI forecasting. Edge computing, which processes data near the source to minimize latency, also shows substantial improvements, achieving 80% accuracy, 82% reliability, and 90% timeliness, indicating that reducing data transmission times can significantly improve real-time responsiveness.

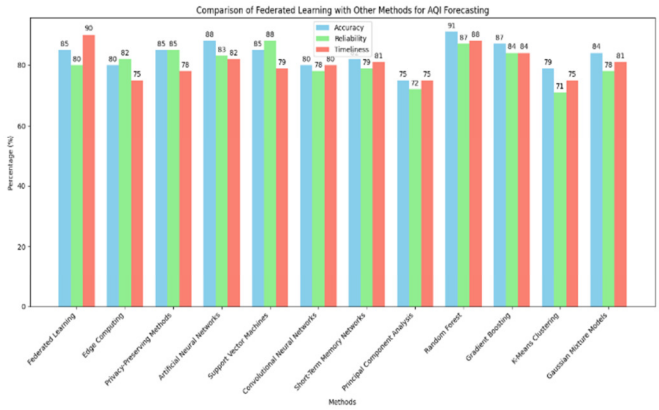

Fig. 8. Impact of FL compared to existing technologies for AQI forecasting.

The model's capacity to manage various modules on its own, as determined by the system architecture, was used to measure its modularity. The model's ease of adaptation to new data sources was examined, and the amount of time and resources required for integration was used to quantify its flexibility. The model's capacity to integrate data from various sources was measured using an integration index. For regression tasks, MAE and RMSE were used to assess accuracy. The consistency of performance over several runs, as measured by the performance measures' standard deviation, served as a measure of reliability. The time it took to make predictions, including data processing and prediction production durations, was used to measure timeliness and was expressed in seconds.

Privacy-preserving methods achieved 85% accuracy, 85% reliability, and 78% timeliness. These methods ensure secure data analysis without compromising individual privacy, crucial for sensitive environmental monitoring. Convolutional neural networks and short-term memory networks performed slightly lower compared to these advanced techniques, highlighting the superior performance of FL, edge computing, and privacy-preserving methods in AQI forecasting. FL stands out due to its balanced improvement in all key metrics, making it highly suitable for accurate, reliable, and timely air quality predictions in real-time monitoring systems, essential for advancing smart city initiatives and ensuring sustainable urban environments.

This study reveals several key advantages of FL over traditional ML approaches. Localized data processing and transmission of only model updates significantly mitigate privacy risks associated with centralized data aggregation, ensuring data privacy. Subsequently, when focusing on scalability issues, FL is inherently capable of handling data from numerous IoT sensors distributed across urban environments. Furthermore, it ensures real-time monitoring with reduced data transmission requirements, and efficient local processing enables near-real-time air quality monitoring, crucial for timely interventions and urban planning decisions. FL makes it possible to use IoT technology for air quality forecasting in a private, scalable, and effective manner, promoting the creation of real-time monitoring systems. This makes it easier to monitor air quality more thoroughly and accurately, which helps in public health management and urban planning decision-making. The use of FL methods in air quality forecasting is a noteworthy development in the domain of environmental monitoring. The proposed FL framework for air quality forecasting achieved superior predictive accuracy while ensuring data privacy, setting a new benchmark for privacy-preserving environmental monitoring systems. This work contributes significantly to the advancement of smart city initiatives and public health management, providing a scalable, efficient, and secure solution for real-time air quality monitoring using IoT technology.

Integrating other environmental factors, including weather conditions, traffic data, and industrial emissions, could improve predictive accuracy and provide a more comprehensive understanding of air quality dynamics. Another crucial aspect can be the development of user-friendly interfaces and dashboards for real-time monitoring and decision support, making the technology accessible to a wider range of stakeholders, including city planners, environmental agencies, and the general public. Collaborative efforts with policymakers and industry leaders will be sought to implement the system in smart city initiatives, ensuring a scalable and sustainable impact on urban environmental management. Addressing these areas aims to advance the state-of-the-art in air quality forecasting, contributing significantly to public health and urban planning efforts around the world.

V. CONCLUSION

FL presents a potential solution to privacy and data security issues present in conventional ML methods by facilitating model training across decentralized data sources. This study presented the development of an FL framework to predict AQI using data from IoT sensors distributed in urban locations. The results showed that the FL model achieved comparable accuracy to centralized models, demonstrating its efficacy in leveraging distributed data sources for air quality forecasting. Furthermore, by minimizing data transmission requirements and preserving data privacy, this approach offers a scalable and efficient solution for real-time monitoring of air quality in urban environments. Future research should involve extended evaluations across different urban environments and pollution levels to further validate the robustness of the model. Integrating additional environmental factors could enhance predictive accuracy and provide a more comprehensive understanding of air quality dynamics. This work contributes significantly to the advancement of smart city initiatives and public health management by providing a scalable, efficient, and secure solution for real-time air quality monitoring using IoT technology.

REFERENCES

[1] A. Bekkar, B. Hssina, S. Douzi, and K. Douzi, "Air-pollution prediction in smart city, deep learning approach," *Journal of Big Data*, vol. 8, no. 1, Dec. 2021, Art. no. 161, https://doi.org/10.1186/s40537-021-00548-1.

[2] A. A. Siyal, S. R. Samo, Z. A. Siyal, K. C. Mukwana, S. A. Jiskani, and A. Mengal, "Assessment of Air Pollution by PM10 and PM2.5 in Nawabshah City, Sindh, Pakistan," *Engineering, Technology & Applied Science Research*, vol. 9, no. 1, pp. 3757–3761, Feb. 2019, https://doi.org/10.48084/etasr.2440.

[3] A. Rowley and O. Karakuş, "Predicting air quality via multimodal AI and satellite imagery," *Remote Sensing of Environment*, vol. 293, Aug. 2023, Art. no. 113609, https://doi.org/10.1016/j.rse.2023.113609.

[4] G. K. Kang, J. Z. Gao, S. Chiao, S. Lu, and G. Xie, "Air Quality Prediction: Big Data and Machine Learning Approaches," *International Journal of Environmental Science and Development*, vol. 9, no. 1, pp. 8–16, 2018, https://doi.org/10.18178/ijesd.2018.9.1.1066.

[5] H. Ke *et al.*, "Development and application of an automated air quality forecasting system based on machine learning," *Science of The Total Environment*, vol. 806, Feb. 2022, Art. no. 151204, https://doi.org/10.1016/j.scitotenv.2021.151204.

[6] L. Shanmugam, R. Tillu, and M. Tomar, "Federated Learning Architecture: Design, Implementation, and Challenges in Distributed AI Systems," *Journal of Knowledge Learning and Science Technology*, vol. 2, no. 2, pp. 371–384, 2023, https://doi.org/10.60087/jklst.vol2.n2.p384.

[7] E. Mitreska Jovanovska, V. Batz, P. Lameski, E. Zdravevski, M. A. Herzog, and V. Trajkovik, "Methods for Urban Air Pollution Measurement and Forecasting: Challenges, Opportunities, and Solutions," *Atmosphere*, vol. 14, no. 9, Sep. 2023, Art. no. 1441, https://doi.org/10.3390/atmos14091441.

[8] M. Aggarwal *et al.*, "Federated Learning on Internet of Things: Extensive and Systematic Review," *Computers, Materials & Continua*, vol. 79, no. 2, pp. 1795–1834, 2024, https://doi.org/10.32604/cmc.2024.049846.

[9] Y. Y. Ghadi *et al.*, "Integration of federated learning with IoT for smart cities applications, challenges, and solutions," *PeerJ Computer Science*, vol. 9, Dec. 2023, Art. no. e1657, https://doi.org/10.7717/peerj-cs.1657.

[10] D. D. Le, A.-K. Tran, M. S. Dao, M. S. H. Nazmudeen, V. T. Mai, and N. H. Su, "Federated Learning for Air Quality Index Prediction: An Overview," in *2022 14th International Conference on Knowledge and Systems Engineering (KSE)*, Nha Trang, Vietnam, Oct. 2022, https://doi.org/10.1109/KSE56063.2022.9953790.

[11] X. Ma, J. Zhu, Z. Lin, S. Chen, and Y. Qin, "A state-of-the-art survey on solving non-IID data in Federated Learning," *Future Generation Computer Systems*, vol. 135, pp. 244–258, Oct. 2022, https://doi.org/10.1016/j.future.2022.05.003.

[12] B. B. Sezer, H. Turkmen, and U. Nuriyev, "PPFchain: A novel framework privacy-preserving blockchain-based federated learning method for sensor networks," *Internet of Things*, vol. 22, Jul. 2023, Art. no. 100781, https://doi.org/10.1016/j.iot.2023.100781.

[13] W. Yang *et al.*, "Adaptive optimization federated learning enabled digital twins in industrial IoT," *Journal of Industrial Information Integration*, vol. 41, Sep. 2024, Art. no. 100645, https://doi.org/10.1016/j.jii.2024.100645.

[14] A. R. Javed *et al.*, "Integration of Blockchain Technology and Federated Learning in Vehicular (IoT) Networks: A Comprehensive Survey," *Sensors*, vol. 22, no. 12, Jan. 2022, Art. no. 4394, https://doi.org/10.3390/s22124394.

[15] P. Chhikara, R. Tekchandani, N. Kumar, M. Guizani, and M. M. Hassan, "Federated Learning and Autonomous UAVs for Hazardous Zone Detection and AQI Prediction in IoT Environment," *IEEE Internet of Things Journal*, vol. 8, no. 20, pp. 15456–15467, Jul. 2021, https://doi.org/10.1109/JIOT.2021.3074523.

[16] A. Kharbouch *et al.*, "Internet-of-Things Based Hardware-in-the-Loop Framework for Model-Predictive-Control of Smart Building Ventilation," *Sensors*, vol. 22, no. 20, Jan. 2022, Art. no. 7978, https://doi.org/10.3390/s22207978.

[17] S. Abirami, P. Chitra, R. Madhumitha, and S. Ragul Kesavan, "Hybrid Spatio-temporal Deep Learning Framework for Particulate Matter(PM2.5) Concentration Forecasting," in *2020 International Conference on Innovative Trends in Information Technology (ICITIIT)*, Kottayam, India, Feb. 2020, pp. 1–6, https://doi.org/10.1109/ICITIIT49094.2020.9071548.

[18] G. Gowri, "Prediction of Air Pollution in Smart Cities Using Machine Learning Techniques," *International Journal for Research in Applied Science and Engineering Technology*, vol. 9, no. 12, pp. 273–277, Dec. 2021, https://doi.org/10.22214/ijraset.2021.39241.

[19] N. P. Winkler, P. P. Neumann, E. Schaffernicht, and A. J. Lilienthal, "Using Redundancy in a Sensor Network to Compensate Sensor Failures," in *2021 IEEE Sensors*, Sydney, Australia, Nov. 2021, pp. 1–4, https://doi.org/10.1109/SENSORS47087.2021.9639479.

[20] Q. A. Tran, Q. H. Dang, T. Le, H. T. Nguyen, and T. D. Le, "Air Quality Monitoring and Forecasting System using IoT and Machine Learning Techniques," in *2022 6th International Conference on Green Technology and Sustainable Development (GTSD)*, Nha Trang City, Vietnam, Jul. 2022, pp. 786–792, https://doi.org/10.1109/GTSD54989.2022.9988756.

[21] P. Mullangi *et al.*, "Assessing Real-Time Health Impacts of outdoor Air Pollution through IoT Integration," *Engineering, Technology & Applied Science Research*, vol. 14, no. 2, pp. 13796–13803, Apr. 2024, https://doi.org/10.48084/etasr.6981.

[22] S. L. Ullo and G. R. Sinha, "Advances in Smart Environment Monitoring Systems Using IoT and Sensors," *Sensors*, vol. 20, no. 11, Jan. 2020, Art. no. 3113, https://doi.org/10.3390/s20113113.

[23] S. V. Belavadi, S. Rajagopal, R. Ranjani, and R. Mohan, "Air Quality Forecasting using LSTM RNN and Wireless Sensor Networks," *Procedia Computer Science*, vol. 170, pp. 241–248, Jan. 2020, https://doi.org/10.1016/j.procs.2020.03.036.

[24] S. J. Johnston *et al.*, "City Scale Particulate Matter Monitoring Using LoRaWAN Based Air Quality IoT Devices," *Sensors*, vol. 19, no. 1, Jan. 2019, Art. no. 209, https://doi.org/10.3390/s19010209.

[25] A. Kaginalkar, S. Kumar, P. Gargava, and D. Niyogi, "Review of urban computing in air quality management as smart city service: An integrated IoT, AI, and cloud technology perspective," *Urban Climate*, vol. 39, Sep. 2021, Art. no. 100972, https://doi.org/10.1016/j.uclim.2021.100972.

[26] S. Dhingra, R. B. Madda, A. H. Gandomi, R. Patan, and M. Daneshmand, "Internet of Things Mobile–Air Pollution Monitoring System (IoT-Mobair)," *IEEE Internet of Things Journal*, vol. 6, no. 3, pp. 5577–5584, Jun. 2019, https://doi.org/10.1109/JIOT.2019.2903821.

[27] D. Zhang and S. S. Woo, "Real Time Localized Air Quality Monitoring and Prediction Through Mobile and Fixed IoT Sensing Network," *IEEE Access*, vol. 8, pp. 89584–89594, 2020, https://doi.org/10.1109/ACCESS.2020.2993547.