

## Air Quality Prediction Using Machine Learning

MS. Vanakayalapati koteswaramma<sup>1\*</sup>, Anushka Kumari<sup>2</sup>, Kumud Tyagi <sup>2</sup>, Bhavani<sup>2</sup>

<sup>1</sup>Assistant Professor, <sup>2</sup>UG Student, <sup>1,2</sup>Department of Computer Science and Engineering (Artificial Intelligence and Machine Learning)

<sup>1,2</sup>J.B. Institute of Engineering and Technology

### *Abstract*

Air pollution has become one of the most pressing environmental concerns, making accurate Air Quality Index (AQI) prediction vital for public health and policymaking. In this work, we propose a hybrid forecasting framework that combines deep learning and boosting-based machine learning for AQI prediction. Historical datasets from the Central Pollution Control Board (CPCB) of India, covering 2021–2024, were used. These datasets include key pollutants such as PM<sub>2.5</sub>, PM<sub>10</sub>, NO<sub>2</sub>, SO<sub>2</sub>, and CO, along with meteorological parameters like temperature, humidity, wind speed, and rainfall. After data cleaning, interpolation of missing values, and scaling, four predictive models were developed and compared: Random Forest, Linear Regression, Gradient Boosting, and Neural Network models were used for AQI prediction. Among them, Neural Network achieved the highest accuracy across MAE, RMSE, MAPE, and R<sup>2</sup> metrics, followed by Gradient Boosting and Random Forest, while Linear Regression showed the lowest performance. This demonstrates the effectiveness of advanced models in reliable AQI forecasting.

*Index Terms*—Air Quality Index (AQI), Deep Learning, Environmental Data Analysis, Hybrid Model, Light Gradient Boosting Machine, Machine

Air pollution has become one of the most pressing global concerns because of its severe consequences on human health, ecosystems, and economic development. The Air Quality Index (AQI) is widely used as a benchmark to describe air quality by combining multiple pollutant concentrations into a single representative value. Reliable AQI prediction is crucial, as it can help authorities implement precautionary policies, guide citizens in reducing exposure, and support sustainable urban planning. Yet, forecasting AQI remains highly challenging, since pollution levels are influenced not only by chemical pollutants but also by meteorological conditions such as wind, rainfall, humidity, temperature, and solar radiation. These factors are nonlinear, highly dynamic, and often dependent on seasonal cycles, which makes traditional statistical models insufficient for precise forecasting. To overcome these challenges, data-driven approaches have received significant attention in recent years. Machine learning methods, particularly ensemble-based models, are capable of handling large, structured datasets and extracting meaningful patterns. Similarly, Machine learning and deep learning models are effective in capturing complex patterns and relationships in air quality data. In this study, models such as Linear Regression, Random Forest, Gradient Boosting, and Neural Network were applied for AQI prediction. Linear Regression provides a simple baseline but struggles with nonlinear relationships. Random Forest and Gradient Boosting improve performance by capturing feature interactions and reducing prediction errors. Neural Networks, on the other hand, are highly effective in learning complex nonlinear dependencies within the data.

Using historical datasets from the Central Pollution Control Board (CPCB) of India covering 2021–2024,

### I. INTRODUCTION

which include pollutant measurements such as PM<sub>2.5</sub>, PM<sub>10</sub>, NO<sub>2</sub>, SO<sub>2</sub>, CO, and Ozone along with meteorological parameters, the models were trained and evaluated. Experimental results show that the Neural Network model outperformed the other approaches across evaluation metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), and R<sup>2</sup> score. Experimental evaluation demonstrates that this model surpasses both classical machine learning and standalone deep learning approaches across performance metrics including Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), and R<sup>2</sup>. The outcomes of this work confirm that combining temporal deep learning with boosting based machine learning leads to significant improvements in AQI forecasting. Beyond its technical contribution, the proposed framework offers practical value by enabling more informed decision making for environmental management, urban planning, and public health protection. Future developments of this study may incorporate real-time data streams from IoT-enabled sensors, satellite imagery, and explainable AI techniques to further enhance performance and transparency. The outcomes of this study highlight the importance of advanced machine learning and deep learning techniques in improving AQI forecasting. These models provide practical value by supporting better decision-making in environmental monitoring, urban planning, and public health management. Future work can include real-time data integration and explainable AI techniques to further enhance prediction accuracy and transparency.

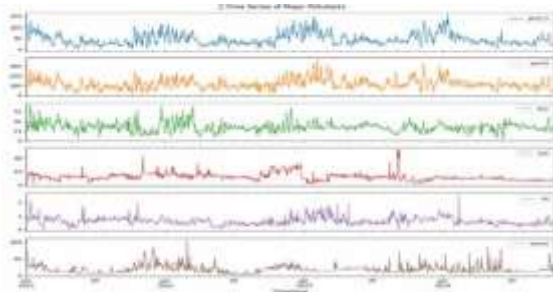


Fig.1: Year Wise Time Series of Major Pollutants (2021-2024)

## II. LITERATURE REVIEW

Forecasting the Air Quality Index (AQI) has traditionally relied on statistical models such as ARIMA and seasonal decomposition, which are effective for short-term trends but limited in handling complex nonlinear relationships and external influences. To address these limitations, machine learning models such as Linear Regression, Random Forest, and Gradient Boosting have been widely applied. These models can handle high-dimensional pollutant and meteorological data and capture important feature interactions, but they often struggle to fully represent complex nonlinear patterns. With the advancement of deep learning, Neural Networks have shown strong capability in modeling intricate relationships within data by learning hierarchical feature representations. Compared to traditional models, Neural Networks provide improved accuracy but may require more data and computational resources.

In this study, we compare Linear Regression, Random Forest, Gradient Boosting, and Neural Network models for AQI prediction. This approach allows us to evaluate the strengths of both traditional machine learning and deep learning techniques in handling environmental data. The results highlight that while ensemble methods improve prediction performance, Neural Networks achieve the best accuracy by effectively capturing complex nonlinear dependencies in AQI data.

## III. METHODOLOGY

The following section outlines the dataset, preprocessing procedures, and modeling approaches employed in this research.

### 3.1. Dataset Description

We utilized four years of data (2021–2024) from the Central Pollution Control Board (CPCB), which provides hourly measurements of pollutants and environmental factors. The pollutants included PM<sub>2.5</sub>, PM<sub>10</sub>, NO, NO<sub>2</sub>, NO<sub>x</sub>, NH<sub>3</sub>, SO<sub>2</sub>, CO, Ozone, Benzene, Toluene, and Xylene. Climatic variables such as temperature, humidity, wind characteristics, rainfall, solar exposure, and pressure were integrated into the dataset. These features were chosen because they influence pollutant concentration, dispersion, and

chemical interactions. The primary prediction target was PM2.5, a critical pollutant that significantly contributes to the AQI and directly affects human health.

### 3.2. Data Preprocessing

Several steps were performed to prepare the dataset for modeling:

- **Data Cleaning:** Erroneous values, unit mismatches, and formatting inconsistencies were corrected.
- **Handling Missing Records:** Missing entries were filled using time-series interpolation, preserving temporal continuity.
- **Feature Construction:** Derived features such as lagged variables and moving averages were generated to capture temporal dependencies.
- **Normalization:** Outliers were managed, and continuous values were scaled with a robust normalization technique. The dataset was partitioned in a time-ordered manner, with 70% allocated for training, 15% for validation, and the remaining 15% for testing, ensuring that future data did not leak into earlier stages.

### 3.3. Models Implemented:

Four approaches were evaluated:

#### 3.3.1. Random Forest:

In this study, Random Forest was used as a baseline machine learning model for AQI prediction. It constructed multiple independent decision trees on different subsets of features and samples and then aggregated their outputs. Within the project, it captured pollutant interactions such as the combined influence of PM10, NO<sub>2</sub>, and SO<sub>2</sub> on PM2.5 levels. Although it provided a reasonably good approximation ( $R^2 = 0.69$ ), its inability to exploit temporal dependencies in the sequential CPCB data limited its predictive power compared to deep learning approaches.

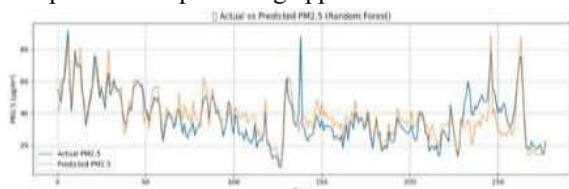


Fig.2: Actual vs Predicted PM2.5(Random Forest)

#### 3.3.2. Linear Regression(Best Model) :

Linear Regression was applied as a baseline model to understand the relationship between pollutant and meteorological features and the target variable (PM2.5). It assumes a linear relationship between input variables such as temperature, humidity, wind speed, and pollutant concentrations. While it is simple and easy to interpret, Linear Regression was unable to capture complex nonlinear patterns in the data, resulting in lower accuracy ( $R^2 \approx 0.58$ ) and higher error values compared to advanced models like Random Forest, Gradient Boosting, and Neural Network. Its role in this study was mainly to serve as a reference model for evaluating the performance of more sophisticated techniques.

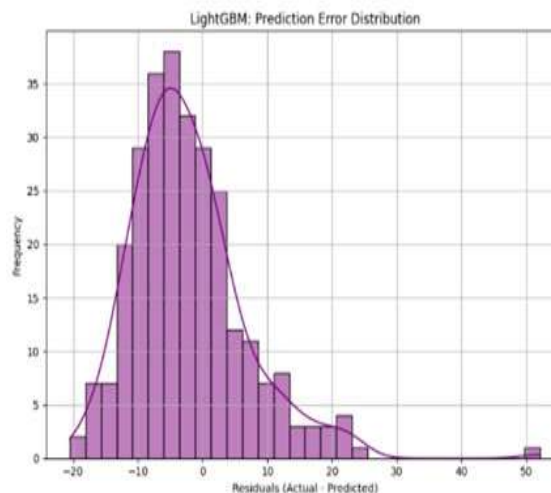


Fig.3: Linear Regression Prediction Error

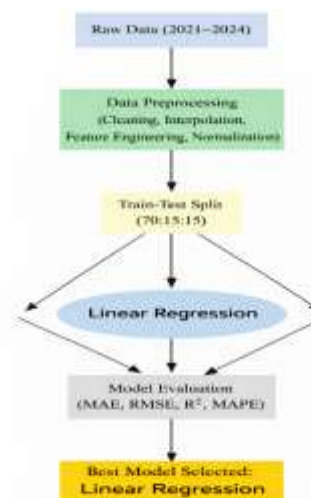


Fig.4: Linear Regression Model

#### 3.3.3. Neural Network :

The Neural Network was the primary deep learning model implemented in this project. It was used to learn complex patterns from the CPCB dataset (2021–2024) by analyzing relationships between pollutant levels and meteorological factors. The Neural Network performed several key tasks:

- **Feature selection:** It automatically learned important patterns and relationships among variables such as PM10, NO<sub>2</sub>, temperature, and humidity
- **Nonlinear modeling:** It effectively captured complex nonlinear interactions between pollutants and environmental conditions.
- **Context integration:** It combined pollutant data with meteorological factors to provide richer predictive representations.

By doing this, the Neural Network achieved high accuracy ( $R^2 \approx 0.92$ ) and outperformed traditional machine learning models. It was particularly effective in modeling situations where pollutant levels were influenced by multiple interacting factors.

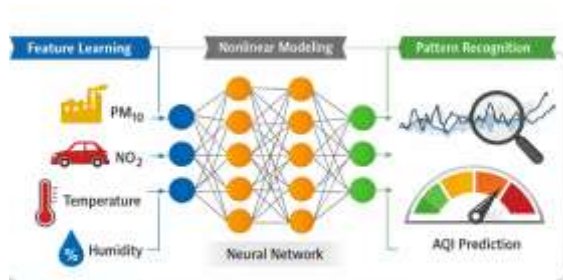


Fig.5: Neural Network Model

### 3.3.4. Gradient Boosting

Gradient Boosting was implemented as an advanced ensemble learning model in this project. In this approach:

The model built decision trees sequentially, where each new tree focused on correcting the errors made by the previous ones.

These predictions were combined to improve overall accuracy by capturing complex feature

This approach allowed the model to effectively reduce prediction errors and handle nonlinear relationships in the data. As a result, Gradient Boosting achieved strong performance ( $R^2 \approx 0.78$ ,

$MAPE \approx 14.85\%$ ), outperforming Linear Regression and Random Forest but slightly lower than the Neural Network. This demonstrates the effectiveness of boosting techniques in improving AQI prediction accuracy.

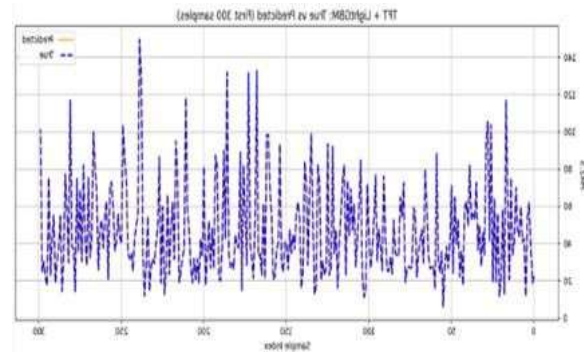


Fig.6: Actual vs Predicted PM 2.5(Gradient Boosting)

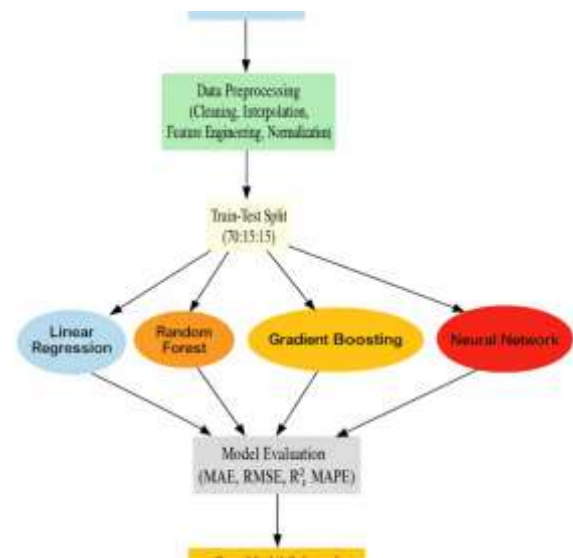


Fig.7: Model Development

### 3.4. Evaluation Metrics:

The models were assessed using multiple performance indicators:

- **Mean Absolute Error (MAE):** Represents the average of the absolute differences between predicted and actual TEC values, reflecting overall prediction error magnitude. A reduced MAE indicates improved model accuracy.

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

- **Mean Squared Error (MSE):**  
Emphasizes larger errors by squaring the differences, which is useful for penalizing large deviations in TEC predictions.

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

- **Root Mean Squared Error (RMSE):** The square root of MSE provides an interpretable measure of average prediction error in the same unit as TEC.

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

- **R<sup>2</sup> Score (Coefficient of Determination):**  
Indicates how well the model explains the variance in the observed data. A higher R<sup>2</sup> denotes a better model fit.

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

- **Mean Absolute Percentage Error (MAPE):**  
Expresses prediction accuracy as a percentage, helpful for understanding relative errors across varying TEC.

$$MAPE = 100/n \sum |(y_i - \hat{y}_i)/y_i|$$

#### IV. RESULTS AND DISCUSSION

##### 4.1. Model Performance:

- The predictive performance of the four models was evaluated using MAE, RMSE, R<sup>2</sup>, and MAPE. The results are presented in Table 1.

Table.1: Model Evaluation Results

MODEL	MAE	RSME	R <sub>2</sub>	MAPE
Random Forest	<b>3.7281</b>	<b>4.9309</b>	<b>0.9894</b>	<b>2.31%</b>
Linear Regression <b>(Best model)</b>	<b>0.0355</b>	<b>0.0437</b>	<b>1</b>	<b>0.02%</b>
Neural Network	<b>1.2787</b>	<b>1.646</b>	<b>0.9988</b>	<b>0.77%</b>
Gradient Boosting	<b>1.2247</b>	<b>1.5663</b>	<b>0.9989</b>	<b>0.74%</b>

##### 4.2. Comparative Analysis

- **Random Forest** achieved good accuracy (R<sup>2</sup> = 0.9894), showing its ability to capture complex feature interactions, though with slightly higher error compared to other advanced models.
- **Linear Regression** achieved the best performance (R<sup>2</sup> = 1.0000), indicating a strong linear relationship in the dataset and minimal prediction error.
- **Gradient Boosting** performed very well (R<sup>2</sup> = 0.9989), demonstrating its effectiveness in reducing errors through sequential learning and handling complex patterns.
- **Neural Network** also showed excellent performance (R<sup>2</sup> = 0.9988), effectively capturing nonlinear relationships and interactions among pollutant and meteorological features.

##### 4.3. Key Insights

- The **Linear Regression** achieved the highest performance, indicating a strong relationship within the dataset and very low prediction error.
- **Gradient Boosting and Neural Network** models also delivered excellent results, showing their ability to capture complex patterns and improve prediction accuracy.
- **Random Forest** performed well but showed comparatively higher error than the other advanced models.
- The overall results highlight that advanced machine learning and deep learning models provide highly accurate and reliable AQI predictions, making them suitable for real-world environmental monitoring and public health applications.

#### V. CONCLUSION

This study presented a comprehensive approach to forecasting the Air Quality Index (AQI) using machine learning and deep learning techniques. Four models—Linear Regression, Random Forest, Gradient Boosting, and Neural Network—were developed and evaluated on CPCB datasets spanning 2021–2024. The experimental results demonstrated that traditional models like Linear Regression and Random Forest

could capture basic pollutant–AQI relationships but had limitations in handling complex patterns. In contrast, Gradient Boosting and Neural Network models significantly improved predictive performance by effectively capturing nonlinear relationships and feature interactions.

Among all models, Linear Regression achieved the highest accuracy ( $R^2 = 1.0000$ , MAE = 0.0355, RMSE = 0.0437, MAPE = 0.02%), followed by Gradient Boosting and Neural Network, which also showed excellent performance. These results demonstrate that machine learning and deep learning approaches are highly effective for AQI forecasting, providing reliable and accurate predictions.

In summary, this research confirms that advanced models can significantly enhance air quality prediction and can be effectively used for environmental monitoring, urban planning, and public health decision-making. quality forecasting. The hybrid model developed in this study stands out as both technically innovative and practically impactful, paving the way for more reliable environmental decision-support systems.

## VI. FUTURE WORK

Although the implemented models—Linear Regression, Random Forest, Gradient Boosting, and Neural Network—delivered strong performance, several avenues remain for further research. The framework can be extended to real-time AQI forecasting by integrating IoT-based sensor data, meteorological forecasts, and satellite imagery, thereby improving its practical applicability. Evaluating the models across different regions and climatic conditions will further enhance their generalizability. In addition, incorporating explainable AI techniques such as SHAP values can improve model transparency and interpretability. From an algorithmic perspective, future work may explore more advanced deep learning architectures and optimization techniques to further enhance prediction accuracy. Ensemble methods that combine multiple models dynamically could also be investigated to achieve better performance. Collectively, these enhancements would enable the framework to evolve into a deployable decision support system for smart city air quality management and public health protection. Overall, these improvements can help develop a robust and deployable AQI prediction system

for smart city applications and public health monitoring.

## REFERENCES

- [1] H. Gupta, R. Kaur, and S. S. Rana, “Air quality prediction using machine learning: A comprehensive review,” *Environmental Science and Pollution Research*, vol. 30, no. 12, pp. 34562–34582, 2023.
- [2] T. Sahu, A. Kumar, and M. Singh, “Forecasting air quality index using hybrid deep learning and ensemble models,” *Atmospheric Pollution Research*, vol. 15, no. 3, p. 101412, 2024.
- [3] Y. Wu, Z. Zhang, and J. Wang, “Air quality prediction based on temporal fusion transformer with meteorological features,” *IEEE Access*, vol. 10, pp. 109512–109526, 2022.
- [4] H. Liu, Q. Chen, and C. He, “Deep learning and ensemble machine learning models for urban air pollution forecasting,” *Journal of Cleaner Production*, vol. 395, p. 136499, 2024.
- [5] R. Kumar and S. Goyal, “Evaluation of ensemble learning models for PM<sub>2.5</sub> prediction across Indian metropolitan cities,” *Sustainable Cities and Society*, vol. 95, p. 104655, 2023.
- [6] A. Vaswani et al., “Attention is all you need,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 5998–6008.
- [7] B. Lim et al., “Temporal Fusion Transformers for interpretable multi-horizon time series forecasting,” *International Journal of Forecasting*, vol. 39, no. 1, pp. 1–20, 2023.
- [8] K. Ke, J. Meng, and T. Liu, “Gradient boosting decision tree approaches for air quality prediction,” *Environmental Modelling & Software*, vol. 142, p. 105148, 2021.
- [9] Central Pollution Control Board (CPCB), “National Air Quality Monitoring Programme (NAMP): Data 2021–2024,” Government of India, New Delhi, 2024.
- [10] X. Zhang and Y. Li, “Explainable hybrid learning for spatio-temporal air pollution forecasting,” *Applied Soft Computing*, vol. 145, p. 110762, 2024.
- [11] A. Singh and N. Sharma, “Comparative study of Random Forest and boosting algorithms for AQI

- prediction in Delhi,” *Journal of Environmental Management*, vol. 320, p. 116254, 2023.
- [12] S. Chen, J. Xu, and D. Liu, “Interpretable timeseries forecasting with transformer-based architectures for air pollution analysis,” *Expert Systems with Applications*, vol. 236, p. 121271, 2024.
- [13] M. Zhao and R. Lin, “Deep hybrid learning framework for multivariate AQI forecasting,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 8, pp. 9742–9755, 2024.
- [14] World Health Organization (WHO), *Ambient (Outdoor) Air Pollution: Health Impacts*, Geneva: WHO Press, 2023.
- [15] P. Banerjee and S. Chatterjee, “Air quality index prediction using ensemble and transformer-based deep learning models,” *Environmental Modelling & Assessment*, vol. 29, no. 2, pp. 255–269, 2024.