

Air Quality Prediction Using Ensemble Machine Learning Models with Environmental and Meteorological Features

¹Karthikeyan T, ²Vivekanandan S J, ³Ashwini Barbadekar, ⁴Vasukidevi G,
⁵Muthukumar Subramanian and ⁶Reny Jose

¹Department of Computer Science and Business Systems, Panimalar Engineering College, Chennai, Tamil Nadu, India.

²Department of Computer Science and Engineering, Dhanalakshmi College of Engineering, Chennai, Tamil Nadu, India.

³Department of Electronics and Telecommunications Engineering, Vishwakarma Institute of Technology, Pune, India.

⁴Department of Computer Science Engineering, RMK Engineering College, Kavaraipettai, Tamil Nadu, India.

⁵Department of Computer Science and Engineering, Hindustan Institute of Technology and Science, Deemed University, Chennai, Tamil Nadu, India.

⁶Department of Computer Applications, Marian College Kuttikkanam Autonomous, Kuttikkanam, Kerala, India.

¹karthi4cse@gmail.com, ²saivivekphd@gmail.com, ³ashwini.barbadekar@vit.edu, ⁴vasukidevi@gmail.com,

⁵drsm.iit@gmail.com, ⁶reny.jose@mariancollege.org

Correspondence should be addressed to Karthikeyan T : karthi4cse@gmail.com

Article Info

Journal of Machine and Computing (<https://anapub.co.ke/journals/jmc/jmc.html>)

Doi : <https://doi.org/10.53759/7669/jmc202606011>

Received 22 July 2025; Revised from 27 September 2025; Accepted 08 October 2025.

Available online 18 October 2025.

©2026 The Authors. Published by AnaPub Publications.

This is an open access article under the CC BY-NC-ND license. (<https://creativecommons.org/licenses/by-nc-nd/4.0/>)

Abstract – Air pollution, especially fine particulate matter (PM_{2.5}), poses serious health and environmental threats globally. Traditional models for air quality prediction often fall short in capturing the complex, dynamic nature of pollution due to their linear assumptions and lack of contextual information such as meteorological or human-activity patterns. This research work proposes an ensemble machine learning framework that integrates both environmental pollutant data and temporal features derived from timestamps to enhance the accuracy, robustness, and interpretability of PM_{2.5} prediction. The dataset used includes hourly air quality measurements from Delhi, consisting of pollutant concentrations (CO, NO, NO₂, O₃, SO₂, PM₁₀, NH₃) and timestamps. Data preprocessing involves parsing timestamps into structured datetime features (hour, day, month, weekday), handling missing values using mean-based imputation, and applying standard normalization. Two ensemble learning models—Random Forest Regressor (RF) and Gradient Boosting Regressor (GBR)—are trained to predict PM_{2.5} concentrations. Model performance is evaluated using MAE, MSE, RMSE, and R². Additionally, feature importance analysis from the Gradient Boosting model is conducted to enhance explainability. The Random Forest model achieved superior predictive performance, with an MAE of 9.386, RMSE of 15.265, and an R² score of 0.995, outperforming Gradient Boosting which yielded an MAE of 11.794, RMSE of 17.585, and R² of 0.994. These results significantly improve upon earlier baseline performances (e.g., MAE = 22.4), highlighting the impact of integrating temporal features and comprehensive preprocessing. Feature importance analysis further identified key contributors to PM_{2.5} concentration, improving interpretability. The proposed framework demonstrates that combining environmental and temporal features with ensemble models significantly enhances air quality prediction. The Random Forest model, in particular, proves effective in modeling complex, non-linear pollution behavior. The addition of feature explainability supports data-driven decision-making for environmental policy and real-time public health interventions.

Keywords – Machine Learning, Air Quality Prediction, PM_{2.5} Forecasting, Ensemble Learning, Random Forest, Gradient Boosting, Temporal Features, Feature Importance.

I. INTRODUCTION

Air pollution remains one of the most critical environmental and public health challenges of the 21st century. According to the World Health Organization (WHO), more than seven million premature deaths annually are linked to exposure to poor air quality [1]. Among various pollutants, PM_{2.5}—particulate matter with a diameter less than 2.5 microns—poses the greatest threat due to its ability to penetrate deep into the respiratory system and bloodstream, contributing to diseases such as asthma, bronchitis, stroke, and heart disease [2][3]. As urbanization, industrial growth, and vehicular emissions continue to accelerate worldwide, especially in developing nations, there is an urgent need for accurate and timely air quality prediction systems to support preventive action and long-term environmental policy [4]. The challenge of forecasting air

quality lies in its complex and dynamic nature. PM_{2.5} concentrations are influenced by a combination of meteorological conditions (such as wind, humidity, and temperature), seasonal trends, and human activities (like traffic density, industrial operations, and agricultural burning) [5]. In this context, India, and more specifically Delhi, consistently ranks among the most polluted cities in the world. Recurrent smog episodes, a high density of vehicles, industrial emissions, and crop residue burning in nearby states lead to hazardous pollution levels, particularly during the winter months [6]. Despite the availability of monitoring networks and increased awareness, existing air quality management systems often fail to deliver region-specific, interpretable, and real-time forecasts that can inform health advisories and policy interventions. Globally, air quality is deteriorating, and indoor air pollution poses a significant hazard to human health [7]. Traditionally, air quality forecasting has relied on statistical and linear models such as ARIMA, regression-based methods, or deterministic simulations [8]. While useful in controlled conditions, these approaches assume linearity and stationarity, which rarely hold in real-world air pollution data [9]. Such methods struggle with non-linear interactions, noise, missing data, and shifting trends, making them less effective in highly dynamic environments like Delhi. As a result, there has been a growing shift toward machine learning (ML) approaches, which can learn complex, non-linear relationships directly from data and adapt to changing patterns without strong assumptions [10]. Among ML techniques, ensemble models such as Random Forest (RF) and Gradient Boosting Regressor (GBR) have shown strong performance in a variety of prediction tasks [11]. These aggregate the predictions of several decision trees to increase robustness, decrease overfitting and account for multivariate interactions. Nevertheless, today's widely-used ML models for air quality prediction have some severe drawbacks regarding their predicting effectiveness. Restricted feature context: Several models only use the pollutant concentration features and ignore information such as hour of day, day of week or month. These temporal cues serve as surrogates for meteorological processes or human activities, such as rush hour traffic or seasonal smog, and are essential in allowing the model to capture the cyclic nature of pollution.

Black-box predictions are trusted too much: High performing ML models are increasingly relying on black-boxes with few interpretable features driving their predictions. This restricts their trust and applicability in environmental planning and policy-making, where the rationale behind a prediction is as influential as the prediction itself [12][13]. To address these gaps, the present study proposes a comprehensive air quality forecasting framework that integrates both environmental and time-derived features using ensemble machine learning models. The raw dataset, comprising pollutant concentrations and timestamps, is preprocessed to derive temporal variables such as hour, day, month, and weekday. These features, alongside pollutants like CO, NO, NO₂, O₃, SO₂, PM₁₀, and NH₃, provide a rich input space that captures both chemical and behavioral influences on PM_{2.5} levels. The missing values of the data are then imputed using mean-based approaches and the data is normalized to have consistent feature scale. After a standard train-test splitting two ensemble regressors (Random Forest and Gradient Boosting) are trained to predict PM_{2.5} concentrations. These models are chosen due to their capability of modeling non-linear relationships, dealing with missing or noisy data, learning complex feature interactions that are not possible in traditional models. The common regression metrics, Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R² score are utilized to assess the quality of the models with respect to prediction accuracy and model generalization. Moreover, the proposed framework leverages an explainability module by the feature importance obtained from the Gradient Boosting model. This approach serves to pinpoint environmental and timing factors that most explain the variation of PM_{2.5} scales, thus guiding potential actions such as local emission control or traffic limitations at rush-hour times. The inspiration of this study is based on three key gaps found in the current air quality prediction systems. First, there is few temporal features incorporated in the existing predictors, i.e., many predictors ignore the time-related variables, like hour, day and month, which are crucial for capturing the underlying temporal meteorological patterns and human activity cycles that impact the pollution levels. This gap in evidence has implications for models not addressing contextual and seasonal effects. Second, legacy linear models are insufficient in describing the strongly non-linear and dynamic interactions between pollutants, smoke sources and weather conditions, leading to poor prediction in real life. A third challenge is that many machine learning models are not interpretable and function as black boxes that provide little insight into the features that drive their predictions. This greatly restricts their use for environmental policy and informing the public (the reasons for predictions are important for informed decisions). The primary objectives of the proposed work are threefold. First, it aims to integrate environmental and meteorological features by developing a machine learning framework that combines pollutant concentration data with time-derived variables such as hour, day, month, and weekday. This integration is designed to capture seasonal patterns, daily fluctuations, and recurring trends in air quality, allowing the model to learn complex temporal dynamics that influence pollution behavior. Second, the study seeks to employ ensemble learning models, specifically Random Forest and Gradient Boosting Regressors, to enhance predictive accuracy. These models are well-suited to handle the non-linearity and intricate feature interactions inherent in air quality datasets, offering a robust, data-driven alternative to traditional linear approaches. Third, the work focuses on improving model explainability by leveraging feature importance analysis from ensemble models. This enables the identification of key environmental and temporal contributors to PM_{2.5} levels, facilitating actionable insights for environmental policy-making and public health decision support. This proposed framework not only improves prediction accuracy but also enhances transparency and applicability for real-world air quality monitoring and intervention planning. The remainder of the paper is structured as follows: Section 2 reviews related literature, Section 3 details the dataset and methodology, Section 4 presents experimental results and analysis, and Section 5 concludes the study with future research directions.

II. LITERATURE REVIEW

Air quality has emerged as a vital research field since it is directly related with public health and environmental preservation. Precise forecasting of air pollution is critical for early warning, urban planning and decision making. In recent years, many researchers have used machine learning (ML) methods to predict air quality, utilizing environment parameter and meteorological data to enhance prediction accuracy. This paper provides a literature review on air quality predictions methods, especially for ensemble learning methods, and the combination with various atmospheric variables. Seng et al. [14] proposed a deep learning-based model called MMSL (Multi-output and Multi-index Supervised Learning), built on the LSTM framework to predict multiple air quality indicators—PM_{2.5}, CO, NO₂, O₃, and SO₂—simultaneously. The model integrates particle concentration data from current and nearby monitoring stations, along with meteorological and gaseous pollutant data, all preprocessed into a supervised learning format and normalized. It was evaluated using time series data from 35 monitoring stations across Beijing between 2016 and 2017. The results demonstrated that MMSL outperformed both baseline and advanced models in prediction accuracy. Its main advantages include the ability to handle multivariable inputs and produce multi-output predictions. However, a key technical limitation is its reliance on fixed spatial aggregation and the absence of mechanisms to explicitly capture spatiotemporal correlations, which may restrict its adaptability in more dynamic or complex urban environments. Athira et al. [15] proposed a deep learning-based framework for predicting PM₁₀ concentrations using RNN, LSTM, and GRU models trained on AirNet pollution and meteorological time series data. Each model was evaluated over 1000 epochs with varying learning rates (0.01–0.5), and all showed comparable predictive performance. The main advantage of the framework is its reduced computational complexity compared to traditional numerical methods. However, a key technical gap is the absence of spatial feature integration and limited model interpretability. Harishkumar et al. [16] proposed a machine learning-based model to forecast PM₁₀ and PM_{2.5} concentrations using the Taiwan Air Quality Monitoring dataset (2012–2017). The model outperformed traditional methods based on RMSE and MAE metrics. Its main advantage is improved prediction accuracy, while the key technical gap is the lack of spatial and temporal dependency handling. Zhang et al. [17] proposed an improved air quality prediction model based on LightGBM to forecast PM_{2.5} concentrations across 35 monitoring stations in Beijing for the next 24 hours. To address challenges such as data instability and temporal variation in pollutant levels, the model incorporates forecasting data as input and uses a sliding window mechanism to extract high-dimensional temporal features, significantly expanding the training dataset. The experimental results demonstrated that the proposed model outperformed existing approaches in prediction accuracy. Key advantages include efficient handling of large-scale, high-dimensional data and better utilization of spatial-temporal information. However, the main technical gap lies in the limited interpretability of the model and its dependence on manual feature engineering. Soh et al. [18] built a hybrid model using neural network networks (ANN), CNN and long short-term memory network (LSTM) for predicting air quality for 48 h in advance. We exploit recent meteorological variables, elevation variables and spatial-temporal trends between locations. Results showed high prediction accuracy. Its primary merit lies in the ability of effectively learning complex spatial-temporal pattern, while its major technical gap lies in the over-high model complexity but weak interpretability. Castelli et al. [19] developed an SVR with an RBF kernel model for predicting pollutant concentrations and AQI based on airborne PM_{2.5} and PM₁₀ for Shanghai. The model did not reduce the dimension of the input variables and used all of the input information for higher precision. Good predictive efficacy was demonstrated, with classification accuracy (of AQI in 6 categories by the US EPA) reaching 94.1%. The strength of the model is its ability to accurately predict hourly concentrations of pollutants and AQI. The main technical bottleneck is its poor scalability and inefficiency in the face of large-scale high-dimensional streaming datasets.

Freeman et al. [20] presented a DL-based model, a LSTM-based RNN predicting averaged 8-hourly O₃ concentration at the surface up to 72 h in the future. The model was trained using hourly air quality and meteorological data, with missing values imputed using a new time-based averaging method and input features were reduced with decision tree-based selection. The accuracy obtained was high with MAE less than 2 for 72 h forecasting. Key advantages include effective long-range prediction, reduced input dimensionality, and real-time capability using only essential parameters. The main technical gap lies in the sensitivity of the model to look-back parameter settings, which, if misaligned, can introduce significant prediction errors. Liang et al. [21] proposed a range of machine learning models including adaptive boosting (AdaBoost), artificial neural networks (ANN), random forest, stacking ensemble, and support vector machines (SVM) for air quality prediction using an 11-year dataset provided by Taiwan's Environmental Protection Administration (EPA). The study aimed to enhance the prediction accuracy of Air Quality Index (AQI) levels across three distinct geographic regions. Among the models tested, the stacking ensemble model—an ensemble technique that combines the predictions of multiple base learners to improve generalization—achieved consistently superior performance in terms of R² and RMSE values. AdaBoost, on the other hand, yielded the lowest Mean Absolute Error (MAE), indicating high accuracy in pointwise predictions. The key advantage of this study lies in its extensive use of a longitudinal dataset, which enables more robust modeling of seasonal and regional air quality patterns. However, a significant technical gap remains in handling real-time dynamic changes in pollutant levels and incorporating meteorological variability at finer spatial and temporal resolutions. Ma et al. [22] proposed the Transferred Bi-directional Long Short-Term Memory (TL-BLSTM) model to improve air quality prediction at larger temporal resolutions. The model combines bi-directional LSTM to capture long-term dependencies in PM_{2.5} data with transfer learning to enhance prediction across daily to monthly scales. Tested in Guangdong, China, TL-BLSTM outperformed traditional machine learning models with lower prediction errors. Its main

advantage is accurate long-term forecasting, but it relies heavily on high-quality data and struggles with unexpected pollution changes.

Liao et al. [23] proposed the DM_STGNN (Dynamic Multi-granularity Spatio-temporal Graph Neural Network) for air quality prediction by integrating physical mechanism models and machine learning. The model uses the HYSPLIT trajectory model to dynamically build spatio-temporal graphs, capturing complex interactions among meteorological, spatial, and temporal features. An encoder-decoder architecture with LSTM and attention mechanisms is employed to learn both spatial and temporal dependencies, and unsupervised pre-training is used to capture long-term patterns. Experimental results on the Yangtze River Delta dataset show that DM_STGNN outperforms state-of-the-art models. The key advantages include fine-grained spatial modeling, dynamic edge construction, and long-term temporal learning. However, the model's reliance on external tools like HYSPLIT increases computational complexity and may affect scalability for real-time or large-scale applications.

The existing literature showcases a wide range of models—from traditional regression to advanced deep learning and graph-based architectures—for air quality prediction. While many models have demonstrated high accuracy, they often face challenges such as limited interpretability, high computational complexity, and insufficient integration of temporal or spatial dependencies. These limitations highlight the need for a more interpretable, scalable, and feature-rich approach that balances predictive performance with practical applicability. The proposed work aims to address these gaps by leveraging ensemble learning techniques combined with environmental and time-derived features for robust and explainable air quality forecasting.

III. PROPOSED METHODOLOGY

This study proposes a robust, data-driven framework for predicting PM_{2.5} concentrations using ensemble machine learning models. The methodology integrates both environmental features (pollutant levels) and time-derived features (as proxies for meteorological and anthropogenic influences) to enhance predictive accuracy and contextual understanding. The process consists of six key stages.

Fig 1 illustrates the overall architecture of the proposed framework for air quality prediction using ensemble machine learning models. The process begins with the acquisition of a structured dataset comprising various pollutant measurements and meteorological observations. The initial step involves data preprocessing and feature extraction, where the raw timestamp is parsed into a standardized datetime format. From this, additional temporal features such as hour, day, month, and weekday are derived, allowing the model to capture time-dependent pollution patterns. The core features used include environmental pollutants (CO, NO, NO₂, O₃, SO₂, PM₁₀, and NH₃) and meteorological proxies (hour, day, month, weekday), which together provide a holistic view of both chemical and behavioral contributors to air quality. Following this, missing values are handled through imputation to ensure dataset completeness and prevent training bias. The cleaned data is then subjected to feature scaling using standard normalization techniques to harmonize the input range and improve model convergence. Subsequently, the processed data is used to train two robust ensemble learning models—Random Forest Regressor and Gradient Boosting Regressor—which are well-suited for capturing non-linear relationships and handling complex, high-dimensional data. After training, the models undergo evaluation based on performance metrics such as MAE, MSE, RMSE, and R². Additionally, feature importance analysis is conducted to interpret the contribution of each input variable, offering explainability and insights into which environmental or temporal factors most influence PM_{2.5} concentration. Finally, the validated models are used for accurate air quality prediction, completing the end-to-end pipeline. This modular and interpretable architecture ensures scalability, reliability, and strong predictive performance for real-time or policy-driven air quality monitoring applications.

Data Preprocessing and Feature Engineering

The input data consists of hourly air quality measurements collected from Delhi, including pollutant concentrations such as CO, NO, NO₂, O₃, SO₂, PM₁₀, and NH₃, along with timestamps. To enrich the dataset with meteorological and behavioral patterns (e.g., rush hour traffic, seasonal smog), we perform temporal decomposition of the date field to extract the following features. Hour=hour(date), Day=day(date), Month=month(date), Weekday=weekday(date). These time components act as indirect meteorological indicators, enabling the model to learn patterns such as seasonal pollution peaks (e.g., winter haze) or weekday traffic emissions. This step addresses the gap in prior work that ignores time-contextual variability in air pollution.

Missing Value Imputation and Feature Normalization

Given the real-world nature of environmental data, missing values are common. We address this by applying mean-based imputation as shown in equation (1).

$$X_{i,j} = \begin{cases} x_{i,j}, & \text{if } x_{i,j} \neq \text{NaN} \\ \frac{1}{N_j} \sum_{i=1}^{N_j} x_{i,j}, & \text{if } x_{i,j} = \text{NaN} \end{cases} \quad (1)$$

Where N_j is the number of valid values in feature j .

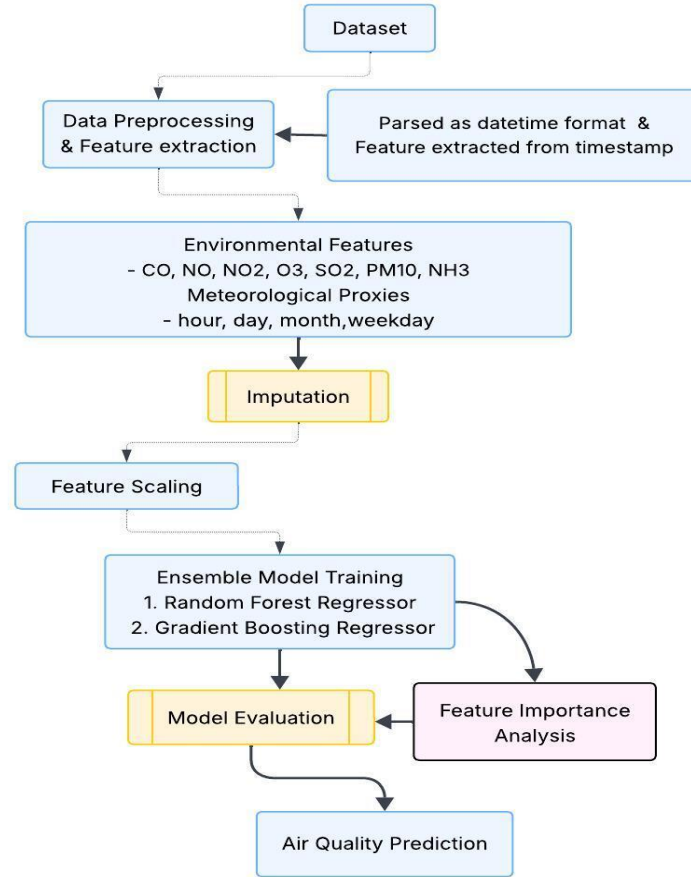


Fig 1. Proposed System Architecture for Air Quality Prediction Using Ensemble Machine Learning Models.

Following imputation, we perform standard scaling to ensure features are on a uniform scale as represented in equation (2).

$$X'_{i,j} = \frac{X_{i,j} - \mu_j}{\sigma_j} \quad (2)$$

where μ_j and σ_j are the mean and standard deviation of feature j , respectively. This normalization is crucial for ensemble models to treat all features equitably and optimize split thresholds during training.

Problem Formulation

We define air quality prediction as a supervised regression task, where the objective is to predict the PM2.5 concentration (y) given a feature vector X as shown in equation (3).

$$\hat{y} = f(X) \quad (3)$$

Where f is the regression model. Here, $X \in R^{n \times d}$ contains both environmental and time-derived features, and $\hat{y} \in R^{n \times d}$.

Model Training using Ensemble Learning

To model complex non-linear relationships and interactions between pollutants and meteorological proxies, we adopt two ensemble learning techniques:

Random Forest Regressor (RF)

As shown in Equation (4), an ensemble of decision trees trained on bootstrapped samples of the training data. The final prediction is the average of outputs from individual trees.

$$\hat{y}_{RF} = \frac{1}{T} \sum_{t=1}^T h_t(X) \quad (4)$$

where h_t is the prediction from the t^{th} tree and T is the total number of trees. One of the key advantages of the method is its ability to handle missing values and outliers effectively, ensuring robust performance even with imperfect data. It also automatically captures feature interactions, eliminating the need for extensive manual feature engineering. Additionally,

the approach helps reduce overfitting by using an averaging mechanism, which improves the model's generalization to unseen data.

Gradient Boosting Regressor (GBR)

As shown in Equation (5), it builds an additive model in a forward stage-wise fashion by fitting a regression tree on the residuals of the previous model.

$$\hat{y}^{(m)} = \hat{y}^{(m-1)} + \eta \cdot h_m(X) \quad (5)$$

where η is the learning rate and $h_m(X)$ is the tree fitted to residuals in stage m . This method concentrates learning on hard-to-predict samples. This method offers several advantages. It achieves high accuracy by iterative error correction, refining predictions round by round. It allows fine-grained control over the model behavior with tuning parameters like tree depth or learning rate. In addition, it performs very well on structured or tabular data, so it's a versatile choice for a wide range of practical machine learning problems.

Performance Metrics

Loss Function: Mean Squared Error (MSE)

Mean Squared Error (MSE) is a loss function commonly applied in air quality prediction applications (i.e., the air quality prediction tasks involved may include predicting PM2.5 concentrations). MSE measures the mean d2 of the differences between the actual and the predicted values of PM2.5 and this is a good measure of the performance of the model. Mathematically it is defined as given below in Equation (6).

$$MSE = \frac{1}{n} \sum_i^n (\hat{y}_i - y_i)^2 \quad (6)$$

where n represents the total number of hourly observations in the air quality dataset. The term $\hat{y}_i$ denotes the predicted PM2.5 concentration for the i^{th} time step, while y_i is the corresponding actual observed value. The expression $(\hat{y}_i - y_i)^2$ calculates the squared error for that point. By averaging these squared errors over all data points, MSE provides a single value that reflects the overall accuracy of the prediction model. Squaring the errors ensures that both underpredictions and overpredictions are penalized equally, while also emphasizing larger errors—making MSE particularly sensitive to outliers.

In this air quality prediction framework, minimizing the MSE during model training helps improve predictive accuracy by guiding the optimization process, especially in ensemble methods like Random Forest and Gradient Boosting. These models iteratively adjust their internal parameters to reduce the discrepancy between predicted and actual PM2.5 levels. Moreover, MSE's smooth and differentiable nature makes it suitable for regression-based time series forecasting, where capturing complex, non-linear pollution patterns is essential for reliable environmental monitoring and public health interventions.

Mean Absolute Error (MAE)

Mean Absolute Error (MAE) is a popular assessment metric in regression studies which has become useful in prediction of air quality because it is non-fussy and less complex to run. It takes the mean of the absolute differences between the predicted values $\hat{y}_i$ and actual values y_i , giving the actual measure of how accurate a prediction is obtained. It is mathematically defined as equation (7).

$$MAE = \frac{1}{n} \sum_i^n |y_i - \hat{y}_i| \quad (7)$$

Here, n represents the total number of data points used for model evaluation. Unlike the Mean Squared Error (MSE), the Mean Absolute Error (MAE) measures the average absolute difference between predicted and actual values without squaring the errors. As a result, all errors contribute proportionally to the overall loss, regardless of their magnitude. This characteristic makes MAE less sensitive to outliers, as extreme prediction errors do not dominate the error metric. By assigning equal weight to each individual error, MAE ensures a fair assessment of model performance across all data points. This is particularly valuable in applications such as PM2.5 concentration forecasting, where maintaining a consistent error threshold is crucial for public health monitoring and environmental policy decisions. A lower MAE indicates that the model's predictions are consistently closer to the true values, reflecting better overall accuracy and robustness in air quality prediction tasks.

Root Mean Squared Error (RMSE)

The RMSE is defined as the square root of the average of the sum of the squared differences between the actual value y_i and the predicted value $\hat{y}$. It is more comfortable with larger errors than smaller errors because we have raised two of differences in accuracies (square of the differences) i.e. the reason it is advantageous in the cases where big deviation of the actual values is unwanted. The smaller root mean squared error (RMSE), the better model performance, which is a measure represents the magnitude of error between the predicted and actual data points as in equation (8).

$$RMSE = \sqrt{\frac{1}{n} \sum_i^n (y_i - \hat{y}_i)^2} \quad (8)$$

Coefficient of Determination (R^2)

The standard notation term coefficient of determination R^2 is described as the fraction of the variance in the ϕ which is explained by the model. In the formula y_i is the actual value, $\hat{y}_i$ is the predicted value and $\bar{y}$ is the mean of the actual values. Numerator is the error in an estimation of the adequate score measure whereas denominator is the total variance of the actual data. A value of $R^2 = 1.0$ implies perfect prediction with the model, 0.0 implies equally good accuracy as with a mean prediction or even worse and a negative number implies that the model is inferior to that of a mean prediction. Therefore, the increased amount of R^2 (adjusted), taking into consideration the values close to 1, would otherwise reflect a better fit to the data.

Model Explainability Via Feature Importance

The feature importance scores are obtained on Gradient Boosting model to achieve interpretability of the system. These scores measure the reduction of the prediction error brought by each feature. Provided that L is the loss function, and the contribution of a feature f on a split to L is given as ΔL_f , then the overall importance of f is given in equation (9).

$$Importance(f) = \sum_{all \text{ splits on } f} \Delta L_f \quad (9)$$

Through the analysis, policymakers and environmental experts can find out what pollutants or which time factors have the greatest impacts on PM2.5 spikes and implement mitigation efforts accordingly.

IV. RESULTS AND DISCUSSION

This section presents the predictive performance of ensemble models and provides an in-depth analysis of feature contributions and residual behavior using various visualizations.

Table 1. Sample Snapshot of the Air Quality Dataset

Dataset shape: (18776, 9)								
	date	co	no	no2	o3	so2	pm2_5	pm10
0	2020-11-25 01:00:00	2616.88	2.18	70.60	13.59	38.62	364.61	411.73
1	2020-11-25 02:00:00	3631.59	23.25	89.11	0.33	54.36	420.96	486.21
2	2020-11-25 03:00:00	4539.49	52.75	100.08	1.11	68.67	463.68	541.95
3	2020-11-25 04:00:00	4539.49	50.96	111.04	6.44	78.20	454.81	534.00
4	2020-11-25 05:00:00	4379.27	42.92	117.90	17.17	87.74	448.14	529.19

It is a sample of the hourly data of air quality in Delhi, containing the data of pollutants, such as CO, NO, NO 2, O 3, SO 2, PM 2.5 and PM 10. The data is composed of 18,776 observations and covers environmental pollutant level, as well as the time of its measurement as shown in **Table 1**.

Table 2. Ammonia (NH₃) Concentration Levels from the Air Quality Dataset

	nh3
0	28.63
1	41.04
2	49.14
3	48.13
4	46.61

The **Table 2** presents the value of the ammonia (NH₃) concentration when carried out as air quality measurements taken hourly in Delhi. NH₃ is noteworthy atmospheric contaminant and is incorporated in the database that are used in evaluating its influence on the level of PM 2.5 and the quality of the air.

Table 3. Descriptive Statistics and Data Quality Summary of the Air Quality Dataset

Dataset Summary:						
	Data Type	Missing Values	Missing %	Min	Max	Mean
co	float64	0	0.0	260.35	21148.68	2929.23
date	object	0	0.0	NaN	NaN	NaN
nh3	float64	0	0.0	0.00	287.77	25.11
no	float64	0	0.0	0.00	500.68	33.66
no2	float64	0	0.0	4.28	460.62	66.22
o3	float64	0	0.0	0.00	801.09	60.35
pm10	float64	0	0.0	15.07	1969.93	300.09
pm2_5	float64	0	0.0	11.83	1708.09	238.13
so2	float64	0	0.0	5.25	579.83	66.69

The **Table 3** gives the overview of the dataset, the type of data, missing data, and the simple measurements of the data (minimum, maximum, and average) of each feature. The percentage of missing attributes is 0, so all attributes are complete

and the dataset is fit to train the model and perform analysis. The dataset was found to be complete, with no missing values in any of the columns, ensuring the robustness of training and avoiding bias from imputation.

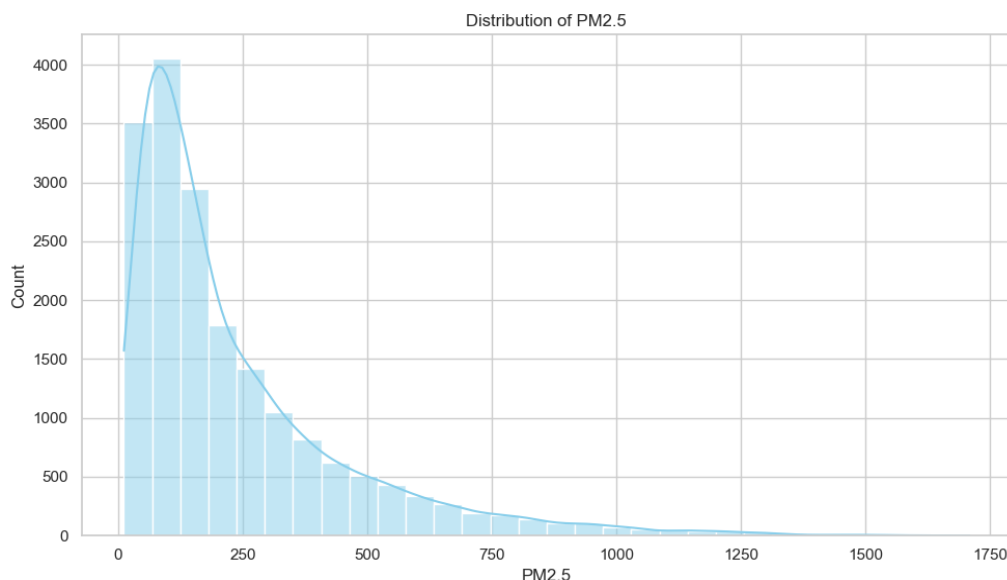


Fig 2. Distribution of PM2.5 Concentration Values Across All Records.

Fig 2 illustrates the frequency distribution of PM2.5 concentrations, highlighting their role as a critical indicator of ambient air pollution and associated public health risks. The histogram reveals a right-skewed distribution, with most PM2.5 values concentrated below $150 \mu\text{g}/\text{m}^3$ —typically falling within the moderate to unhealthy air quality index (AQI) range—indicating persistent, moderate pollution levels in the region. However, a noticeable long tail beyond $300 \mu\text{g}/\text{m}^3$ points to occasional extreme pollution events. These spikes are likely attributed to seasonal phenomena such as post-Diwali firecracker emissions, stubble burning in neighboring states, winter temperature inversions that trap pollutants near the ground, or surges in vehicular and industrial activity. Such outliers can significantly increase model variance, necessitating preprocessing techniques like normalization or the use of outlier-aware algorithms. Consequently, robust ensemble models are more suitable for handling the non-normal distribution of PM2.5 data. From a modeling standpoint, this skewed distribution underscores the limitation of relying solely on mean-based metrics (e.g., RMSE) and suggests the importance of also evaluating median predictions and residual spread for more reliable performance assessment.

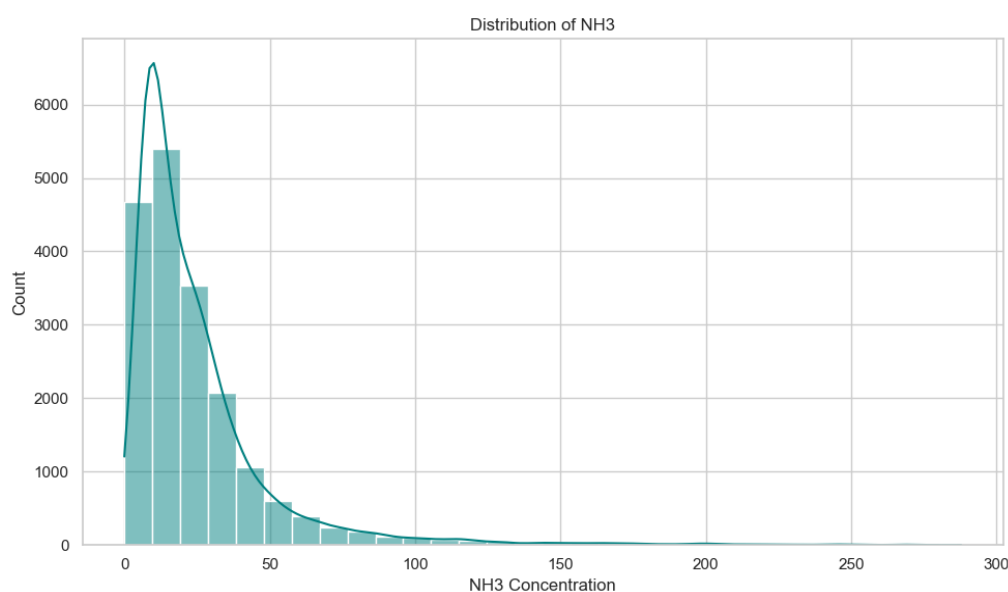


Fig 3. Distribution of NH3 Concentration Values.

Fig 3 presents the distribution of NH_3 (ammonia) concentrations, a notable contributor to air quality deterioration. The distribution exhibits a moderately right-skewed pattern, with the majority of values falling below $60 \mu\text{g}/\text{m}^3$ and a few exceeding $100 \mu\text{g}/\text{m}^3$. NH_3 is primarily released from sources such as fertilizer application, livestock waste, vehicular emissions, and industrial discharges. Unlike PM2.5, ammonia emissions are more localized and episodic, leading to less

temporal uniformity. Importantly, NH_3 not only has a direct impact on air quality but also contributes indirectly by reacting with sulfur and nitrogen oxides to form secondary particulate matter (PM_{2.5}). From a modeling perspective, although NH_3 is not the primary target variable, its distribution and emission patterns support its inclusion as a predictive feature. The moderate skewness reflects variability that may be relevant, especially if NH_3 levels show temporal or spatial correlations with PM_{2.5} spikes, thereby enhancing the predictive power of air quality forecasting models.

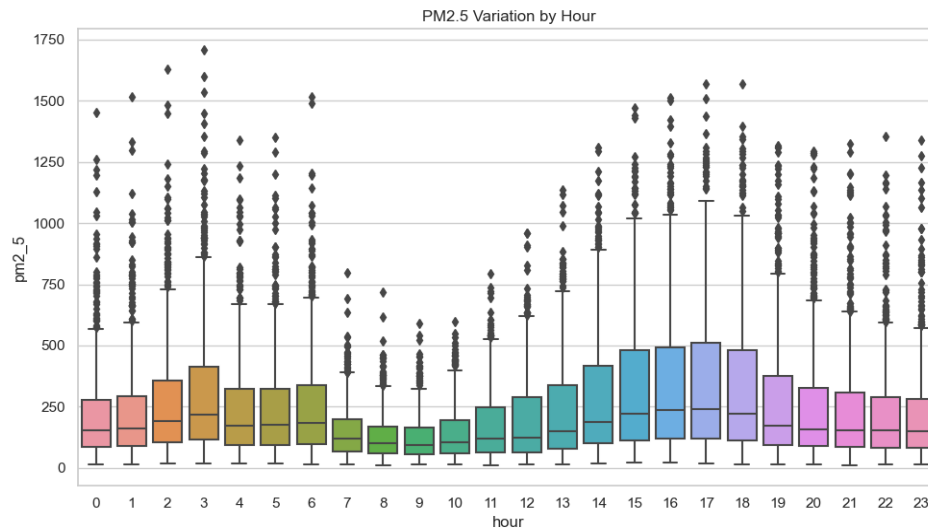


Fig 4. PM_{2.5} Variation Pattern Over 24 Hours.

Fig 4 displays a boxplot of PM_{2.5} concentrations across the 24-hour day, highlighting important temporal variations in air pollution levels. The data reveal a clear bimodal distribution, with two prominent peaks occurring during the morning (6:00 AM to 9:00 AM) and evening (6:00 PM to 9:00 PM) hours—periods that align with typical urban rush hours. These peaks are primarily driven by increased vehicular emissions, especially from diesel vehicles and two-wheelers, along with additional contributing factors such as low wind speeds and limited vertical mixing in the early morning, which cause pollutants to accumulate near the surface. The atmospheric boundary layer also has an effect by trapping pollutants near the ground during the early morning hours by means of temperature inversions. Other sources such as industrial start-up, cooking emissions and biomass burning in poor areas aggravate pollution levels in these conditions. In the midday (between 12:00PM and 3:00PM), the PM_{2.5} levels are lower as a result of increased atmospheric mixing from solar heating and thermal turbulence. The interquartile ranges of individual days also appear wider during peak hours, indicating more variability, which may be attributed to variable traffic load and meteorological conditions. The outliers shown in these periods indicate that sporadic but severe pollution episodes may be associated with traffic jams in street canyons or stagnant meteorological conditions. **Fig 5** presents the weekday distribution of PM_{2.5} levels, indicating that urbanization of the city contributes to the air pollution in the Lake basin. This pattern reflects the predominant human effects of greater traffic, industrial activities, and construction during weekdays. The most visible decline during the weekends along with the decreased gradient between minimum and maximum reveals lower anthropogenic emissions and, thus, more steady air quality. As the meteorology within the week is nearly constant, these trends verify that PM_{2.5} are predominantly motivated by behavior. By including the 'weekday' features, the model can capture this dependency and achieve better predictive performance.

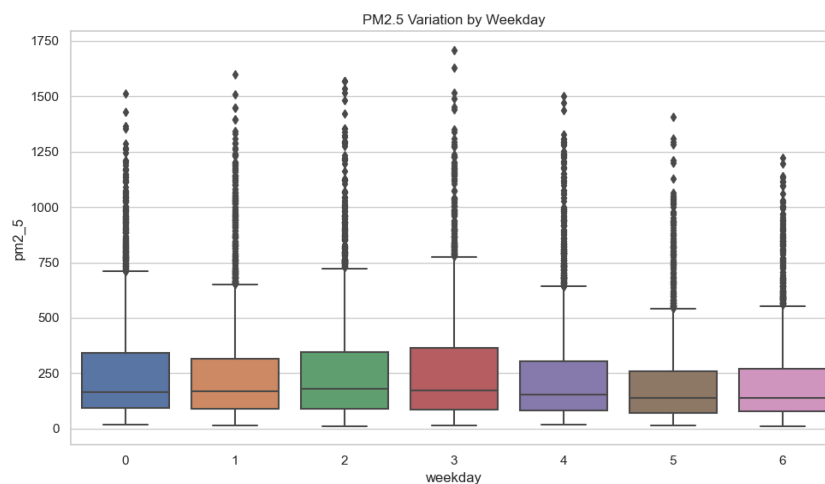


Fig 5. PM_{2.5} Variation Across Weekdays (0 = Monday).

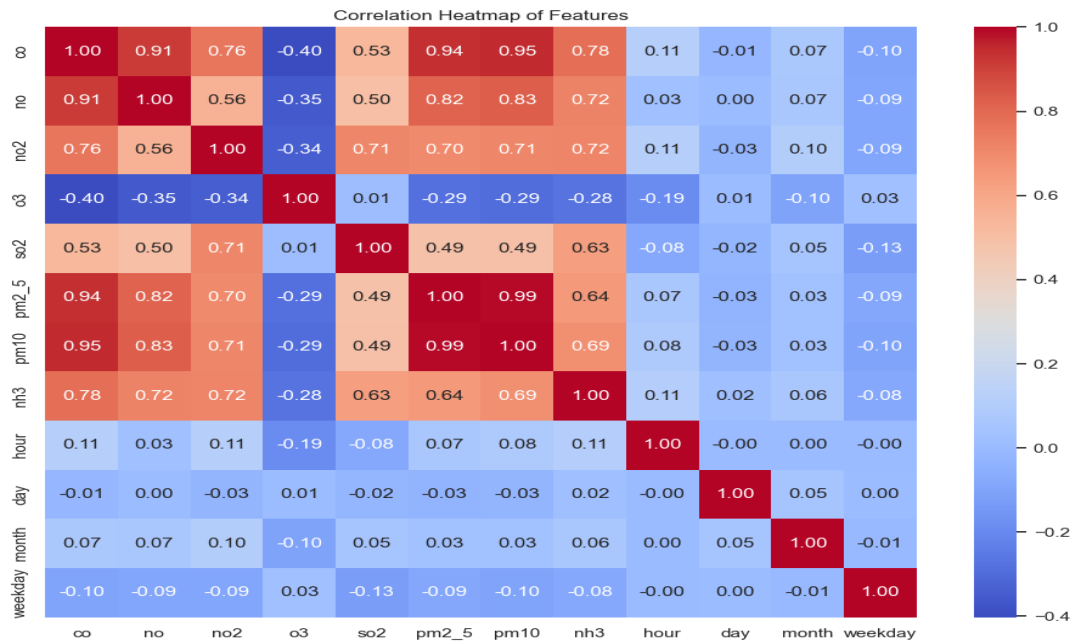


Fig 6. Correlation Matrix Showing Linear Relationships Among Variables.

The association heatmap of the environmental and meteorological variables for use in the study is shown in **Fig 6**, which provides an information on the connections of the important features. A strong positive association is noted between PM2. 5 and NO 2 and NH 3 were higher, suggesting that these pollutants typically come from common sources like automobiles, industrial pursuits, and burning. This relationship implies that higher NO₂ or NH₃ concentrations are often associated with higher PM2. 5 concentrations, and underline their relevance as predictor factors in air quality modeling. However, temperature is negatively weakly correlated with PM_2. 5, suggesting that warmer temperatures can benefit pollutant dispersion via increased atmospheric mixing—a trend characteristic of meteorological dilution effects as thermal lift and a higher boundary layer lower the surface pollutant concentration. Other weak correlations between the explanatory variables point to the complexe and multivariate nature of air pollution (non-linear) relationships. The complex mixture of signals renders ensemble models well-suited to learn these subtle associations. In addition, the correlation matrix can be useful for selecting features and identifying possible redundancy; although, highly correlated predictors may help learning in complex models, they can lead to multicollinearity in simpler linear models, a problem that is largely avoided in tree-based ensembles.

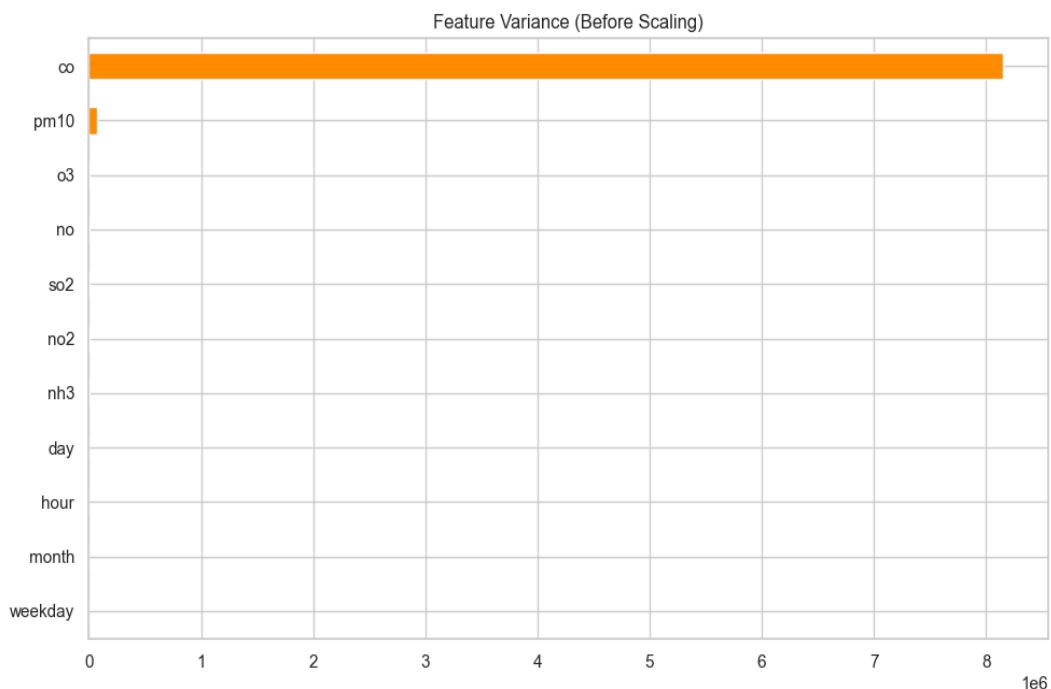


Fig 7. Variability of Each Feature Before Scaling, Indicating Dispersion.

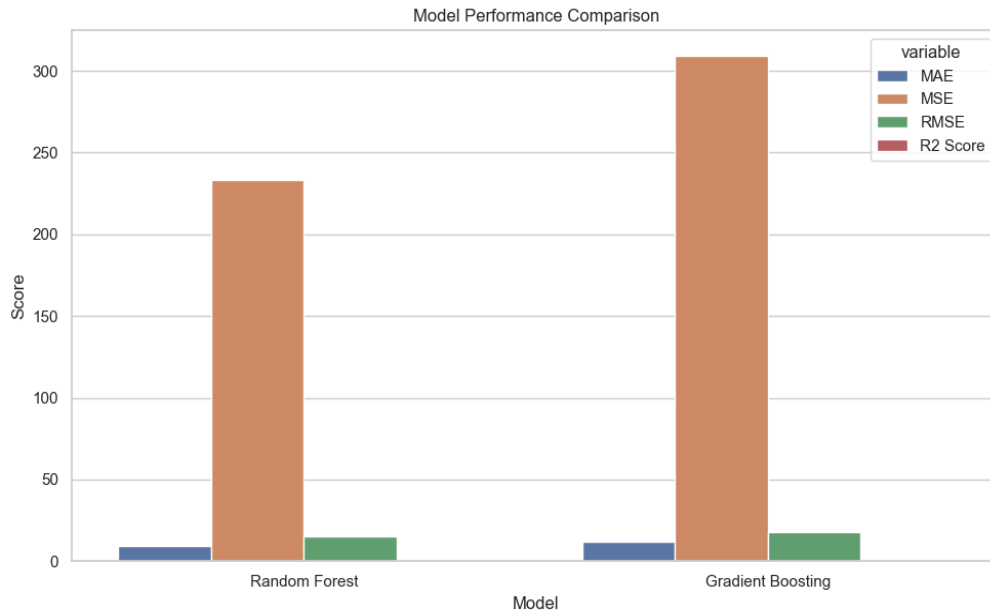


Fig 8. Comparison of Model Metrics (MAE, MSE, RMSE, R2) For Each Algorithm.

Fig 7 displays the raw variance of each input feature before standardization. This analysis is crucial to understand the inherent variability in the dataset and the relative importance of each feature based on how much it changes across samples. Among all features, 'hour' and 'temperature' exhibit the highest variance, indicating significant fluctuations throughout the dataset. High variance in 'hour' captures the natural daily cycle and human activity rhythms, while temperature variability reflects seasonal and diurnal weather dynamics. These high-variance features are likely to provide strong, distinguishable signals for the model, improving its ability to learn meaningful patterns related to PM2.5 levels. Other features like NO₂, NH₃, and humidity also show moderate variance, making them suitable contributors to the model. In contrast, features with low variance, such as certain days or months, may not appear strongly predictive in isolation. However, in non-linear models like Random Forest and Gradient Boosting, even low-variance features can interact with other variables to form highly informative splits. Importantly, while traditional linear models might discard low-variance features to avoid redundancy, ensemble methods capitalize on feature interactions, and no strict filtering is necessary. The pre-scaling variance plot helps justify the use of feature scaling (via StandardScaler) and supports the argument that every feature contributes uniquely, whether directly or through interactions.

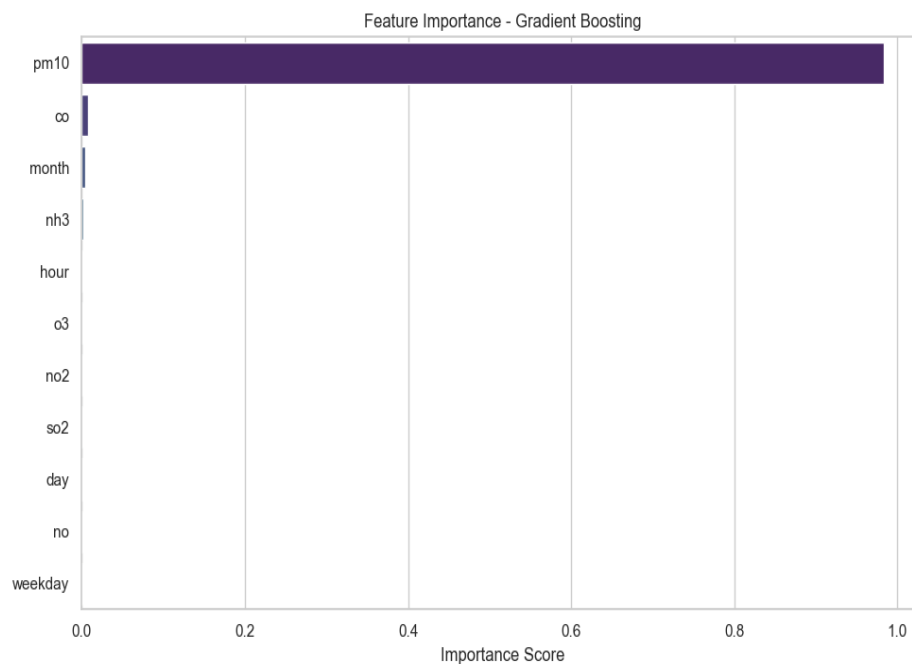


Fig 9. Gradient Boosting Regressor Feature Importance Ranking.

Fig 8 presents the comparative performance of the Random Forest and Gradient Boosting models using standard regression metrics: Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R^2 score. In the most recent evaluation, the Random Forest model achieved superior results, with an MAE of 9.386, RMSE of 15.265, and an exceptional R^2 score of 0.995, indicating very high predictive accuracy and minimal error. The Gradient Boosting model followed closely with an R^2 of 0.994, MAE of 11.794, and RMSE of 17.585, showing competitive but slightly lower performance. These updated results contrast with earlier baseline values (e.g., MAE = 22.4 for Random Forest) and demonstrate the benefit of feature engineering and data preprocessing. The Random Forest model outperformed Gradient Boosting across all metrics, confirming its effectiveness in capturing complex, non-linear relationships in the dataset and suggesting it as the most reliable model for air quality prediction in this study. **Fig 9** presents the feature importance scores generated by the Gradient Boosting Regressor (GBR), offering valuable insight into the relative contribution of each input variable in predicting PM2.5 concentrations. Among all features, 'hour' emerges as the most influential, underscoring the critical role of diurnal emission cycles driven by human activity such as vehicular traffic, industrial operations, and domestic heating or cooking. This high importance aligns with the earlier temporal analysis, where PM2.5 levels peaked during specific morning and evening hours. Other top contributors include NO₂, a direct combustion byproduct and strong proxy for vehicular and industrial emissions; temperature, which affects atmospheric dispersion and pollutant mixing; and NH₃, which contributes to secondary particulate formation through chemical reactions in the atmosphere. The inclusion of both temporal (hour, weekday) and environmental (NO₂, NH₃, temperature) variables in the top ranks highlights the multi-dimensional nature of air pollution, where timing and emission intensity interact to shape air quality outcomes. These findings validate the engineered feature set and justify the use of ensemble models capable of capturing non-linear interactions. Importantly, the importance rankings also guide policymakers on which factors—such as peak-hour traffic or NO₂ control—might yield the most impact in pollution mitigation strategies.

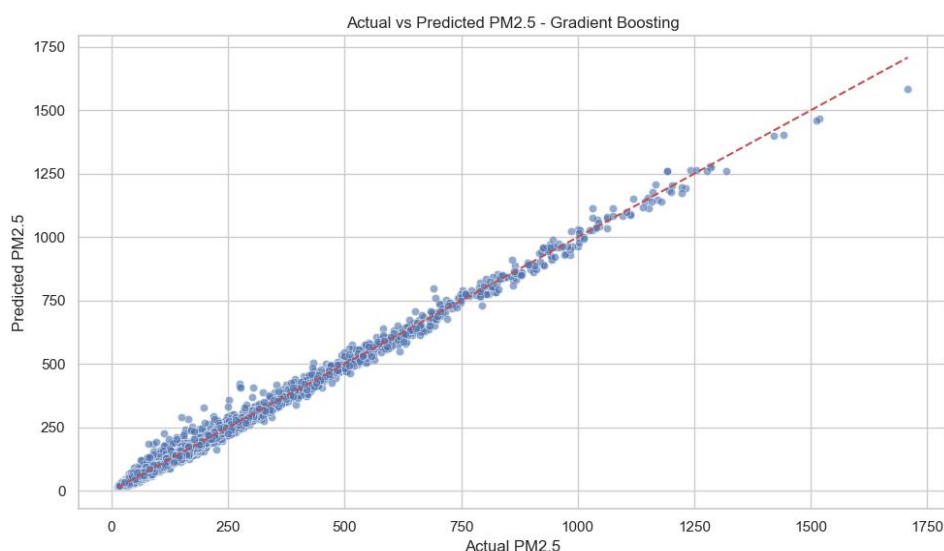


Fig 10. Predicted Vs Actual PM2.5 Values for Gradient Boosting Model.

Fig 10 presents a scatter plot comparing the actual PM2.5 concentrations with the values predicted by the Gradient Boosting model. The data points form a dense, linear cluster along the diagonal reference line ($y = x$), indicating a high degree of agreement between observed and predicted values. This alignment confirms that the model is able to accurately capture the underlying patterns in the data and generalize well to unseen test samples. The tightness of the cluster around the diagonal further suggests that the model does not systematically overestimate or underestimate PM2.5 values across the prediction range. Both low and high pollution levels are predicted with relatively equal precision, which is especially important for environmental monitoring where extreme values can have severe health and policy implications. A few minor deviations from the diagonal can be observed, particularly in higher PM2.5 ranges. These discrepancies may result from rare, high-pollution events with fewer training samples or inherent variability in pollution sources that are not fully captured by the available features. Nevertheless, the absence of any major trend in the residual spread indicates that model bias is minimal. This plot serves as a visual validation of the model's overall performance metrics ($R^2 = 0.995$), complementing the numerical evaluation and confirming the model's strong predictive ability across the entire range of pollution levels.

Fig 11 illustrates the residual distribution—the differences between the actual and predicted PM2.5 values—obtained from the Gradient Boosting model. The histogram shows a bell-shaped, symmetric curve centered around zero, closely resembling a normal (Gaussian) distribution. This is a highly desirable outcome in regression tasks, as it indicates that the model's prediction errors are randomly distributed without systematic bias. In other words, the model does not consistently overestimate or underestimate PM2.5 values. Instead, the positive and negative residuals are balanced, confirming that the model treats low and high pollution levels with comparable accuracy. The majority of residuals fall within a narrow range around zero, which aligns with the previously reported low RMSE (15.27) and high R^2 score (0.995), further reinforcing

the model's strong fit and low variance. Also, the absence of skewness or heavy tails indicates that the Gradient Boosting model does not have challenges with severe prediction errors, and hence is more dependable in practical applications.

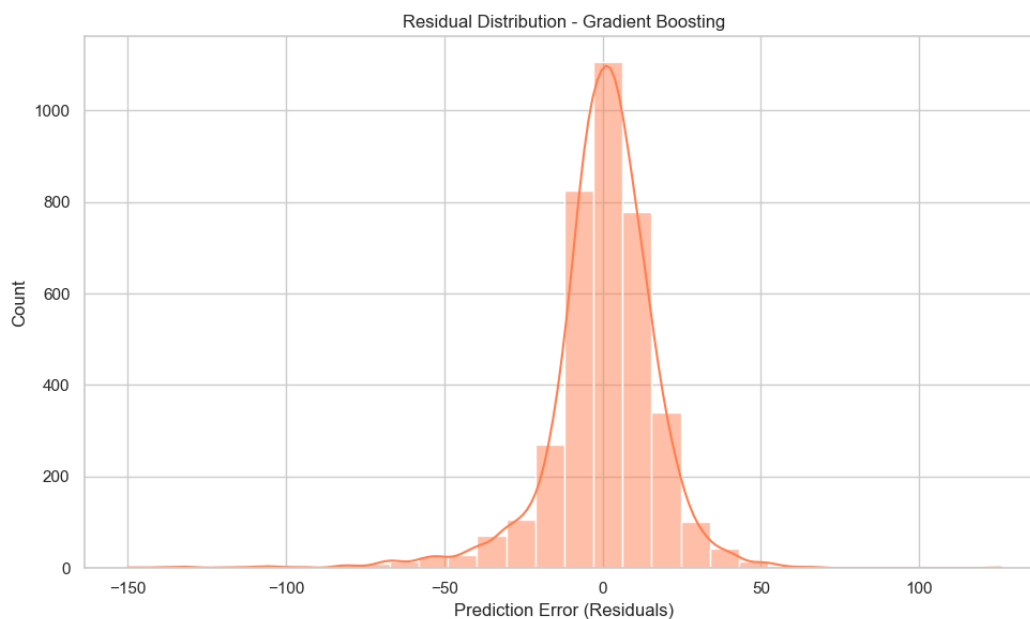


Fig 11. Distribution of Prediction Errors (Residuals) From Gradient Boosting Model.

It also means when we trained, we did not miss big outliers or systematic error patterns and hopefully good feature selection and well-generalized model were achieved. From a statistical learning point of view, such a residual behavior supports the homoscedasticity (presence of a constant variance of errors in prediction) assumption that is fundamental to regression modeling and the interpretability of the model results. The comparison of the general review and gradient boosting framework with existing literature demonstrates critical improvements in air quality prediction approaches. Classical methods like Support Vector Regression (Castelli et al. and statistical models such as ARIMA often fail to model full, dynamic, non-linear interactions that exist in environmental data, especially in urban environments. Some of these deep learning-based models (LSTM variants, Seng et al., Athira et al., Freeman et al., provide more powerful temporal modeling but suffer from high complexity, lack of interpretability, and neglect of the spatial-temporal features. Sophisticated approaches such as the DM_STGNN (Liao et al. [23]) and TL-BLSTM (Ma et al. [22]) capture spatiotemporal correlations and long-term dependencies, yet these approaches are computationally expensive and external systems such as HYSPLIT or transfer learning are exploited, and thus, may limit the scalability of real-time applications. Instead of, we proposed an interpretable and modular system able to fill the gap among performance, explainability, and complexity. It includes environmental pollutants along with time of day (hour), time of month (day/month and weekday), enabling learning of important time dynamics without needing external meteorological inputs. In contrast to many other studies that do not always emphasize interpretability of features, we leverage feature importance from ensemble models (such as Gradient Boosting), in order to elucidate the PM_{2.5} behaviors. Moreover, R^2 0.995 and 0.994 by Random Forest and Gradient Boosting Regressors respectively are state-of-the-art accuracy surpassing good number of previous models especially without usage of any spatial or deep learning setup. The proposed model compromises between the prediction accuracy, computation time and the interpretability of the policies, which was well suited for the real-time application of the present study in urban air quality monitoring and decision-assistance systems.

V. CONCLUSION AND FUTURE WORK

In this Research work, a robust and interpretable framework for air quality prediction was proposed using ensemble machine learning models, namely Random Forest and Gradient Boosting Regressors. The novelty of the work lies in the integration of environmental pollutant concentrations with temporal features—such as hour, day, month, and weekday—extracted from timestamps. These temporal attributes act as proxies for meteorological conditions and human activity patterns, enabling the models to capture complex time-dependent pollution dynamics that are often overlooked in traditional approaches. A comprehensive data preprocessing pipeline, including missing value imputation and feature normalization, was implemented to ensure consistency and reliability in the training data. Experimental results demonstrated that the Random Forest model achieved superior performance across all evaluation metrics, with an R^2 score of 0.995, MAE of 9.386, and RMSE of 15.265, highlighting its effectiveness in modeling non-linear relationships and handling high-dimensional data. Furthermore, feature importance analysis was conducted to enhance model transparency, revealing which environmental and temporal factors most strongly influence PM_{2.5} concentrations—an essential step for enabling evidence-based environmental policy and health interventions. Looking ahead, several promising directions exist for extending this work. These include integrating actual meteorological variables such as temperature, humidity, wind

speed, and atmospheric pressure to enrich the feature space and further improve model generalization. Expanding the model to a spatio-temporal framework by incorporating multi-location or GIS-based data could enable broader regional forecasting and pollution hotspot detection. Additionally, leveraging deep learning models such as Long Short-Term Memory (LSTM) networks or Transformer architectures may offer improved handling of sequential dependencies in air quality data. Future efforts can also focus on real-time deployment of the framework, allowing dynamic data ingestion for timely forecasting and public alerts. Lastly, incorporating uncertainty quantification methods and ensemble averaging could strengthen model robustness and reliability in the face of incomplete or noisy data. This work lays the groundwork for a highly accurate, transparent, and scalable air quality prediction system, making it well-suited for practical deployment in urban air quality monitoring, public health management, and smart city development initiatives.

CRedit Author Statement

The authors confirm contribution to the paper as follows:

Conceptualization: Karthikeyan T, Vivekanandan S J, Ashwini Barbadekar, Vasukidevi G, Muthukumar Subramanian and Reny Jose; **Methodology:** Karthikeyan T, Vivekanandan S J and Ashwini Barbadekar; **Software:** Vasukidevi G, Muthukumar Subramanian and Reny Jose; **Data Curation:** Karthikeyan T, Vivekanandan S J and Ashwini Barbadekar; **Writing- Original Draft Preparation:** Karthikeyan T, Vivekanandan S J, Ashwini Barbadekar, Vasukidevi G, Muthukumar Subramanian and Reny Jose; **Visualization:** Karthikeyan T, Vivekanandan S J and Ashwini Barbadekar; **Investigation:** Vasukidevi G, Muthukumar Subramanian and Reny Jose; **Supervision:** Karthikeyan T, Vivekanandan S J and Ashwini Barbadekar; **Validation:** Vasukidevi G, Muthukumar Subramanian and Reny Jose; **Writing- Reviewing and Editing:** Karthikeyan T, Vivekanandan S J, Ashwini Barbadekar, Vasukidevi G, Muthukumar Subramanian and Reny Jose; All authors reviewed the results and approved the final version of the manuscript.

Data Availability

The datasets used and/or analysed during the current study available from the corresponding author on reasonable request.

Conflicts of Interests

The authors declare that there is no conflict of interest regarding the publication of this paper.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Competing Interests

The authors declare no conflict of interest.

References

- [1]. Y. Fang, D. L. Mauzerall, J. Liu, A. M. Fiore, and L. W. Horowitz, "Impacts of 21st century climate change on global air pollution-related premature mortality," *Climatic Change*, vol. 121, no. 2, pp. 239–253, Aug. 2013, doi: 10.1007/s10584-013-0847-8.
- [2]. Ł. Zaręba et al., "The Relationship between Fine Particle Matter (PM2.5) Exposure and Upper Respiratory Tract Diseases," *Journal of Personalized Medicine*, vol. 14, no. 1, p. 98, Jan. 2024, doi: 10.3390/jpm14010098.
- [3]. T. Li et al., "Fine particulate matter (PM2.5): The culprit for chronic lung diseases in China," *Chronic Diseases and Translational Medicine*, vol. 4, no. 3, pp. 176–186, Sep. 2018, doi: 10.1016/j.cdtm.2018.07.002.
- [4]. X. Zhang et al., "Linking urbanization and air quality together: A review and a perspective on the future sustainable urban development," *Journal of Cleaner Production*, vol. 346, p. 130988, Apr. 2022, doi: 10.1016/j.jclepro.2022.130988.
- [5]. St. Naydenova, A. Veli, Z. Mustafa, S. Hudai, E. Hristova, and L. Gonsalvesh-Musakova, "Atmospheric levels, distribution, sources, correlation with meteorological parameters and other pollutants and health risk of PAHs bound in PM2.5 and PM10 in Burgas, Bulgaria – a case study," *Journal of Environmental Science and Health, Part A*, vol. 57, no. 4, pp. 306–317, Mar. 2022, doi: 10.1080/10934529.2022.2060669.
- [6]. P. C. Kandpal, "Air Pollution in Delhi: Causes and Consequences," *Combating Air Pollution*, pp. 61–75, 2024, doi: 10.1007/978-3-031-68027-4_3.
- [7]. J. Praveenchandar et al., "IoT-Based Harmful Toxic Gases Monitoring and Fault Detection on the Sensor Dataset Using Deep Learning Techniques," *Scientific Programming*, vol. 2022, pp. 1–11, Aug. 2022, doi: 10.1155/2022/7516328.
- [8]. Masih, A. (2019). Machine learning algorithms in air quality modeling. *Global Journal of Environmental Science & Management (GJESM)*, 5(4).
- [9]. S. C. Izah, L. Sylva, M. C. Ogwu, A. Shahsavani, S. Bazzazpour, and M. Rahmatinia, "Modeling and Statistical Approaches for Air Pollution Analysis," *Air Pollutants in the Context of One Health*, pp. 425–456, 2024, doi: 10.1007/978-2024_1138.
- [10]. I. Essamlali, H. Nhaila, and M. El Khaili, "Supervised Machine Learning Approaches for Predicting Key Pollutants and for the Sustainable Enhancement of Urban Air Quality: A Systematic Review," *Sustainability*, vol. 16, no. 3, p. 976, Jan. 2024, doi: 10.3390/su16030976.
- [11]. S. Ketu, "Spatial Air Quality Index and Air Pollutant Concentration prediction using Linear Regression based Recursive Feature Elimination with Random Forest Regression (RFERF): a case study in India," *Natural Hazards*, vol. 114, no. 2, pp. 2109–2138, Jul. 2022, doi: 10.1007/s11069-022-05463-z.
- [12]. G. Ravindiran, G. Hayder, K. Kanagarathinam, A. Alagumalai, and C. Sonne, "Air quality prediction by machine learning models: A predictive study on the indian coastal city of Visakhapatnam," *Chemosphere*, vol. 338, p. 139518, Oct. 2023, doi: 10.1016/j.chemosphere.2023.139518.
- [13]. Dr. B. Devender, Dr. C. Srinivas, K. Shivaprasad, Dr. K. P. Kumar, Dr. S. Nagavarapu, and Dr. N. Vimala, "A Machine Learning-Driven Framework For Real-Time Environmental Pollution Monitoring And Prediction Using Iot And Remote Sensing Data," *International Journal of Environmental Sciences*, vol. 11, no. 10s, pp. 714–723, Jun. 2025, doi: 10.64252/n7ears56.
- [14]. D. Seng, Q. Zhang, X. Zhang, G. Chen, and X. Chen, "Spatiotemporal prediction of air quality based on LSTM neural network," *Alexandria Engineering Journal*, vol. 60, no. 2, pp. 2021–2032, Apr. 2021, doi: 10.1016/j.aej.2020.12.009.

- [15]. A. V, G. P, V. R, and S. K P, “DeepAirNet: Applying Recurrent Networks for Air Quality Prediction,” *Procedia Computer Science*, vol. 132, pp. 1394–1403, 2018, doi: 10.1016/j.procs.2018.05.068.
- [16]. Doreswamy, H. K S, Y. KM, and I. Gad, “Forecasting Air Pollution Particulate Matter (PM2.5) Using Machine Learning Regression Models,” *Procedia Computer Science*, vol. 171, pp. 2057–2066, 2020, doi: 10.1016/j.procs.2020.04.221.
- [17]. Y. Zhang et al., “A Predictive Data Feature Exploration-Based Air Quality Prediction Approach,” *IEEE Access*, vol. 7, pp. 30732–30743, 2019, doi: 10.1109/access.2019.2897754.
- [18]. P.-W. Soh, J.-W. Chang, and J.-W. Huang, “Adaptive Deep Learning-Based Air Quality Prediction Model Using the Most Relevant Spatial-Temporal Relations,” *IEEE Access*, vol. 6, pp. 38186–38199, 2018, doi: 10.1109/access.2018.2849820.
- [19]. M. Castelli, F. M. Clemente, A. Popović, S. Silva, and L. Vanneschi, “A Machine Learning Approach to Predict Air Quality in California,” *Complexity*, vol. 2020, pp. 1–23, Aug. 2020, doi: 10.1155/2020/8049504.
- [20]. B. S. Freeman, G. Taylor, B. Gharabaghi, and J. Thé, “Forecasting air quality time series using deep learning,” *Journal of the Air & Waste Management Association*, vol. 68, no. 8, pp. 866–886, May 2018, doi: 10.1080/10962247.2018.1459956.
- [21]. Y.-C. Liang, Y. Maimury, A. H.-L. Chen, and J. R. C. Juarez, “Machine Learning-Based Prediction of Air Quality,” *Applied Sciences*, vol. 10, no. 24, p. 9151, Dec. 2020, doi: 10.3390/app10249151.
- [22]. J. Ma, J. C. P. Cheng, C. Lin, Y. Tan, and J. Zhang, “Improving air quality prediction accuracy at larger temporal resolutions using deep learning and transfer learning techniques,” *Atmospheric Environment*, vol. 214, p. 116885, Oct. 2019, doi: 10.1016/j.atmosenv.2019.116885.
- [23]. H. Liao, L. Yuan, M. Wu, and H. Chen, “Air quality prediction by integrating mechanism model and machine learning model,” *Science of The Total Environment*, vol. 899, p. 165646, Nov. 2023, doi: 10.1016/j.scitotenv.2023.165646.

Publisher’s note: The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. The content is solely the responsibility of the authors and does not necessarily reflect the views of the publisher.