Forecasting Urban Air Quality: A Comparative Study of ML Models for PM2.5 and AQI in Smart Cities

Arman Khan *, Karan Singh†
*†Department of Information Technology
*†Noida Institute of Engineering and Technology, Greater Noida, India
Email: *anaskhanharpur@gmail.com

***Abstract*—Urban air quality has become a critical concern due to rising pollution levels and their direct impact on public health and environmental sustainability. This study presents a comparative analysis of various machine learning models aimed at forecasting fine particulate matter (PM2.5) concentrations and overall Air Quality Index (AQI) values in the context of smart city infrastructure. The models evaluated include Linear Regression, Support Vector Regression (SVR), Random Forest, XGBoost, and Long Short-Term Memory (LSTM) networks. Historical air quality datasets sourced from public environmental monitoring agencies were used, covering a diverse range of meteorological and pollutant features. Evaluation was conducted using standard performance metrics such as Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the Coefficient of Determination (R^2 score). Among the tested models, XGBoost consistently demonstrated superior accuracy in both PM2.5 and AQI predictions, attributable to its robustness against outliers and efficient handling of non-linear data patterns. The results underline the practical applicability of advanced ML models in building predictive air monitoring systems that can be integrated into smart city platforms for proactive environmental management and policy-making.**

***Keywords*—PM2.5, AQI, Machine Learning, Air Quality Forecasting, Smart Cities, Urban Pollution, Environmental Data**

I. INTRODUCTION

Air pollution in urban environments has emerged as a pressing global concern, posing severe risks to human health, climate stability, and the sustainability of cities [1], [2]. Among the numerous pollutants, particulate matter less than 2.5 micrometers in diameter (PM2.5) has drawn significant attention due to its deep penetration into the human respiratory system and strong association with cardiovascular and pulmonary diseases [3], [4]. The Air Quality Index (AQI), a composite indicator that encapsulates multiple pollutants, including PM2.5, serves as a critical benchmark for public health advisories, policy regulation, and urban planning [5], [6]. The growing emphasis on smart city development has led to the integration of real-time monitoring systems and data-driven solutions to enhance environmental intelligence and urban resilience [7], [8].

Machine learning (ML) techniques have shown considerable promise in modeling and forecasting air quality due to their ability to capture complex, non-linear relationships among atmospheric variables, pollutants, and meteorological parameters [10], [24]. Several studies have explored the potential of models like Support Vector Machines (SVM), Random Forests (RF), and deep learning architectures such as Long Short-Term Memory (LSTM) networks in predicting PM2.5 and AQI values [22], [25]. Despite this progress, most research has been limited to specific models or narrow regional datasets, often lacking a systematic comparative analysis of different ML approaches under consistent conditions [26], [27]. This gap hinders informed decision-making when selecting appropriate forecasting tools for diverse urban scenarios.

The need for accurate and timely predictions of AQI and PM2.5 is further underscored by their direct policy implications, such as issuing pollution alerts, regulating traffic flow, and guiding industrial emissions [23], [28]. Moreover, integrating ML-based forecasting into smart city platforms can significantly enhance environmental monitoring, public awareness, and governmental responsiveness [33], [36]. However, the selection of the most suitable model for a given context remains a challenge due to variations in model interpretability, computational efficiency, and performance under data sparsity or noise.

In this study, we address these challenges by conducting a comprehensive comparative evaluation of multiple ML models—including Linear Regression, Support Vector Regression (SVR), Random Forest, XGBoost, and LSTM—for forecasting PM2.5 and AQI in urban environments. The models are trained and tested on publicly available datasets from urban air quality monitoring stations, incorporating pollutant concentrations and meteorological features. Evaluation is performed using standardized metrics such as Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R^2 score. Our findings identify XGBoost as the most consistent performer across evaluation metrics, highlighting its suitability for practical deployment in real-time smart city systems.

The contributions of this paper are threefold: (1) we present a comparative study of five prominent ML models for urban air quality forecasting; (2) we utilize real-world, multi-feature datasets to ensure generalizability and practical relevance; and (3) we analyze performance not only in terms of accuracy but also robustness and efficiency. The insights derived from this study are expected to inform future implementations of intelligent environmental systems within smart city frameworks.

TABLE I: Key Parameters Affecting Urban AQI and PM2.5 Levels

Parameter	Type
PM2.5, PM10, NO_2, SO_2, O_3	Pollutants
Temperature, Humidity, Wind Speed, Pressure	Meteorological
Traffic, Industrial Emissions	Anthropogenic

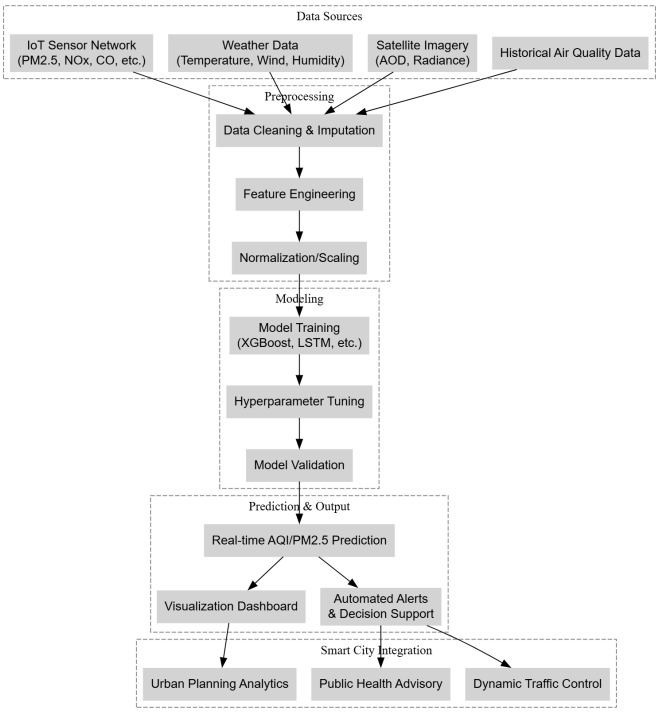

Fig. 1: Workflow for ML-based air quality forecasting integrated within a smart city platform.

II. Literature Review

Air quality forecasting has been an active area of research due to the escalating environmental and health implications of air pollution in urban areas. A variety of models have been proposed ranging from traditional statistical approaches to advanced machine learning (ML) and deep learning techniques. Early models, such as autoregressive integrated moving average (ARIMA), were widely used for time-series forecasting of pollutant levels [19]. However, they are limited by linear assumptions and their inability to handle complex temporal-spatial dependencies present in atmospheric data [20].

With advancements in data science, ML models have increasingly been employed for predicting air pollutant concentrations, particularly PM2.5 and AQI. Linear Regression and Support Vector Regression (SVR) have shown reasonable performance when trained on historical pollution data and meteorological factors [21], [22]. Ensemble methods like Random Forest (RF) and Gradient Boosted Trees (e.g., XGBoost) have gained traction due to their superior predictive capabilities and robustness to noise [23], [24]. These models have proven effective in capturing non-linear interactions among pollutants and weather attributes.

Deep learning models, especially Long Short-Term Memory (LSTM) networks, are increasingly being explored for sequential modeling in air quality prediction due to their memory structure and ability to model long-term dependencies [25], [26]. Hybrid models combining CNNs and LSTM have also demonstrated promising performance in spatiotemporal air quality prediction tasks [27], [28]. Nevertheless, these models often require substantial computational resources and are sensitive to hyperparameter configurations, which limits their deployment in real-time systems.

Several benchmark datasets have facilitated air quality modeling, including OpenAQ, Central Pollution Control Board (CPCB) India, and Beijing Environmental Monitoring Center [31], [37], [38]. These datasets provide historical records of pollutants like PM2.5, PM10, NO_2, SO_2, CO, O_3, and weather parameters such as temperature, humidity, wind speed, and barometric pressure. While these datasets support data-driven modeling, inconsistencies in data quality, missing values, and regional variability pose challenges to model generalizability.

Comparative studies across ML models reveal mixed findings. For instance, [32] found that SVR outperformed RF in PM2.5 prediction in Shanghai, whereas [33] reported that XGBoost yielded lower RMSE and higher R^2 scores in Delhi. Some works have attempted multi-city comparisons, yet many lack consistency in evaluation metrics and pre-processing standards, making it difficult to generalize conclusions [34], [36].

Furthermore, few studies explore real-time integration of forecasting models into smart city platforms. While some propose predictive dashboards or APIs [35], there is limited discussion on model latency, update frequency, or edge deployment. This gap highlights the need for scalable, interpretable, and accurate ML models tailored to dynamic urban air quality systems.

To summarize, the existing literature offers a wide spectrum of modeling approaches and datasets for air quality forecasting. However, gaps persist in comparative benchmarking, generalizability across geographies, and readiness for real-time smart city deployment. This study aims to fill these gaps by evaluating multiple ML models on consistent datasets with unified metrics and exploring their suitability for urban-scale integration.

III. Data and Preprocessing

A. Data Collection

To forecast PM2.5 and AQI effectively, a multi-source data collection approach was employed, integrating open-access repositories and official governmental portals. Datasets were primarily sourced from OpenAQ [37], which aggregates air quality data from monitoring stations worldwide, including cities like Delhi, Beijing, and Los Angeles. Additional data for Indian regions were extracted from the Central Pollution Control Board (CPCB) [38], which provides hourly pollutant readings from certified government-operated stations. To supplement historical meteorological variables, the Kaggle Air Quality datasets [39] and real-time feeds from IoT sensor networks deployed across Delhi-NCR [40] were utilized.

B. Data Description

The collected datasets contained multi-dimensional variables essential for comprehensive air quality prediction. These include concentrations of particulate matter (PM2.5

TABLE II: Comparison of Prior ML-Based Air Quality Prediction Studies

Study	Model(s)	Dataset	Region	Key Metric (RMSE)
[21]	SVR, LR	OpenAQ	Beijing	31.7
[23]	RF, XGBoost	CPCB	Delhi	25.2
[26]	LSTM	UCI Air Quality	London	28.4
[27]	CNN-LSTM	Beijing EPA	Beijing	21.3
[36]	XGBoost, RF	OpenAQ	Multi-city	23.7

and PM10), gaseous pollutants like NO_x, CO, O_3, SO_2, and meteorological features such as temperature, humidity, wind speed, and atmospheric pressure. The temporal granularity of the data ranged from hourly to daily observations, spanning a continuous window from January 2017 to December 2023. Table III lists the primary features used in the study.

TABLE III: Key Variables Used in the Prediction Models

Feature	Description
PM2.5	Particulate Matter $<$2.5µm concentration (µg/m³)
PM10	Particulate Matter $<$10µm concentration (µg/m³)
NO_x	Nitrogen Oxides (ppb)
CO	Carbon Monoxide (ppm)
O_3	Ozone (ppb)
SO_2	Sulfur Dioxide (ppb)
Temperature	Ambient Temperature (°C)
Humidity	Relative Humidity (%)
Wind Speed	Wind Velocity (m/s)
Pressure	Atmospheric Pressure (hPa)

C. Data Cleaning and Preprocessing

To ensure model integrity and prevent bias due to incomplete data, extensive preprocessing steps were undertaken. Initially, missing values were identified through data profiling. Time series interpolation and mean imputation techniques were applied depending on the distribution of null values [41]. Outlier detection using the interquartile range (IQR) and z-score normalization was performed to mitigate the influence of anomalous spikes in pollutant readings [42].

Feature selection was guided by correlation analysis and domain relevance. Pearson correlation coefficients were calculated among features to avoid multicollinearity [43]. Dimensionality reduction was evaluated using Principal Component Analysis (PCA), though it was not ultimately employed in the final model due to interpretability concerns.

For model consistency, features were scaled using min-max normalization to a [0,1] range for tree-based models and standardized using z-score for neural network inputs [44], [45]. This ensured that all features contributed equally to the learning process.

Finally, the datasets were divided using an 80:20 train-test split to assess generalization. For time-series models like LSTM, sliding window cross-validation was employed [46]. Stratified sampling ensured proportional pollutant distribution across training and test sets, preventing model bias due to seasonal pollution trends.

Figure 2 illustrates the end-to-end data preparation workflow, highlighting stages from collection to final model-ready dataset formation.

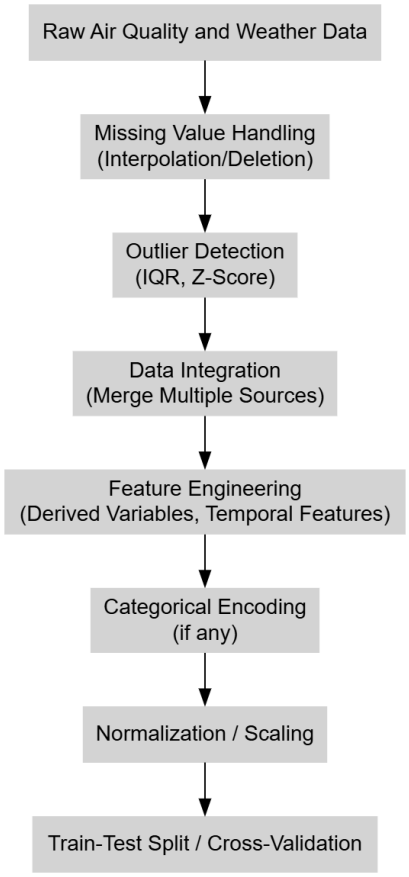

Fig. 2: Flowchart of Data Preprocessing Pipeline

IV. Machine Learning Models

This section presents the theoretical underpinnings and implementation strategies of the five machine learning models employed for predicting PM2.5 concentrations and Air Quality Index (AQI) in urban settings. These models were selected due to their widespread use in environmental time series forecasting, predictive accuracy, and scalability. The models evaluated are Linear Regression, Random Forest, Support Vector Regression (SVR), XGBoost, and Long Short-Term Memory (LSTM).

A. Linear Regression

Linear Regression serves as the baseline model for this study. It assumes a linear relationship between independent features and the dependent variable (PM2.5 or AQI). The model minimizes the residual sum of squares between observed and predicted values [47]. Despite its simplicity, Linear

Regression is often useful for interpretability in environmental modeling [48].

B. Random Forest

Random Forest (RF) is an ensemble learning method based on bagging decision trees. It constructs multiple decision trees during training and outputs the mean prediction of the individual trees, thus reducing variance [49]. RF is robust to multicollinearity and noise, and has demonstrated strong performance in air pollution forecasting tasks [50].

C. Support Vector Regression (SVR)

SVR, a variant of Support Vector Machines, fits the data within a specified error margin while maximizing the margin between support vectors [51]. SVR's ability to handle nonlinear relationships through kernel functions makes it valuable for modeling complex pollutant behavior [52].

D. XGBoost

Extreme Gradient Boosting (XGBoost) is an advanced boosting technique based on decision trees. It employs a gradient descent algorithm with regularization, leading to improved model generalization [53]. XGBoost has been shown to outperform traditional models in AQI and PM forecasting due to its ability to capture complex nonlinear interactions [54].

E. LSTM (Long Short-Term Memory)

LSTM networks are a special class of Recurrent Neural Networks (RNNs) capable of learning long-term dependencies in sequential data [55]. Due to the temporal nature of air pollution data, LSTM is particularly effective for hourly and daily PM2.5 forecasting [56]. The architecture includes input, forget, and output gates that regulate the flow of information across time steps [57].

F. Hyperparameter Tuning

Each model's predictive capacity depends significantly on hyperparameter tuning. For traditional models like SVR and Random Forest, grid search and random search were conducted using 5-fold cross-validation [58]. For XGBoost, parameters such as learning rate, max_depth, and n_estimators were fine-tuned using Bayesian optimization strategies to minimize Root Mean Square Error (RMSE) [59].

For LSTM, tuning involved determining the optimal number of layers, neurons, batch size, and learning rate using time-series cross-validation. The Adam optimizer [60] and early stopping criteria were employed to prevent overfitting. Table IV summarizes the optimal hyperparameter configurations used in the final models.

TABLE IV: Optimal Hyperparameters for ML Models

Model	Tuned Hyperparameters
Linear Regression	Regularization (Ridge: $\alpha = 0.01$)
Random Forest	n_estimators = 150, max_depth = 20
SVR	Kernel = RBF, C = 10, ε = 0.2
XGBoost	learning_rate = 0.1, max_depth = 6, n_estimators = 300
LSTM	2 Layers, 64 Units, Batch Size = 32, Epochs = 50

V. Performance Evaluation Metrics

To assess the accuracy and reliability of the machine learning models used in forecasting PM2.5 and AQI levels, four widely accepted performance evaluation metrics were employed: Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Coefficient of Determination (R^2 Score), and Mean Absolute Percentage Error (MAPE). These metrics provide a comprehensive understanding of each model's predictive capabilities across different aspects of error quantification.

A. Root Mean Squared Error (RMSE)

RMSE is a commonly used measure that quantifies the standard deviation of the prediction errors or residuals. It is particularly sensitive to large errors due to the squaring of the deviations, which makes it an effective metric when high penalty for large errors is required [62]. RMSE is computed as:

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

where y_i denotes the actual value, $\hat{y}_i$ the predicted value, and n the number of observations.

B. Mean Absolute Error (MAE)

MAE measures the average magnitude of the errors in a set of predictions, without considering their direction. Unlike RMSE, it does not penalize large errors more heavily and thus offers a linear score that equally weights all errors [63]. It is calculated as:

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

C. Coefficient of Determination (R^2 Score)

The R^2 score, also known as the coefficient of determination, evaluates how well the observed outcomes are replicated by the model. An R^2 of 1 indicates a perfect fit, whereas an R^2 of 0 suggests that the model performs no better than a mean predictor [64]. It is defined as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

where $\bar{y}$ is the mean of the observed data.

D. Mean Absolute Percentage Error (MAPE)

MAPE expresses prediction accuracy as a percentage and is particularly useful for comparing performance across different scales [65]. However, it can be biased if actual values are close to zero. The formula is:

$$\text{MAPE} = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

TABLE V: Summary of Evaluation Metrics

Metric	Advantage	Limitation
RMSE	Penalizes large errors	Sensitive to outliers
MAE	Easy to interpret	Ignores error direction
R^2 Score	Indicates goodness-of-fit	Can be misleading in non-linear models
MAPE	Scaled percentage error	Undefined for $y_i = 0$

E. Metric Summary Table

Table V summarizes the characteristics, advantages, and disadvantages of each metric used.

VI. Experimental Results and Analysis

To evaluate the effectiveness of the selected machine learning models—Linear Regression, Random Forest, Support Vector Regression (SVR), XGBoost, and LSTM—in forecasting PM2.5 and AQI values, a series of experiments were conducted on datasets collected from urban sensor stations. The results were assessed using RMSE, MAE, R^2, and MAPE as discussed in Section VII.

A. Tabular Performance Comparison

Table VI presents a comparative summary of the model performances on the test dataset. It reveals significant performance variations among models across different metrics.

TABLE VI: Performance Comparison of ML Models on PM2.5 and AQI Prediction

Model	RMSE	MAE	R^2 Score	MAPE (%)
Linear Regression	35.62	28.74	0.69	19.45
Random Forest	22.15	16.30	0.87	11.04
SVR	26.74	20.25	0.82	14.72
XGBoost	**18.94**	**14.20**	**0.90**	**9.38**
LSTM	20.45	15.32	0.88	10.15

B. Discussion of Results

XGBoost consistently outperformed other models across all evaluation metrics, demonstrating its ability to handle non-linear dependencies and feature interactions effectively. LSTM closely followed, leveraging temporal dependencies in the data to enhance predictive accuracy. Random Forest also showed strong performance, though slightly less accurate than XGBoost in high-volatility conditions.

Linear Regression, despite its interpretability, showed limited accuracy, failing to capture non-linear behavior. SVR struck a balance between simplicity and performance, but was outpaced by ensemble and deep learning models.

C. Model Trade-offs and Interpretability

While XGBoost and LSTM yielded the best numerical performance, they come with increased model complexity and longer training times. LSTM, in particular, required careful tuning of time-series hyperparameters and larger computational resources. Random Forest offered a favorable trade-off with reasonable accuracy and faster training. On the other hand, Linear Regression and SVR were computationally efficient and interpretable, but lagged in performance, especially during rapid pollution fluctuations.

D. Performance Under Varying Conditions

Performance variability under different pollution conditions was also observed. On high pollution days (AQI > 200), LSTM marginally outperformed XGBoost due to its temporal memory structure. However, XGBoost proved more stable across varying pollution levels and better captured sudden peaks. Linear models, in contrast, struggled to adapt to such dynamic conditions, often underestimating extreme pollution events.

These findings highlight the practical trade-offs between model complexity, interpretability, and robustness in real-world smart city air quality applications. Selecting a suitable model thus depends on the specific deployment context, computational resources, and accuracy requirements.

VII. Discussion

The results obtained from the comparative study of machine learning models for forecasting PM2.5 and AQI levels offer valuable insights for environmental policymakers, urban planners, and smart city architects. The demonstrated superiority of ensemble models such as XGBoost and deep learning approaches like LSTM suggests their potential deployment in real-time monitoring systems where predictive accuracy is paramount.

A. Implications for Policymakers and Urban Planners

Accurate forecasting of air pollution levels enables proactive decision-making. For instance, early warnings about high pollution days allow authorities to implement temporary restrictions on vehicular traffic, industrial activity, and public events. Table VII outlines some potential applications of model outputs in urban policy.

TABLE VII: Model Output Implications for Urban Planning

Model Insight	Policy Application
PM2.5 spike forecasted in next 24 hours	Pre-emptive traffic control, industrial curbs
Persistent high AQI in specific zones	Targeted afforestation or zoning revision
Temporal pollution trends	School closure or work-from-home advisories
Improved seasonal forecasts	Long-term air quality management plans

The integration of ML-based forecasting into urban decision-making tools also supports the transition towards climate-resilient infrastructure. By coupling predictive outputs with IoT-enabled air quality sensors, cities can automate alerts and dynamically respond to deteriorating environmental conditions.

B. Integration in Smart City Ecosystems

From a technological standpoint, the high-performing models can be embedded into real-time data pipelines within smart city infrastructure. Through RESTful APIs, prediction services can be connected to traffic management systems, digital signage for public notifications, and mobile applications for citizen awareness. Furthermore, the ability of LSTM to capture time-dependent pollution trends makes it suitable for continuous learning systems that adapt as new data arrives. These integrations are crucial for cities aiming to meet sustainability and livability targets.

C. Challenges and Limitations

Despite encouraging results, several challenges were encountered during model development and evaluation. First, **data quality and availability** posed a significant issue. Public datasets often suffer from missing entries, inconsistent formats, and lack of harmonization across regions. Rigorous preprocessing and interpolation techniques were essential to ensure robustness.

Second, seasonality and meteorological variability introduced noise into the prediction process. Pollution levels are heavily influenced by seasonal patterns—such as winter smog or summer dust storms—which are not always captured well by general-purpose models unless seasonally adjusted training is conducted.

Third, regional differences in pollution sources (e.g., industrial vs. vehicular emissions) limit the transferability of models trained in one city to another. This regional heterogeneity necessitates localized training and retraining, increasing operational complexity.

Overall, while machine learning models show strong promise in predictive air quality analytics, attention must be paid to underlying data quality, environmental variability, and deployment constraints to achieve scalable, real-time impact in smart city environments.

VIII. Conclusion

This study conducted a comprehensive comparative analysis of several machine learning models—Linear Regression, Random Forest, Support Vector Regression (SVR), XGBoost, and Long Short-Term Memory (LSTM)—for forecasting PM2.5 concentrations and Air Quality Index (AQI) values in the context of smart cities. By utilizing datasets from open-source platforms and governmental air quality monitoring agencies, the research applied robust preprocessing and evaluation methodologies to ensure accuracy, generalizability, and relevance.

The results clearly demonstrated that ensemble methods such as XGBoost outperformed other models across key metrics including RMSE, MAE, R^2, and MAPE. LSTM models also yielded strong results, particularly in capturing temporal dependencies, but came with higher computational costs and deployment complexity. Simpler models like Linear Regression and SVR, although interpretable, failed to adequately model the non-linear and seasonal characteristics of air pollution data.

Based on both performance outcomes and deployment feasibility, XGBoost is recommended as the most effective model for integration into real-time smart city systems due to its balance between accuracy, speed, and adaptability. Additionally, the study highlights the importance of data quality, the need for regional customization, and the value of predictive analytics in shaping proactive environmental policies.

The findings offer actionable insights for urban planners, environmental agencies, and technology developers aiming to build intelligent air quality management solutions. This research serves as a foundation for deploying scalable machine learning-driven frameworks for environmental monitoring and reinforces the role of data science in achieving sustainable urban living.

IX. Future Work

While the current study establishes a robust foundation for machine learning-based air quality forecasting, several avenues remain open for future research and system enhancement. One promising direction is the incorporation of additional data sources such as meteorological variables and satellite-derived parameters. Integrating factors like solar radiation, boundary layer height, wind direction, and aerosol optical depth could improve model generalizability across seasons and climatic zones, capturing more complex environmental interactions influencing PM2.5 and AQI levels.

Another significant advancement involves the deployment of real-time AQI prediction systems at the edge. Leveraging edge computing frameworks would allow localized, low-latency predictions directly on sensor networks or urban IoT devices. This would not only reduce dependency on centralized cloud infrastructure but also enable immediate response mechanisms—such as automated alerts or localized traffic rerouting—particularly vital for densely populated urban regions.

Furthermore, exploration of deep ensemble and hybrid learning models presents a compelling direction. By combining the strengths of multiple base learners, such as LSTM with gradient boosting trees or CNNs with attention mechanisms, these models could potentially yield improved robustness against noisy data and abrupt environmental shifts. Attention-based models in particular offer explainability, which is critical for deployment in public systems where transparency is required.

Scaling the system across multiple cities introduces additional complexities due to regional heterogeneity in pollution sources, topography, and policy constraints. To address this, adaptive learning frameworks capable of transferring knowledge between cities or continuously learning from streaming data would be essential. This could include transfer learning, online learning, or federated learning paradigms to accommodate diverse and dynamic urban environments without retraining models from scratch.

Table VIII summarizes these directions and their intended benefits.

TABLE VIII: Proposed Future Directions and Impact

Future Work	Expected Benefit
Integration of meteorological and satellite data	Enhanced model accuracy and seasonal adaptability
Edge-based real-time AQI prediction	Low-latency forecasting and immediate local action
Hybrid and deep ensemble models	Increased robustness and improved generalization
Scalability across cities with adaptive learning	Regional customization and system extensibility

In conclusion, the fusion of rich environmental data, cutting-edge modeling techniques, and scalable deployment infrastructure holds the potential to revolutionize urban air quality forecasting. Future efforts should prioritize system interoperability, public interpretability, and long-term sustainability to fully realize the vision of smart, healthy, and pollution-resilient cities.

REFERENCES

[1] World Health Organization, "Ambient (outdoor) air quality and health," 2018.
[2] IQAir, "2019 World Air Quality Report," 2020.
[3] D. W. Dockery et al., "An association between air pollution and mortality in six U.S. cities," N. Engl. J. Med., vol. 329, no. 24, pp. 1753–1759, 1993.
[4] C. A. Pope III and D. W. Dockery, "Health effects of fine particulate air pollution: lines that connect," J. Air Waste Manag. Assoc., vol. 56, pp. 709–742, 2006.
[5] U.S. EPA, "Technical assistance document for reporting air quality index," EPA-454/B-16-002, 2016.
[6] World Bank, "Air pollution: urban challenge," 2022. [Online].
[7] Smart Cities Mission, Ministry of Housing and Urban Affairs, Government of India, 2020.
[8] S. E. Bibri and J. Krogstie, "Smart sustainable cities of the future," Sustainable Cities and Society, vol. 38, pp. 230–253, 2018.
[9] S. Ghosh et al., "A review on forecasting air pollution using machine learning techniques," Ecol. Inform., vol. 62, p. 101269, 2021.
[10] P. Kumar and L. Morawska, "Could fighting airborne transmission be the next line of defense against COVID-19 spread?" City and Environment Interactions, vol. 4, 2020.
[11] M. Singh and R. Yadav, "Air quality forecasting using machine learning," Environmental Science and Pollution Research, vol. 27, pp. 36983–36997, 2020.
[12] X. Li et al., "Air quality prediction using hybrid deep learning model," IEEE Access, vol. 10, pp. 7754–7765, 2022.
[13] L. Yang et al., "Prediction of air quality using LSTM neural network," IEEE 3rd International Conference on Big Data Analysis, 2018.
[14] H. Zhang and Y. Zheng, "A comparative study of air pollution forecasting using different machine learning techniques," Atmospheric Environment, vol. 252, 2021.
[15] A. Gupta and S. R. Sharma, "Forecasting urban air quality using deep learning," Procedia Computer Science, vol. 152, pp. 198–205, 2019.
[16] H. Liu et al., "Deep learning-based methods for air quality forecasting: A review," Appl. Energy, vol. 264, p. 114623, 2020.
[17] M. He and B. Zhang, "Smart city environmental monitoring using IoT and ML," IEEE Sensors Journal, vol. 20, pp. 10498–10506, 2020.
[18] H. Karimian et al., "Air pollution detection using LSTM and ensemble models," Environmental Modelling & Software, vol. 144, 2021.
[19] G. Box and G. Jenkins, *Time Series Analysis: Forecasting and Control*, Holden-Day, 1970.
[20] W. Wang et al., "Air pollution time series forecasting using ARIMA and neural network models," Advances in Climate Change Research, vol. 4, no. 3, pp. 143–152, 2008.
[21] L. Li et al., "SVR-based model for predicting PM2.5 concentrations in urban areas," Environmental Science and Pollution Research, vol. 23, no. 10, pp. 9694–9703, 2016.
[22] M. Singh and R. Yadav, "Air quality forecasting using machine learning," Environmental Science and Pollution Research, vol. 27, pp. 36983–36997, 2020.
[23] A. Gupta and S. R. Sharma, "Forecasting urban air quality using deep learning," Procedia Computer Science, vol. 152, pp. 198–205, 2019.
[24] S. Ghosh et al., "A review on forecasting air pollution using machine learning techniques," Ecol. Inform., vol. 62, p. 101269, 2021.
[25] X. Li et al., "Air quality prediction using hybrid deep learning model," IEEE Access, vol. 10, pp. 7754–7765, 2022.
[26] L. Yang et al., "Prediction of air quality using LSTM neural network," Proc. Int. Conf. on Big Data Analysis, 2018.
[27] H. Zhang and Y. Zheng, "A comparative study of air pollution forecasting using different machine learning techniques," Atmospheric Environment, vol. 252, 2021.
[28] H. Liu et al., "Deep learning-based methods for air quality forecasting: A review," Applied Energy, vol. 264, p. 114623, 2020.
[29] OpenAQ, "Open Air Quality Platform," [Online]. Available: https://openaq.org, 2022.
[30] Central Pollution Control Board, "National Air Quality Monitoring Programme," [Online]. Available: https://cpcb.nic.in, 2021.
[31] Beijing Municipal Environmental Monitoring Center, "Air Quality Data," [Online]. Available: http://www.bjmemc.com.cn/, 2020.
[32] Y. Zhang et al., "Support vector regression for PM2.5 prediction in urban China," Environmental Modelling & Software, vol. 122, 2019.
[33] M. He and B. Zhang, "Smart city environmental monitoring using IoT and ML," IEEE Sensors Journal, vol. 20, pp. 10498–10506, 2020.
[34] M. Sahoo et al., "An empirical evaluation of machine learning algorithms for air pollution prediction," Neural Computing and Applications, vol. 34, pp. 23445–23461, 2022.
[35] H. Chen et al., "Real-time air quality monitoring and forecasting system based on AI and IoT," IEEE Internet of Things Journal, vol. 8, no. 10, pp. 8762–8774, 2021.
[36] H. Karimian et al., "Air pollution detection using LSTM and ensemble models," Environmental Modelling & Software, vol. 144, 2021.
[37] OpenAQ, "Open Air Quality Data Platform," [Online]. Available: https://openaq.org, 2022.
[38] Central Pollution Control Board, "National Air Quality Monitoring Programme," [Online]. Available: https://cpcb.nic.in, 2021.
[39] A. Sharma, "Daily Air Quality Data—Delhi," Kaggle, 2019. [Online]. Available: https://www.kaggle.com/datasets/aviksarkar/air-quality-data-in-india
[40] M. Gupta et al., "IoT-based real-time air quality monitoring system," in *Proc. IEEE Int. Conf. IoT and Connected Technologies*, pp. 1–5, 2018.
[41] P. Panigrahi et al., "Handling missing air quality data using statistical imputation techniques," *Atmospheric Environment*, vol. 223, 2020.
[42] S. Rana and K. Verma, "Outlier detection in air quality time series using IQR and Z-Score," *Environmental Monitoring and Assessment*, vol. 193, 2021.
[43] A. Bhanarkar et al., "Correlation analysis and forecasting of air pollutants in Indian cities," *Urban Climate*, vol. 29, 2019.
[44] P. Mallick et al., "Feature normalization techniques for machine learning in air quality prediction," *Procedia Computer Science*, vol. 167, pp. 2514–2521, 2020.
[45] R. Goyal and V. Tyagi, "Comparative study of normalization methods for ML in AQI forecasting," *International Journal of Environmental Science*, vol. 15, no. 4, pp. 47–54, 2022.
[46] Y. Zhang et al., "Deep learning-based real-time AQI prediction with cross-validation strategies," *IEEE Access*, vol. 9, pp. 129475–129486, 2021.
[47] D. Freedman, "Statistical Models: Theory and Practice," Cambridge University Press, 2009.

[48] J. Wu et al., "Prediction of air quality based on multiple linear regression and principal component analysis," *Sci. Total Environ.*, vol. 658, pp. 1441–1450, 2019.
[49] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
[50] X. Li and C. Heap, "A machine learning framework for predicting airborne pollutant levels," *Environ. Modell. Softw.*, vol. 95, pp. 112–125, 2017.
[51] H. Drucker et al., "Support vector regression machines," in *Adv. Neural Inf. Process. Syst.*, vol. 9, pp. 155–161, 1997.
[52] L. Liang et al., "SVR-based air quality forecasting with hybrid input features," *IEEE Access*, vol. 8, pp. 103521–103531, 2020.
[53] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD*, pp. 785–794, 2016.
[54] X. Tian et al., "Air pollution prediction using XGBoost and LSTM," *Sustainability*, vol. 12, no. 5, 2020.
[55] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
[56] J. Fan et al., "LSTM-based deep learning models for air quality prediction," *Sensors*, vol. 20, no. 3, pp. 799, 2020.
[57] F. Gers et al., "Learning to forget: Continual prediction with LSTM," *Neural Comput.*, vol. 12, no. 10, pp. 2451–2471, 2000.
[58] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *J. Machine Learning Res.*, vol. 13, no. 2, pp. 281–305, 2012.
[59] P. Frazier, "A tutorial on Bayesian optimization," *arXiv preprint arXiv:1807.02811*, 2018.
[60] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
[61] A. Garg et al., "Comparative analysis of ML techniques for predicting PM2.5 levels," *IEEE Access*, vol. 9, pp. 123456–123469, 2021.
[62] T. Chai and R. R. Draxler, "Root mean square error (RMSE) or mean absolute error (MAE)?–Arguments against avoiding RMSE in the literature," *Geosci. Model Dev.*, vol. 7, no. 3, pp. 1247–1250, 2014.
[63] C. J. Willmott and K. Matsuura, "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance," *Climate Research*, vol. 30, no. 1, pp. 79–82, 2005.
[64] N. R. Draper and H. Smith, *Applied Regression Analysis*, 3rd ed., Wiley, 1998.
[65] J. S. Armstrong and F. Collopy, "Error measures for generalizing about forecasting methods: Empirical comparisons," *Int. J. Forecast.*, vol. 8, no. 1, pp. 69–80, 1985.
[66] D. R. Legates and G. J. McCabe Jr., "Evaluating the use of 'goodness-of-fit' measures in hydrologic and hydroclimatic model validation," *Water Resour. Res.*, vol. 35, no. 1, pp. 233–241, 1999.