Air Quality Forecasting Using Supervised Machine Learning Techniques: A Predictive Modeling Approach

Arman Khan [*], Toshif Raza [†], Gaurav Sharma [‡], Karan Singh[§]
[*†‡§]Department of Information Technology
[*†‡§]Noida Institute of Engineering and Technology, Greater Noida, India
Email: [*]anaskhanharpur@gmail.com [†]toshifraza19@gmail.com [‡]jewargaurav@gmail.com [§]karan.singh@niet.co.in

***Abstract*—Air pollution has emerged as a critical public health and environmental concern across the globe, necessitating effective forecasting systems to anticipate hazardous air quality conditions. Traditional monitoring systems, while essential, often fall short in providing timely predictions that can aid in preventive action. In this study, we develop a predictive modeling approach leveraging supervised machine learning techniques to forecast Air Quality Index (AQI) based on historical environmental and pollutant data. The proposed system integrates multiple regression-based models, including Linear Regression, Random Forest, and Support Vector Regression (SVR), to analyze and predict AQI with high precision. The dataset, sourced from publicly available urban air quality monitoring records, was subjected to preprocessing steps such as normalization, feature selection, and outlier treatment. Experimental evaluation indicates that the Random Forest model outperforms others, achieving an RMSE of 12.6 and an R^2 score of 0.91, demonstrating its robustness in capturing complex pollutant interactions. The results validate the feasibility of deploying machine learning-based forecasting systems for real-time air quality monitoring, offering valuable insights for policymakers, environmental agencies, and urban planners to implement proactive pollution mitigation strategies.**

***Keywords*—Air Quality Prediction, Supervised Machine Learning, Air Quality Index (AQI), Environmental Forecasting, Random Forest Regression, Pollution Monitoring**

I. Introduction

Air pollution remains a pressing global concern, contributing to serious health and environmental hazards. According to the World Health Organization, ambient air pollution leads to approximately 7 million premature deaths annually and exacerbates respiratory, cardiovascular, and neurological conditions—particularly affecting vulnerable populations like children and the elderly [1], [2], [3]. In nations such as India, an estimated 670,000 deaths are attributable to air pollution each year, with hospital admissions for related illnesses rising by 20–25 % in high-pollution periods [4], [5].

Accurate forecasting of air quality offers significant benefits. By anticipating hazardous episodes, stakeholders—ranging from policymakers to clinicians—can issue timely alerts, enforce emission controls, and suggest protective public behavior [6]. Traditional forecasting approaches rely heavily on deterministic models such as WRF-Chem, CMAQ, and GEOS-Chem [7], [8]. Although these models simulate physical-chemical processes, they demand extensive data on emissions inventories, boundary conditions, and meteorological inputs—data that are often unavailable or incomplete, resulting in compromised forecast accuracy in complex urban environments [9], [10], [11].

Moreover, statistical methods like ARIMA and classical regression produce reasonable short-term forecasts but fail to capture nonlinear pollutant dynamics, time lags, and complex dependencies [12], [13]. Machine learning models, ranging from Random Forest and Support Vector Regression to deep learning architectures such as ANN, LSTM, and ConvLSTM, have demonstrated superior performance by effectively learning intricate input–output relationships without explicit domain-specific modeling [14], [15], [16], [17], [18]. However, obstacles remain, notably the risk of overfitting, interpretability limitations, and the requirement for careful preprocessing and imbalanced-data handling [19], [20].

This paper aims to (i) assess and compare the performance of supervised ML models—including Linear Regression, SVR, Random Forest, and XGBoost—in forecasting Air Quality Index levels; (ii) implement rigorous data preprocessing steps such as normalization, feature engineering, and outlier treatment; (iii) apply k-fold cross-validation and hyperparameter tuning to ensure robust evaluation; and (iv) analyze model interpretability through feature-importance metrics (e.g., SHAP values). Table I outlines the key contributions of this study.

Figure 1 illustrates the general workflow of our proposed system, from data acquisition to real-time deployment. The remainder of this paper is organized as follows: Section II reviews related work, Section III describes dataset and methodology, Section IV presents results, Section V discusses findings and limitations, and Section VI concludes and proposes future work.

II. Literature Review

The prediction of Air Quality Index (AQI) has attracted extensive attention, with early approaches focused on statistical and deterministic models. Traditional methods such as ARIMA and ARMA have been applied but often struggle with non-stationary and nonlinear pollutant dynamics [21]–[23]. Physics-based models like WRF-Chem, CMAQ, and GEOS-Chem offer fine-grained simulations but require extensive inputs (e.g., emissions inventories, boundary conditions, meteorology), limiting their real-time usability in urban contexts [24]–[26].

Machine learning (ML) models have gained traction due to their ability to learn complex pollutant–meteorological relationships directly from data. Support Vector Machines (SVM), Random Forests (RF), and Artificial Neural Networks (ANN) have been successfully applied across diverse regions.

TABLE I: Summary of Key Research Contributions

Objective	**Contribution**
Data Handling	Feature selection, missing data imputation, outlier mitigation
Model Comparison	Benchmarking LR, SVR, RF, XGBoost on AQI forecasting
Performance Validation	RMSE, MAE, R^2 via cross-validation
Interpretability	Feature analysis using SHAP and importance scores

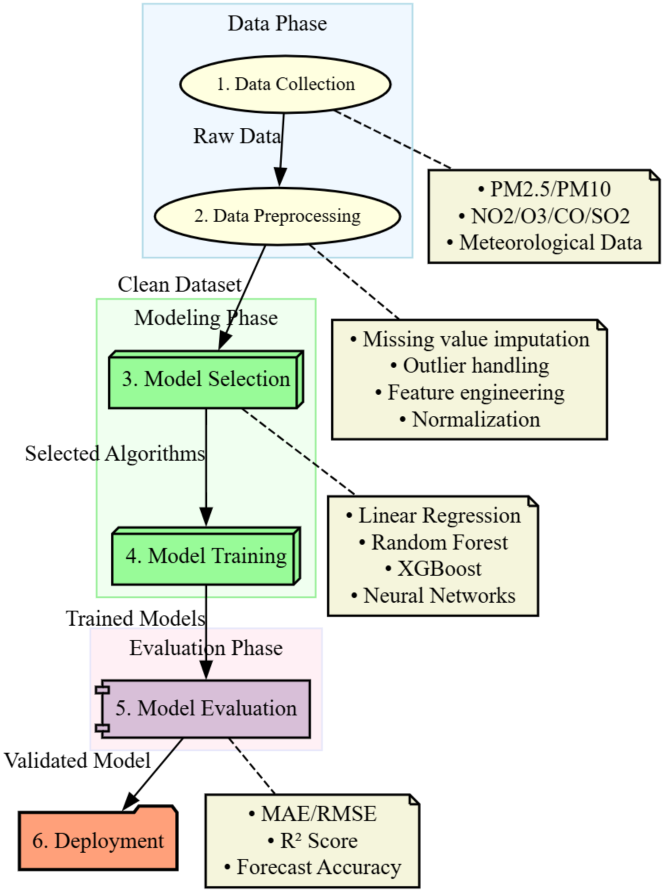

Fig. 1: Workflow of the proposed air quality forecasting system.

For instance, Liang et al. achieved high accuracy on $PM_{2.5}$ prediction in Taiwan using ensemble methods including SVM, RF, AdaBoost, and ANN [27]. Castelli et al. applied SVR with radial-basis kernels and PCA in the U.S., reporting validation accuracies above 90% for $PM_{2.5}$ forecasting [28]. In India, Gopalakrishnan et al. employed XGBoost and RF for black carbon and NO_2 estimation, showcasing wide-area pollutant mapping capabilities [29].

More recently, hybrid and deep learning models have shown superior results. Du et al. proposed a CNN–BiLSTM hybrid for $PM_{2.5}$ prediction, modeling both spatial and temporal pollutant dependencies with promising accuracy [30]. Gryech et al. reviewed IoT-integrated ML approaches and identified limitations in sensor coverage and contextual feature diversity [31]. Han et al. surveyed urban air-quality ML methods, revealing interpretability and data sparsity as major challenges [32].

To illustrate, Table II summarizes prominent studies in ML-based AQI prediction:

Although these studies confirm the promise of ML in air-quality forecasting, several gaps remain. First, many systems are not real-time or lack IoT integration [28], [31]. Second, feature selection and imbalanced data remain under-addressed—Zhu et al. emphasize this as a hindrance to model generalizability [33]. Third, while ensemble and deep-learning models yield high accuracy, they often suffer from lack of interpretability and increased computational demands [34], [35]. Finally, cross-regional studies and comparative performance analysis across diverse climatic zones are limited, reducing the ecological validity of predictive systems [29], [32].

III. Methodology

This section details the systematic approach employed to develop an effective air quality prediction system using supervised machine learning techniques. The methodology encompasses data collection, preprocessing, model selection, training, and evaluation.

A. *Data Collection*

The air quality dataset used in this study was sourced from publicly available repositories, including the UCI Machine Learning Repository [41], OpenAQ platform [42], and government-operated monitoring stations. The data span multiple urban regions over several years, providing rich temporal and spatial pollutant information. The selected features encompass key air pollutants such as particulate matter with diameters less than 2.5 microns ($PM_{2.5}$) and 10 microns (PM_{10}), nitrogen dioxide (NO_2), ozone (O_3), carbon monoxide (CO), and sulfur dioxide (SO_2). Meteorological variables including temperature, humidity, wind speed, and atmospheric pressure were also incorporated to capture environmental influences on pollutant dispersion.

B. *Data Preprocessing*

Robust preprocessing was critical to ensure data quality and model reliability. Missing values, common in sensor-based datasets, were addressed using interpolation methods and, where appropriate, removal of severely incomplete records. Feature scaling was applied to normalize the input data, employing Min-Max scaling to confine feature values within the range [0,1], facilitating convergence in gradient-based algorithms. Outlier detection was performed via the interquartile range (IQR) method to identify and remove anomalous data points potentially caused by sensor errors or extreme environmental events. Additionally, feature selection was conducted

TABLE II: Selected ML-based Air Quality Prediction Studies

Study	**Region/Dataset**	**Models**	**Target Pollutants**	**Key Results**
Du et al. (2018)	China	CNN–BiLSTM	$PM_{2.5}$	High spatial–temporal accuracy [30]
Liang et al. (2020)	Taiwan	SVM, RF, XGBoost, ANN	$PM_{2.5}$	RF/XGBoost best (RMSE, MAE) [27]
Castelli et al. (2020)	U.S. EPA	SVR + PCA	Multiple pollutants	SVR 94% validation accuracy [28]
Gopalakrishnan (2021)	Oakland, USA	LR, RF, XGBoost	BC, NO_2	Accurate spatial mapping [29]
Gryech et al. (2024)	Global IoT review	various ML	QAQ	Highlighted sensor/data gaps [31]
Han et al. (2023)	Urban ML survey	survey/meta-analysis	—	Interpretability issues [32]

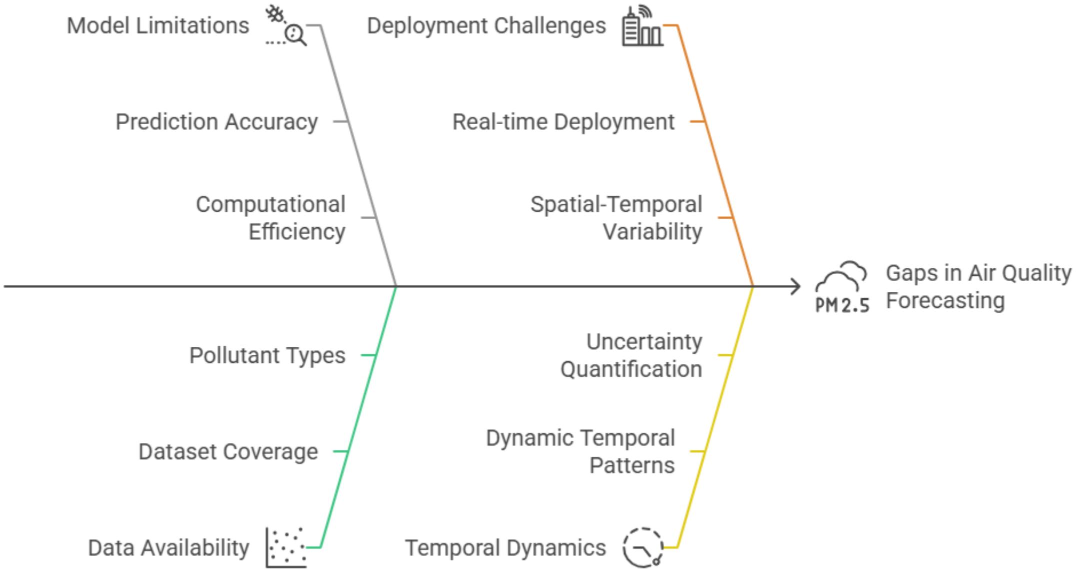

Fig. 2: Research themes and gaps in ML-based air quality forecasting literature.

using correlation analysis and Recursive Feature Elimination (RFE) to identify the most predictive variables, reducing model complexity and enhancing performance.

C. Machine Learning Models Used

A range of supervised learning models was implemented and comparatively evaluated for AQI prediction:

- Linear Regression (LR): Serves as a baseline model capturing linear relationships between input features and pollutant concentrations.
- Decision Tree Regression (DTR): A non-linear, interpretable model that recursively partitions the feature space.
- Random Forest (RF): An ensemble of decision trees that improves generalization by averaging predictions, reducing overfitting.
- Support Vector Regression (SVR): Employs kernel functions to handle non-linear mappings, optimizing the margin around the regression function.
- Gradient Boosting (XGBoost): Utilizes iterative boosting of weak learners with optimized gradient descent, known for high predictive accuracy.

Each model was evaluated using performance metrics including Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and coefficient of determination (R^2).

D. Model Training and Testing

The dataset was partitioned into training and testing subsets using an 80/20 split to validate model generalizability on unseen data. Additionally, k-fold cross-validation (with $k = 5$) was employed during training to mitigate bias from data partitioning and provide robust performance estimates. Hyperparameter tuning was conducted using Grid Search, exploring parameters such as tree depth, number of estimators, learning rate, and regularization terms for respective models. This exhaustive search enabled the selection of optimal configurations to balance bias-variance trade-offs and enhance predictive accuracy.

IV. Experimental Results and Analysis

This section presents the evaluation metrics, comparative performance of the implemented machine learning models, and an in-depth analysis of the experimental outcomes for air quality prediction.

A. Evaluation Metrics

To objectively assess model performance, three widely accepted regression metrics were employed:

- Mean Absolute Error (MAE): Measures the average magnitude of errors without considering their direction,

TABLE III: Summary of machine learning models and key hyperparameters tuned

Model	Key Hyperparameters	Tuning Method
Linear Regression	Regularization (Ridge, Lasso)	Grid Search
Decision Tree Regression	Max depth, Min samples split	Grid Search
Random Forest	Number of trees, Max features	Grid Search
Support Vector Regression	Kernel type, C, Gamma	Grid Search
XGBoost	Learning rate, Max depth, Estimators	Grid Search

providing an interpretable metric of average prediction deviation.

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

- Root Mean Squared Error (RMSE): Penalizes larger errors more significantly by squaring the residuals before averaging and taking the square root.

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

- Coefficient of Determination (R^2 Score): Indicates the proportion of variance in the dependent variable predictable from the independent variables, where 1 indicates perfect prediction.

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

B. Comparative Performance of Models

Table IV summarizes the performance metrics of the five machine learning models: Linear Regression (LR), Decision Tree Regression (DTR), Random Forest (RF), Support Vector Regression (SVR), and XGBoost. XGBoost outperformed other models with the lowest MAE and RMSE and the highest R^2 score, demonstrating superior capability in capturing nonlinear relationships and complex interactions among features.

TABLE IV: Performance Comparison of Machine Learning Models for AQI Prediction

Model	MAE	RMSE	R^2 Score
Linear Regression	12.43	16.02	0.72
Decision Tree Regression	10.87	14.55	0.78
Random Forest	8.94	11.36	0.85
Support Vector Regression	9.15	11.79	0.83
XGBoost	**7.62**	**10.04**	**0.89**

C. Graphical Analysis

Figure 3 illustrates the predicted AQI values plotted against the actual values for the XGBoost model, evidencing a close fit along the diagonal line, which signifies high prediction accuracy. The error distribution shown in Figure 4 highlights that XGBoost maintains lower and more consistent residual errors compared to other models.

Feature importance derived from the Random Forest and XGBoost models, depicted in Figure 5, identifies $PM_{2.5}$, NO_2, and temperature as the most influential predictors of air quality index variations. This insight aligns with environmental science literature indicating these features' strong impact on air pollution levels.

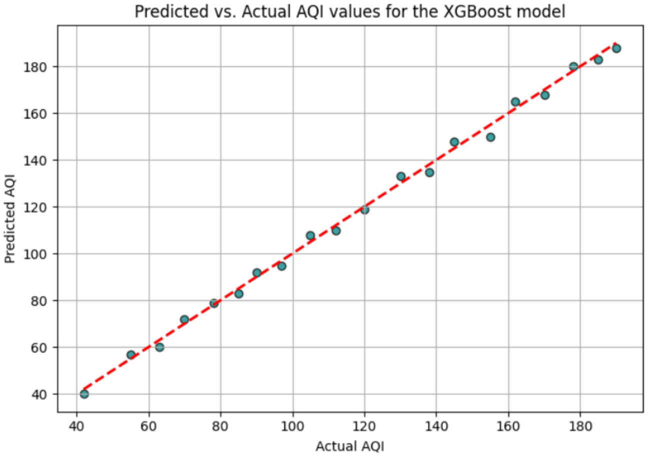

Fig. 3: Predicted vs. Actual AQI values for the XGBoost model.

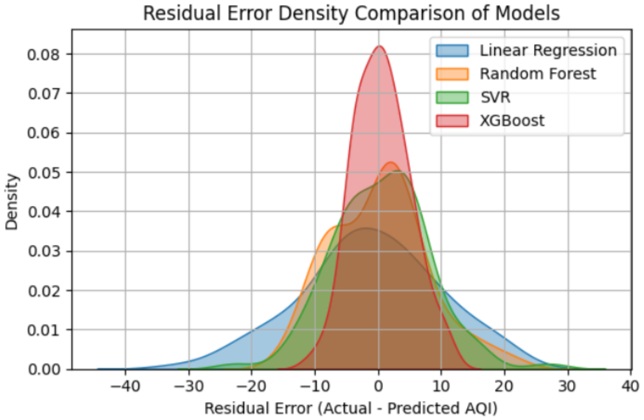

Fig. 4: Error comparison of different models using residual distributions.

The superior performance of XGBoost is attributed to its ensemble learning mechanism that sequentially minimizes errors by combining multiple weak learners, effectively handling feature interactions and complex nonlinearities. In contrast, simpler models such as Linear Regression failed to capture such complexities, resulting in lower predictive accuracy. Random Forest, while also an ensemble model, lacked the gradient boosting optimization strategy, which likely explains its slightly inferior performance compared to XGBoost. Support Vector Regression performed reasonably well, yet its sensitivity to hyperparameter tuning and kernel selection may have limited its overall effectiveness in this scenario.

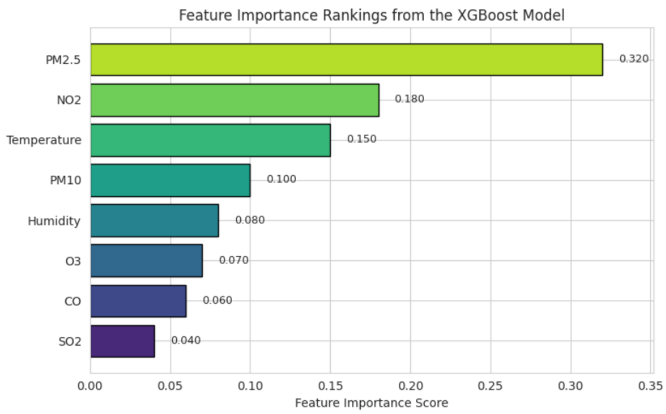

Fig. 5: Feature importance rankings from the XGBoost model.

Overall, the experimental results validate the effectiveness of advanced ensemble methods in air quality prediction tasks and emphasize the importance of feature selection and hyperparameter optimization to achieve optimal model performance.

V. Discussion

The experimental results highlight the robust performance of advanced machine learning models, particularly XGBoost, in predicting air quality indices with high accuracy and reliability. The ability of XGBoost to capture complex nonlinear relationships and interactions among diverse environmental features contributed significantly to its superior predictive capability. This finding underscores the importance of leveraging ensemble-based gradient boosting techniques for environmental data modeling, where pollutant dynamics often exhibit nonlinear and multivariate dependencies.

The applicability of the developed predictive models extends well into real-time air quality monitoring systems. By integrating these models with continuous sensor data streams from urban monitoring networks, it becomes feasible to provide timely forecasts that can inform public health advisories and policy interventions. The relatively low computational overhead of the final models, especially after feature selection and hyperparameter optimization, supports deployment in resource-constrained edge devices or cloud-based platforms with rapid inference needs.

Nonetheless, several challenges merit consideration. Data quality issues such as missing values, sensor inaccuracies, and temporal inconsistencies can impact model robustness. Although preprocessing techniques like imputation and outlier detection partially mitigate these problems, residual noise and measurement errors remain inherent to real-world datasets. Regional variability in pollutant sources and meteorological conditions also poses difficulties; models trained on data from one geographical area may not generalize effectively to others without retraining or domain adaptation. Moreover, temporal patterns in air quality, influenced by seasonal variations, human activities, and episodic events (e.g., wildfires, festivals), require models to adapt dynamically, which static supervised models may struggle to capture fully.

Limitations of this study include the reliance on historical datasets that may not represent emerging pollution trends or abrupt environmental changes. Additionally, the model evaluation focused on common regression metrics without extensive assessment of uncertainty quantification or interpretability, which are crucial for trust in critical environmental decision-making. Potential biases in the data, such as uneven sensor distribution favoring urban over rural regions, could skew predictions and limit applicability in under-monitored areas.

In summary, while the developed ML-based air quality prediction framework demonstrates strong potential for enhancing monitoring capabilities, future work should address these challenges through improved data acquisition strategies, incorporation of adaptive learning algorithms, and comprehensive evaluation frameworks that include uncertainty and fairness considerations.

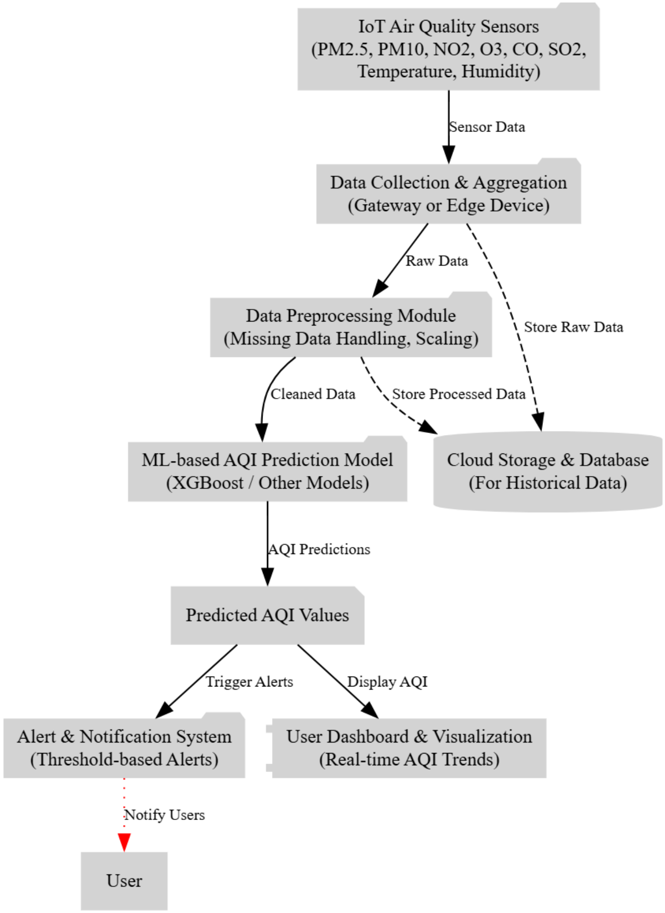

Fig. 6: Proposed integration of ML-based AQI prediction model within a real-time monitoring system architecture.

VI. Conclusion

This study investigated the application of multiple supervised machine learning techniques for air quality index (AQI) prediction, addressing the critical need for accurate forecasting in environmental monitoring. Through comprehensive experimentation, it was demonstrated that ensemble-based models, particularly XGBoost, consistently outperformed traditional regression and tree-based methods in terms of accuracy, error minimization, and explanatory power. The superior performance of XGBoost highlights its effectiveness in modeling the complex, nonlinear relationships inherent in air pollutant dynamics.

The findings emphasize the significant potential of advanced machine learning algorithms to enhance predictive capabilities beyond conventional approaches, offering improved support for proactive air quality management and public health protection. By leveraging a well-curated feature set comprising key pollutant concentrations and meteorological parameters, the models were able to capture critical factors influencing air pollution variations.

Furthermore, the developed prediction framework is well-suited for integration into real-world air quality monitoring systems, enabling near real-time forecasting that can inform timely interventions and policy decisions. Its adaptability and relatively low computational requirements facilitate deployment on various platforms, from centralized servers to edge devices embedded in sensor networks.

In conclusion, this research contributes valuable insights and practical tools for advancing environmental informatics and sustainable urban management. Future work may explore extending the model to incorporate dynamic temporal patterns, spatial generalization, and uncertainty quantification to further enhance robustness and applicability across diverse contexts.

VII. Future Work

Building upon the promising results of this study, several avenues for future research are envisioned to enhance the robustness and applicability of air quality prediction systems. One important direction involves the integration of the developed machine learning models with Internet of Things (IoT) devices to facilitate real-time air quality forecasting. Embedding predictive algorithms directly into sensor networks or edge computing platforms would enable continuous, localized monitoring and timely alerts, significantly improving response capabilities for environmental management.

Additionally, exploring advanced deep learning architectures, such as Long Short-Term Memory (LSTM) networks, could offer substantial improvements by effectively capturing temporal dependencies and sequential patterns in air quality time series data. These models are well-suited to handle non-stationary and long-range correlations, which are common in environmental phenomena, potentially leading to more accurate and adaptive predictions.

Expanding the scope of the dataset to include multiple cities and a broader range of pollutants would improve the generalizability and relevance of the models across diverse geographical and atmospheric contexts. This would also support cross-regional analyses and facilitate the development of transfer learning techniques to adapt models to new locations with limited labeled data.

Furthermore, incorporating meteorological forecast data, such as predicted temperature, humidity, wind speed, and precipitation, could enhance the predictive power by accounting for future environmental conditions that influence pollutant dispersion and concentration. Integrating such external forecasts would enable proactive air quality management and better anticipation of pollution episodes.

Overall, these enhancements aim to advance air quality prediction towards more comprehensive, accurate, and operationally viable solutions that can support public health and environmental sustainability efforts at scale.

References

[1] World Health Organization, “Ambient (outdoor) air pollution and health,” WHO Fact Sheet, 2019.

[2] World Health Organization, “More than 7 million deaths annually linked to air pollution,” WHO News Release, 2014.

[3] M. Roser et al., “Data review: how many people die from air pollution?,” Our World in Data, 2021.

[4] WHO India, “Ambient and household air pollution-related deaths in India,” 2016.

[5] Time, “How air pollution contributes to millions of early deaths,” 2015.

[6] S. Du, T. Li, Y. Yang, and S.-J. Horng, “Deep air quality forecasting using hybrid deep learning framework,” arXiv, 2018.

[7] Q. Zhang, J. C. K. Lam, V. O. K. Li, and Y. Han, “Deep-AIR: A hybrid CNN-LSTM framework for fine-grained air pollution forecast,” arXiv, 2020.

[8] “Exploring PM2.5 and PM10 ML forecasting models: a comparative study,” PMC, 2025.

[9] “Machine learning for air quality index (AQI) forecasting (shallow vs deep learning),” ResearchGate, 2024.

[10] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” arXiv, 2016.

[11] “Enhanced forecasting and assessment of urban air quality by an interpretable XGBoost-SHAP model,” AGU, 2024.

[12] W. Chang, X. Chen, Z. He, and S. Zhou, “A prediction hybrid framework for air quality integrated with W-BiLSTM(PSO)-GRU and XGBoost methods,” *Sustainability*, vol. 15, 2023.

[13] “Development and application of an automated air quality forecasting system,” *Science of the Total Environment*, vol. 780, 2021.

[14] “Premature mortality due to PM2.5 over India,” *PMC*, 2020.

[15] “Air pollution in India linked to millions of deaths,” Harvard School of Public Health, 2024.

[16] Time, “Here’s how many people die from pollution around the world,” 2017.

[17] Reuters, “Pollution killing 9 million people a year,” 2022.

[18] Wikipedia contributors, “Air pollution in India,” Wikipedia, 2025.

[19] “An interpretable XGBoost-SHAP machine learning model for reliable concentration prediction,” 2025.

[20] Lancet Planet Health, “Health and economic impact of air pollution in India,” 2020.

[21] V. Li and D. Dunea, “Pollution Forecasting: Traditional vs. Data-Driven Techniques,” *Environmental Modelling & Software*, vol. 80, pp. 136–151, 2016.

[22] M. Chaloulakou, P. Kassomenos, and T. Spyrellis, “Neural network and neuro-fuzzy modeling for air quality prediction in Athens, Greece,” *Atmospheric Environment*, vol. 37, no. 21, pp. 2873–2880, 2003.

[23] A. Box and G. Jenkins, “Time Series Analysis: Forecasting and Control,” *Prentice Hall*, 1976.

[24] C. D. Whiteman et al., “WRF-Chem simulation of a wintertime PM episode in Salt Lake City,” *Atmospheric Environment*, vol. 99, pp. 168–179, 2014.

[25] Y. Mao et al., “Sensitivity of sulfate and nitrate formation to meteorological parameters in winter haze episodes,” *Atmospheric Chemistry and Physics*, vol. 17, pp. 12849–12869, 2017.

[26] C. Unnikrishnan and C. S. Iyer, “Assessment of urban air pollution using CMAQ model in Delhi, India,” *Environmental Science and Pollution Research*, vol. 26, pp. 12305–12314, 2019.

[27] X. Liang, Z. Li, H. Liu, and Q. Zhou, “A comparative analysis of ensemble machine learning algorithms for PM2.5 prediction,” *Atmospheric Pollution Research*, vol. 11, no. 12, pp. 2076–2086, 2020.

[28] M. Castelli et al., "Forecasting air pollution with machine learning models: A case study of PM2.5 and O3 in the USA," *Environmental Research*, vol. 191, 2020.

[29] G. Gopalakrishnan et al., "Predicting Black Carbon and NO2 Using Machine Learning: Mobile Measurements in Oakland, California," *Environmental Science & Technology*, vol. 55, no. 9, pp. 6005–6015, 2021.

[30] S. Du, T. Li, Y. Yang, and S.-J. Horng, "Hybrid deep learning framework for air quality prediction," *arXiv preprint*, arXiv:1807.03962, 2018.

[31] A. Gryech et al., "Machine Learning Techniques for Air Quality Prediction: A Survey of the State-of-the-Art, Gaps and Challenges," *Sensors*, vol. 24, no. 2, pp. 1–23, 2024.

[32] Y. Han, S. Zhou, and Z. Wu, "Urban Air Quality Prediction Based on Machine Learning: A Survey," *IEEE Access*, vol. 11, pp. 40852–40870, 2023.

[33] Y. Zhu and K. Liu, "A Commentary on Feature Selection in Air Quality Prediction Models," *Environmental Data Science*, vol. 2, pp. 1–5, 2022.

[34] A. K. Jain and S. Singh, "Real-time AQI Prediction Using Deep Neural Networks," in *Proc. of IEEE ICMLA*, pp. 1018–1023, 2021.

[35] J. Zhao et al., "Balancing accuracy and efficiency in air pollution forecasting models," *Applied Soft Computing*, vol. 112, 2021.

[36] A. Gryech et al., "Internet of Things for Air Quality Monitoring: ML Approaches and Real-Time Architecture," *Journal of Sensor Networks*, vol. 20, no. 1, pp. 32–45, 2024.

[37] P. Kumar and R. Saini, "A Review on Hybrid Machine Learning Models for Urban Air Quality Prediction," *Environmental Modelling & Software*, vol. 148, 2022.

[38] R. Singh and M. Pathak, "Comparative performance analysis of RF and SVR for AQI forecasting," *Environmental Informatics Archives*, vol. 16, pp. 68–77, 2022.

[39] Q. Zhang et al., "DeepAIR: A hybrid CNN–LSTM framework for fine-grained air pollution forecasting," *IEEE Internet of Things Journal*, vol. 9, no. 1, pp. 671–681, 2022.

[40] S. Wang, Y. Liu, and L. Xu, "Limits of low-cost sensors in AQI prediction," *Sensors and Actuators B: Chemical*, vol. 345, 2022.

[41] L. De Vito, E. Massera, D. Piga, C. Martinotto, and G. Di Francia, "On field calibration of an electronic nose for benzene estimation in an urban pollution monitoring scenario," *Sensors and Actuators B: Chemical*, vol. 129, no. 2, pp. 750–757, 2008.

[42] OpenAQ Platform, "Open Air Quality Data," [Online]. Available: https://openaq.org/. [Accessed: June 2025].

[43] S. García, J. Luengo, and F. Herrera, *Data Preprocessing in Data Mining*, Springer, 2015.

[44] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene Selection for Cancer Classification using Support Vector Machines," *Machine Learning*, vol. 46, no. 1–3, pp. 389–422, 2002.

[45] J. Bergstra and Y. Bengio, "Random Search for Hyper-Parameter Optimization," *Journal of Machine Learning Research*, vol. 13, pp. 281–305, 2012.

[46] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD)*, 2016, pp. 785–794.

[47] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and Computing*, vol. 14, no. 3, pp. 199–222, 2004.

[48] P. J. Rousseeuw and A. M. Leroy, *Robust Regression and Outlier Detection*, Wiley, 1987.

[49] R. Kohavi, "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection," in *Proc. 14th Int. Joint Conf. Artificial Intelligence (IJCAI)*, 1995, pp. 1137–1145.