VAWKUM Transactions on Computer Sciences
http://vfast.org/journals/index.php/VTCS@ 2025, ISSN(e): 2308-8168, ISSN(p): 2411-6335
Volume 13, Number 1, January-June 2025 *pp:106-119*

Harnessing Machine Learning for Accurate Smog Level Prediction: A Study of Air Quality in India

Sahil Jatoi [1*], **Bushra Abro** [2], **Sanam Narejo** [1], **Yaqoob Ali Baloch** [2], **Kehkashan Asma** [2]

[1]Department of Computer Systems Engineering, Mehran University of Engineering and Technology Jamshoro 76062, Sindh, Pakistan; [2]Department of Electronic Engineering, Mehran University of Engineering and Technology Jamshoro 76062, Sindh, Pakistan

Keywords: *Air Quality Index (AQI), Machine Learning Models, Smog Prediction, Automation and Environmental Monitoring.*

Journal Info:
Submitted:
March 02, 2025
Accepted:
April 26, 2025
Published:
May 06, 2025

Abstract

Accurate prediction of smog concentrations is needed to mitigate the harm of AP on public health and the environment. This research proposes a new method to combine machine learning (ML) models with live data from Central Pollution Control Board (CPCB) to fill in the smog prediction accuracy gaps. The data consist of hourly AQI readings from different towns in India which were preprocessed to adjust for missing values and normalize data before ML models. The algorithms were tested with 8 ML algorithms, and hyper-parameter settings were tuned using the GridSearchCV method. The results show that XG Boost Regressor (XGBR) and Extra Tree Regressor (ETR) models significantly surpass other ML algorithms and traditional techniques with better accuracy on predicting smog. These results are useful for policymakers and environmental agencies to implement sustainable air quality management.

***Correspondence author email address:** sahiljatoi744@gmail.com
DOI: 10.21015/vtcs.v13i1.2077

1 Introduction

Environmental issues especially related to smog are a prime challenge that urban areas are facing in 21st century. The aim of this study is to develop and evaluate machine learning models for predicting smog levels in India using real-time data from the Central Pollution Control Board (CPCB). The objectives of this study are: To compare the performance of several machine learning models, to identify the key features affecting air quality predictions, and to apply hyperparameter tunning techniques to improve model performance. The consequences of harmful substances affect badly the city inhabitants and even their morale. The research written by [1], tries to demonstrate the detrimental impact that smog has on health, such as breathing diseases, cardiac problems, and shortened

life expectancy.

The World Health Organization (WHO) indicate these observations that suggests 4.2 million deaths annually due to AP, with smog being one of the most problematic sources, are often premature and preventable [2]. It's irrefutable that smog and poor AQ represent a serious health risk, especially in urban contexts. Many studies have shown that an increase in AP correlates with worsening health outcomes, including respiratory and cardiovascular diseases [3]. While urbanization and industrialization have tremendous potential for positive global development and growth. Rapid industrialization has made access to clean air burdensome for many, primarily in developing countries like India where poor AQ is a consequence of the region efforts to rapidly urbanize and modernize. To mitigate these pollution trends and predict smog formation more proactively, the Indian government has been actively monitoring AQ with open data for years, through its CPCB platform. This includes daily updates on key air pollutants, providing information that's crucial for modeling, estimating, predicting the impacts of AQ on human health, and finding ways to enhance AQ [4][5].

The research experimented by [6] provides a detailed analysis of the correlation between AP and health in urban environments. Another article published in the Journal of the American College of Cardiology, smog exposure represents an elevated risk for cardiovascular diseases such as heart attacks and strokes [7]. Figure 1 illustrates a map of India featuring 30 states from where the data is being collected to predict smog.

Figure 1. Map of Data Collection Sites

Each state shown is represented by attractive figure 1 to give the specific states their own individual cultural or geographical identity. The states shown include Andra Pradesh, Arunachal Pradesh, Assam, Bihar, Chandigarh, Chhattisgarh, Delhi, Gujarat, Haryana, Himachal Pradesh, Jharkhand, Karnataka, Kerala, Madhya Pradesh, Maharashtra, Manipur, Meghalaya, Mizoram, Nagaland, Odisha, Puducherry, Punjab, Rajasthan, Sikkim, Tamilnadu,

Telangana, Tripura, Uttar Pradesh, Uttarakhand, and West Bengal. The pollution data collected from these states cover almost all diverse regions of India.

The hourly AQI recorded in the survey are often fed back into prediction models to make them more accurate. The paper by [8] states that forecasting the levels of smog is important for authorities to plan preventive measures to mitigate smog, including restricting traffic and issuing public health advisories. The paper by [9] describes a ML approach to smog prediction, which tackled the complexities of AP via the use of previously gathered real-time AP data. The study by [10] in which the researchers demonstrates the application of a deep learning (DL) approach to forecasting the level of smog, a concern for public health and the management of AQ. The predictions that are made from the underlying data are discussed considering the limitations of the traditional approach and the benefits of the advanced approach, using ML to create better models that reflect the complex relationships between a variety of irrelevant pollutants, atmospheric conditions, and other factors.

Due to the growing global concern for the economic and social consequences of unhealthy levels of AP. The forecasting of the AQI has become an important topic in urban environment research. Shih et al. [11] published a paper where a ML based AQI warning system was proposed to predict future AQ and identify both peak and low values. This included the integration of ML in detection systems. In the research paper, Shah et al. [12] introduced an intelligent real-time detection and notification system integrated with sensor integration and artificial intelligence in a distributed computing framework for enhanced detection of real-time data. A similar study carried out by Osman et al. [13] included the fusion of real-time and predictive analytics of AQI with the IoT that integrated the different communication modules (ZigBee and Wifi) to monitor environments and collected data from various nearby sensors to support the proposed method for environmental monitoring system. Furthermore, Kow et al. [14] introduced a convolutional neural network (CNN) of image-based DL architecture for AQ estimation in a distributed computing framework using a regression classifier in 2022. More recently, Liu et al. [15] focused on the AQI forecasting using genetic algorithm ensemble of extreme learning machines (GAELM) for predicting future AQ using ML algorithms. Ravindran et al. [16] illustrated the ML models for predicting the AQI in a coastal city of India named Visakhapatnam. In another research paper, Hardini et al. [17] presented the 'image-based models and CNNs' for AQ prediction. Similarly, Hardini et al. [18] presented the use of Ensemble ML for AQ prediction. Another research paper by Morapedi et al. [19] illustrated the 'ML techniques' for predicting the AP due to particulate matter (PM2.5) in South African cities by using 'AP datasets (2007 - 2015) from the region'.

Chen et al. [20] developed part of their research as an improved predictor for the 'hourly PM2.5 concentrations' that used a 'causal CNN' for short-term predictions. Zhang et al. [21] in their research paper utilized the ML algorithms for 'data-driven developments to improve the deterministic AP forecasts' in Greater Stockholm, Sweden. Masseran et al. [22] agreed with other researchers that predicting the AQI or AP and the classification prediction of unhealthy AP events in terms of their severity classes depends on suitable ML techniques. Kuo et al. [23] showed us how to predict dengue fever using ML models by considering different influencing factors, mainly meteorological factors, a vector index, and AQIs as an early warning system for public health. It is worthy to mention also the work of Jitkajornwanich et al. [24] who investigated the estimation of AQI with satellite data. Based on detecting data, having challenges, and different types of satellite images, researchers in the field of AQ have focused on predicting the AQI based on the satellite image using ML techniques and hybrid models that estimate some parameters involved.

The next section will outline our methodology by detailing the data sources, preprocessing, and predictive models. The subsequent section will present the results of our analysis. The study will ultimately be concluded with a summary of the findings and limitations of this research.

Below are the briefly explained salient features and aspects of this research work:

- Preprocessed data by computing AQI values and addressing anomalies, missing data, and outliers.

- Employed various algorithms named XGBR and ETR with hyperparameter tunning.
- Comparable protocols of building and deploying different ML models on real-time AQ data published by the Government of India to predict smog with a high level of accuracy.

The research bottleneck has been the lack of comparative ML model efficacy in predicting smog in real time, specifically in India's various environment scenarios. While some previous research employs different ML techniques, few provide a detailed comparative view that is specific to India's specific air quality environment. Unlike previous studies that focused on global AQ patterns, this research is unique in its focus on the specific AQ challenges of India. Moreover, while prior wok has largely applied traditional ML techniques, this study explores advanced methods such as hyperparameter tunning and ensemble models to improve predictive accuracy for India's urban and rural environments. The paper fills this gap by comparing several ML models for smog prediction with higher precision for India's urban monitoring needs. This paper offers a valuable contribution on India's AQ monitoring problem. In contrast to other research, which focus broadly on global AQ patterns, our work adjusts ML models to the uncertainty and heterogeneity of India's urban and rural data. It is also unique in combining real-time data from the CPCB with hyperparameter-tuned ensemble models like XGBR and ETR etc. The predictive performance is better than the known methods and shows the possibility of individualized ML methods in extremely diverse locations.

2 Methodology

This analysis gathered data for three consecutive months from the CPCB official website: https://cpcb.nic.in/real-time-air-quality-data/. The dataset used in this study is taken from the CPCB, which provides real-time AQI data across various regions in India. The data covers hourly AQI readings from 30 states in India, spanning three months, from January to March 2024. It includes various pollutants such as PM2.5, PM10, O3, CO, and NO2. The first was that the dataset didn't have AQI values, which are essential to calculate air pollution levels. AQI was factored into the dataset, using a common AQI formula to deal with this. Prior to the AQI, data was preprocessed (cleaning data, removing duplicates, and handling missing values). This was necessary to make the data valid and stable for the next step which was to apply machine learning algorithms to further analyze and predict air quality trends.

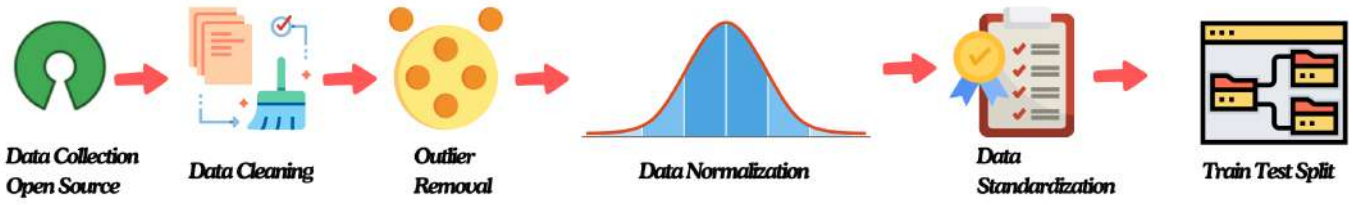

Figure 2. Data Preprocessing Workflow

Figure 2 represents the preprocessing workflow of the smog prediction data. It consists of many essential steps that help to prepare the data for the accurate modeling of the smog. It starts with open-source data collection. After that, the cleaning of missing values, and outliers are removed. The calculation of the AQI for each point using the standardized formula discussed above. The data is normalized and scaled. It involves the alteration of numbers into the common scale which helps to train the model. Finally, the data is split into two parts training and testing data which is required by the model.

The categorical AQI values are computed as follows:

$$\text{AQI} = \left(\frac{\text{Pollutant Concentration}}{\text{Pollutant Standard}} \right) \times 100 \tag{1}$$

This formula is applied to every data point to create an equivalent scale of AQ across different locations. Based on standards like these, the data is scaled and normalized. The raw data was applied to this formula to get AQI value for ML models. Before training and machine learning models, a correlation matrix was generated to extract the relationships between the various air pollutants and the AQI. The correlation matrix was calculated to identify significant predictors of smog levels, that allowed us to select the most relevant features for model training. The dataset is pre-processed to train and test the ML model. Several ML algorithms are used for prediction of smog. These algorithms consist of Linear Regression (LR), Decision Tree Regressor (DTR), Random Forest Regressor (RFR), K Nearest Neighbors Regressor (KNNR), Ada Boost Regressor (ABR), ETR, XGBR, and Support Vector Regressor (SVR). For models like LR, the relationship between the input features and the AQI is assumed to be linear. However, AQ data often exhibits complex, non-linear relationships, that may make this model less effective compared to more flexible models like RFR or XGBR that can capture these non-linear relationships.

Linear Regression (LR): One of the simplest and most interpretable models that is used for a regression task is the LR model. With this model, it is assumed that the target variable is linearly related to the input features according to the equation:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \tag{2}$$

Where y is the predicted value, β_0 is the intercept, and β_1 are the coefficients of the input features x_1. Although incredibly simple, linear regression may fail to capture non-linear relationships among the data.

Decision Tree Regressor (DTR): It is a non-linear model that splits the data into sub-groups based on the value of a feature variable, recursively partitioning the data to minimize variance within each subset and assigning the prediction for a given observation to the mean value of the corresponding target variable in the terminal node.

Random Forest Regressor (RFR): It exploits the fact that ensemble methods can outperform each other by combining 'best predictions'. The prediction equation for RF is:

$$\hat{y} = \frac{1}{N} \sum_{i=1}^{N} T_i(x) \tag{3}$$

Where $T_i(x)$ represents the prediction from the i-th tree.

K Nearest Neighbors Regressor (KNNR): It is a non-parametric method that predicts the target variable based on the closest k training examples. The prediction is the average of the target values of the KNN, determined by a distance metric such as Euclidean distance. The equation is:

$$\hat{y} = \frac{1}{k} \sum_{i=1}^{k} y_i \tag{4}$$

Where y_i are the target values of the KNNs.

AdaBoost Regressor (ABR): It is an ensemble method that combines weak learners to create a strong predictor. It adjusts the weights of the training data based on the errors of previous learners, focusing more on difficult-to-predict examples. The prediction is given by:

$$F(x) = \sum_{m=1}^{M} \alpha_m h_m(x) \tag{5}$$

Where $h_m(x)$ are the weak learners and α_m are their corresponding weights.

Extra Tree Regressor (ETR): It is similar to RF but uses random splits for improved diversity. This method averages predictions from multiple extremely randomized trees, enhancing prediction accuracy and robustness. It uses the same prediction equation as RF.

XGBoost Regressor (XGBR): It is an optimized gradient boosting method known for its high performance. It builds a series of trees, one after the other, each fixing the mistakes of the previous ones. The equation is:

$$\hat{y} = \sum_{k=1}^{K} f_k(x) \tag{6}$$

Where f_k are regression tree functions. XGB employs regularization to avoid overfitting and better generalization.

Support Vector Regressor (SVR): It utilizes support vector machines which are used to perform a regression task given the values of the input data points and the target responses. The prediction function is:

$$f(x) = \sum_{i=1}^{n} \alpha_i K(x, x_i) + b \tag{7}$$

Where K is the kernel function, α_i are the support vectors, and b is the bias term.

To further improve these models' performance, we performed hyperparameter tuning and grid search with cross-validation. To optimize the performance of ML models, hyperparameter tuning was carried out using `GridSearchCV`. This process involved adjusting key parameters such as the number of estimators, learning rate, and max depth for different models, as each model had different hyperparameters. The best performing configurations were selected based on evaluation metrics: R^2, MSE, and RMSE. This is accomplished by systematically changing the parameters of a model to find which combination of parameters produces the best predictive performance. In this case, the usage of R-Squared (R^2), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Root Mean Squared Log Error (RMSLE) is performed to assess and compare performance.

The evaluation metrics used are defined as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{8}$$

Where y_i is the actual value, $\hat{y}_i$ is the predicted value, and $\bar{y}$ is the mean of the actual values.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{9}$$

Where n is the number of data points, and y_i and $\hat{y}_i$ are the actual and predicted values, respectively.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{10}$$

Where RMSE is the square root of the Mean Squared Error.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{11}$$

Where MAE represents the Mean Absolute Error.

$$RMSLE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\log(1+y_i) - \log(1+\hat{y}_i)\right)^2} \quad (12)$$

Where RMSLE calculates the Root Mean Squared Logarithmic Error.

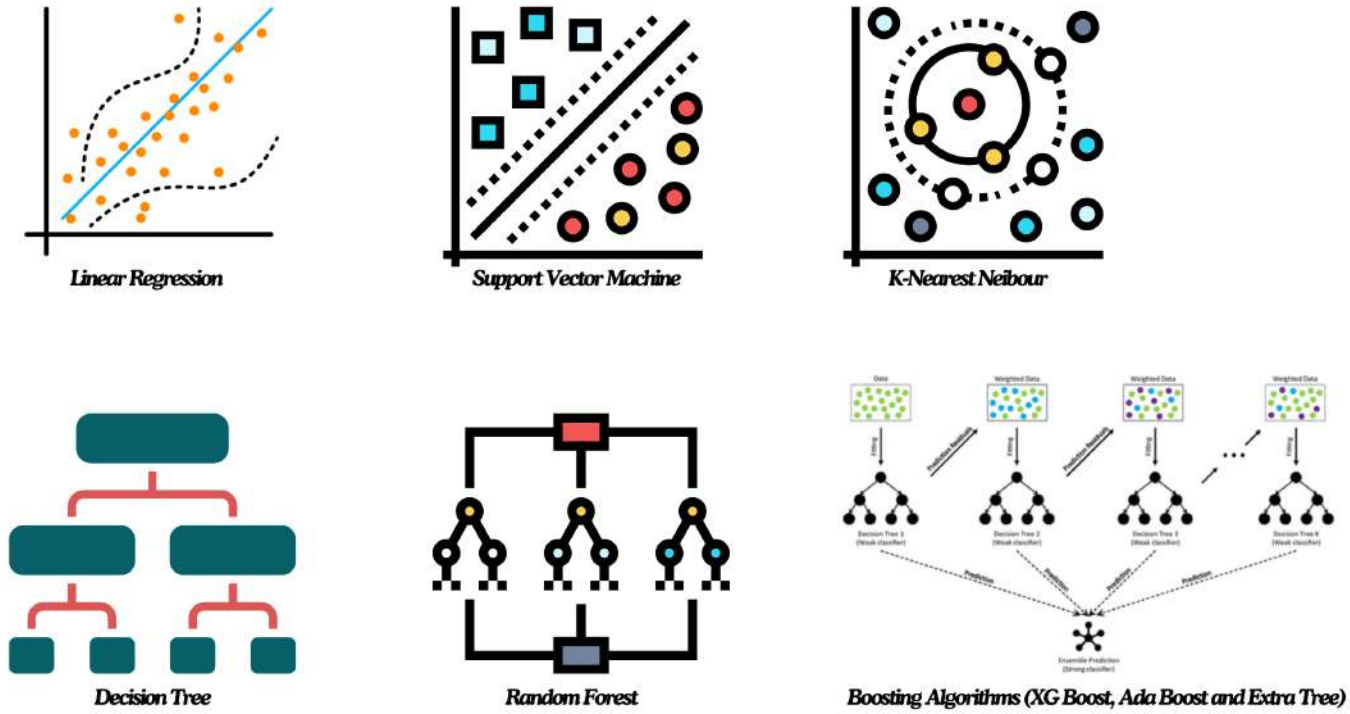

Figure 3. Machine Learning Models Overview

Figure 3 depicts various ML models used to predict smog in the study. To begin with, LR is shown with a straight line, which outlines a simple linear relationship between one value and the prediction of another value. Furthermore, the DTR is represented as a tree that is used to make predictions by collecting a set of future outcomes from a node. Meanwhile, the RFR has a symbol of multiple trees. Additionally, the KNNR is illustrated by network nodes, which are clustered based on the nearest neighbors. The boosting methods, being the collection of weak learners learning, classified problems for stronger predictions are for ABR, ETR, and XGBR.

3 Results

The ML models were trained on a built-in computer running on an Intel i7 9th generation processor, 16GB of RAM, and an NVIDIA RTX 2060 Super GPU. This powerful configuration allowed us to train models efficiently and train each model hyperparameter tuned in 20 minutes. High-performance hardware and highly optimised algorithms meant that the models were trained quickly and accurately and with high precision. This configuration saved a lot of time in training and was flexible enough to play with various parameters and calibrate models for the best performances.

Figure 4 illustrates the hyperparameters tunning of ML models. The process of tuning the hyper-parameter starts with the selection of parameters. Followed by cross-validation, where different sets of parameters are considered to verify the performance. Next step, the model evaluation is performed under R2 metric. The last

stage called as optimization step indicates a choice of the best performing parameters to refine the accuracy and robustness of inputs.

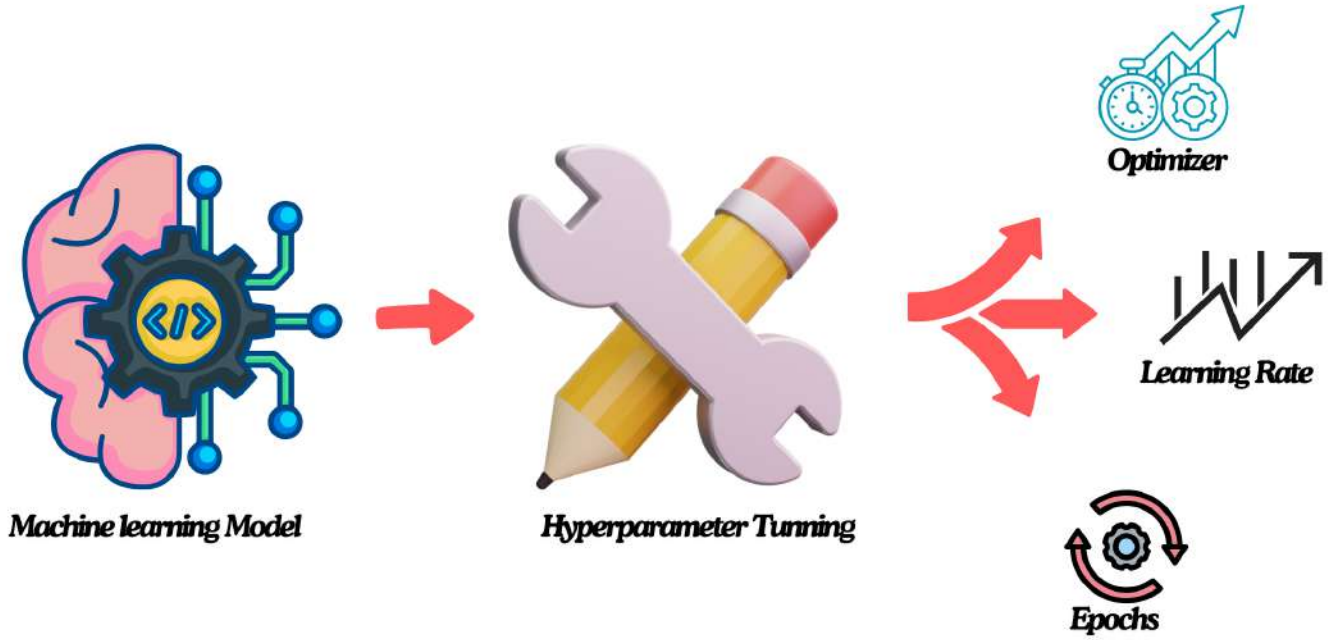

Figure 4. Hyper-parameter Tuning Process

Figure 5 shows the Heatmap of AQI levels in India on a geographical scale. The AQI values for the most elevated areas are highlighted in yellow to purple, with purple representing the highest values. The graph plots this information on a map of India to illustrate the distribution of air pollution problems. It gives a visual guide of areas with air pollution, along a gradient color spectrum that can be easily understood.

Figure 5. Geographical Heatmap of AQI Levels

The results of each ML model are measured by metrics of R2, MSE, RMSE, MAE, and RMSLE that precisely describe the accuracy and error rates of predicted smog levels. The AQ data used in this dataset is real-time in real-world at different places in India, so there is a lot of variability and different types of data. To overcome this complexity and variability, several ML models are used to ensure the best accuracy in predicting smog levels. The

insights drawn from the ML models highlighted their effectiveness and limitations in addressing the objectives of the study. Simpler models such as Linear Regression (LR) and K-Nearest Neighbors Regressor (KNNR) also were used but performed less than ensemble approaches. LR presumes a linear correspondence between input features and target variables, which does not apply to the nonlinear and complex interactions observed in air quality measurements. Similarly, KNNR was more reliant on finding the right k-value and noise sensitivity when applied to high-dimensional data, hence had a lower accuracy and error rate. These limitations emphasize the need for ensemble methods that can better account for data heterogeneity. Table 1 below depicts the metrics obtained from different ML models which are used in this research.

Table 1. Model Performance Metrics

Model	Mean Squared Error	Root Mean Squared Error	Mean Absolute Error	R-Squared	Root Mean Squared Log Error
Linear Regression	1.1e-27	3.407e-14	2.6e-14	1.00	5.8e-15
Decision Tree Regressor	8.0e-03	9.0e-02	1.5e-02	0.99	7.0e-03
Random Forest Regressor	1.1e-02	1.0e-01	2.1e-02	0.99	1.0e-02
K-Nearest Neighbors Regressor	1.7e01	4.1	1.5	0.98	1.6e-01
AdaBoost Regressor	8.0e-01	8.9e-01	4.8e-01	0.99	1.0e-01
Extra Trees Regressor	4.0e-03	6.5e-02	1.0e-02	1.00	3.0e-03
XG Boost Regressor	1.7e-02	1.3e-01	3.6e-02	0.99	7.0e-03
Support Vector Regressor	3.1e01	5.5	2.8	0.97	3.1e-01
Extra Trees Regressor (Tuned)	4.3e-01	6.5e-01	3.0e-01	0.99	9.3e-02
Random Forest Regressor (Tuned)	7.3e-01	8.6e-01	3.5e-01	0.99	9.6e-02
XG Boost Regressor (Tuned)	1.7e02	1.3e01	5.7	0.99	3.9e-01

Figure 6 shows the comparison of various ML models for 5 metrics MSE, RMSE, MAE, RMSLE, and R2. In each subplot each metric value is plotted with bars, and there are also the best performing models marked. It can be observed that Extra Trees Regressor and XG Boost Regressor performed extremely well in several aspects. To further validate our findings, other baseline models including Decision Tree Regressor (DTR), Random Forest Regressor (RFR), and Support Vector Regressor (SVR) networks were used. These models had slightly higher accuracy compared with simpler models, but not as much as XGBR and ETR. While XGBR performed well in this study, its performance could be considered overly idealized, as the validation with more diverse dataset and longer time periods would help assess the model's generalizability and ensure its robustness under varying conditions. DTR and RFR achieved R2 of 0.99, whereas SVR attains 0.97 mainly because of their inability to use the real-time variability in the data. RFR outperforms other models due to its ability to handle high-dimensional data and capture complex, non-linear relationships. As an ensemble method, it aggregates the predictions from multiple decision trees, which improves its generalization and robustness against over fitting. The importance of key parameters such as PM2.5 and NO2 were found to be significant in determining the AQI levels, in line with findings from similar studies.

It is also important that hyper-parameter tuning also improves performance. Grid search and cross-validation were used to determine the best hyper-parameter configuration among the several ML models. These included varying the tree count or number of estimators or the number of trees in the forest (for RFR and ETR), varying the learning rate and the number of boosting rounds (in XGBR) and varying the value of k (for KNNR). The best performing configuration sharply reduced both MSE and RMSE, hence the effect of hyper-parameter tuning to be irreplaceable. To visually compare the performances among these models, the graphical representation between the actual and predicted AQI values of the best performing models (XGBR and ETR) is shown in Figure 5. Figure 5 depicts a coincidence between the predicted and actual AQI values which signifies the high accuracy of the model.

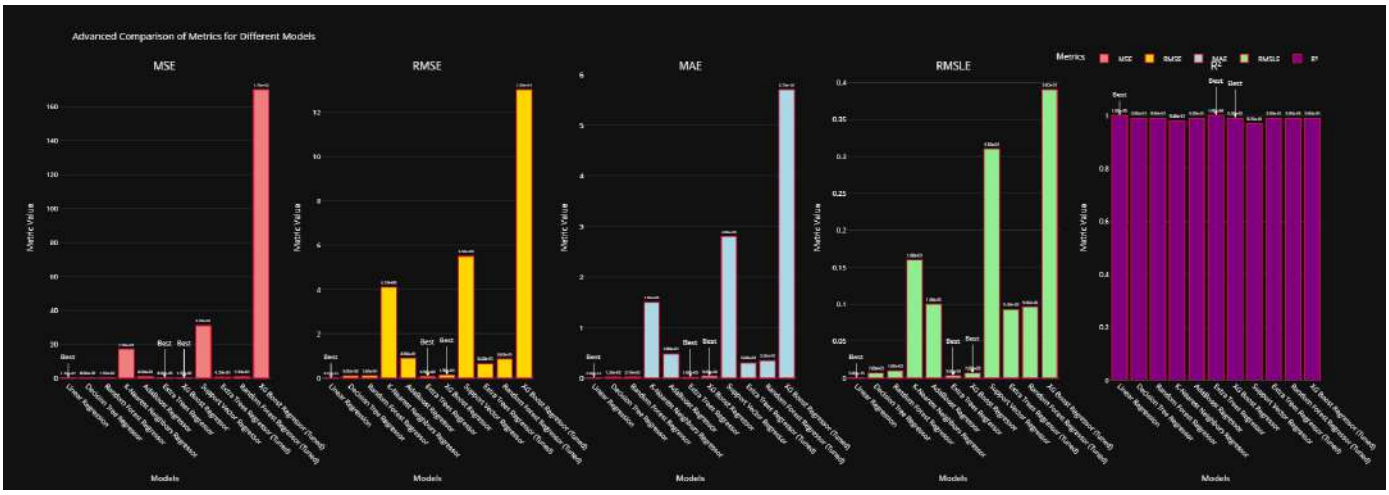

Figure 6. Bar Chart of Various Model Performance with Various Performance Metrics

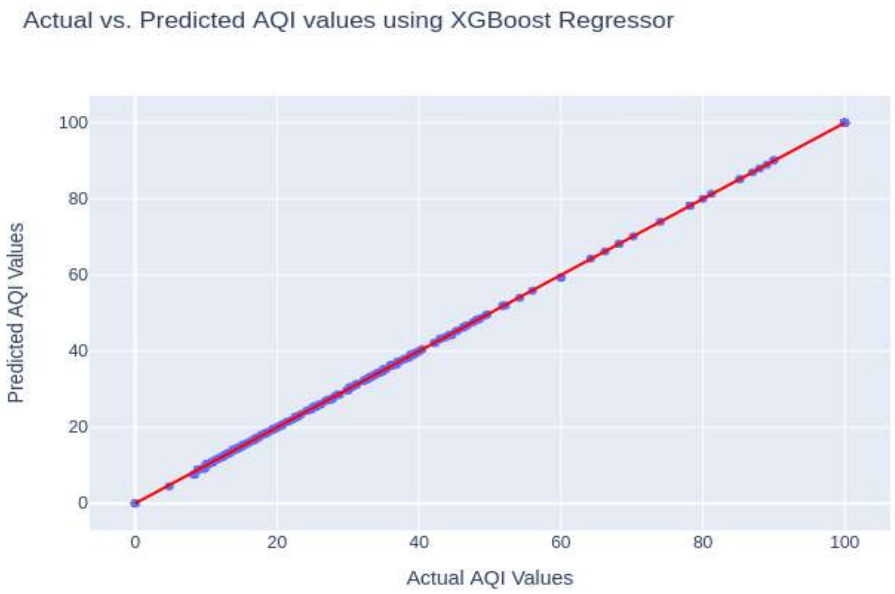

Figure 7. Actual vs. Predicted AQI values using XGBoost Regressor

Figure 7 records the actual versus the predicted AQI values for the ETR. As for XGBR, the predicted values are near to the actual ones, which denotes the high robustness of the model. Figure 8. Actual Vs. Predicted AQI Values for ETR.

Table 2 compares our results with previously published works, it is seen that there is a drastic increase in prediction performance. Earlier models were reported to reach R2 value of around 0.87 to 0.95 and RMSE value between 2.0 to 4.0. But our best performing models (XGBR and ETR) reached R2 values of 0.99966 and 0.9582 respectively and RMSE values of 0.65600 and 1.4455 respectively. The key factor behind these comparatively better validation results was the use of advanced ML techniques, robust preprocessing techniques, and extensive

Figure 8. Actual vs. Predicted AQI Values for Extra Tree Regressor

hyperparameter tuning. One of the major things that contributed to good prediction in our case was the process of hyperparameter tuning. Finally, the best performing model was the XGBR with the bare minimum amount of error. Another strong model was the ETR. In the end, we conclude that ensemble methods can do well with smog prediction which helps better manage and prevent smog issues as well as promote public health.

Table 2. Comparison Table

SNO	Feature Selected	Model Technique	Evaluation Metrics	Reference
1	NO2, O3, PM10, PM2.5	LSTM, Hierarchical GRU	RMSE (5.35 μg/m³)	[24]
2	PM10, PM2.5, CO, NO2, SO2, NOx, NO	KNN, GNB, SVM, RF, XGBoost	Accuracy: 90%	[25]
3	Temperature, CH4, CO, NMHC, NO, NO2, NOx, O3, PM10, PM2.5, RH, SO2	RF, SVM, and ANN	Accuracy: 97%	[26]
4	PM2.5	CNN, GRU, LSTM	Accuracy: 99%	[27]
5	PM2.5, PM10, SO2, CO	PCA, VGG-16	Accuracy: 85%	[28]
6	CO, SO2, NO2, O and PM	DT, Gradient Boosted Tree, RF	Accuracy: 82%	[29]
7	O3, PM2.5, PM10, SO2, CO, And NO2	ARIMA, ARIMAX and RNN Models	Accuracy: 75%	[30]
8	Encoder, STAA-LSTM Network	AAD, RMSE, MAE, and R2	Accuracy: 37%	[31]
9	Number of Occupants, Area Per Person, Outdoor Temperature, Outer Wind Speed, Relative Humidity, and Air Quality	ANN, SVM, DT, GPR, LR, EL, Optimized GPR, Optimized EL, Optimized DT and Optimized SVM	Accuracy: 98%	[32]
10	PM2.5, PM10, O3, CO, NO2, SO2	LGBM, LSTM, WeightedC, LRC, and RFC	Precision: 97.5% & F1 Score: 93.3%	[33]
11	PM2.5, PM10, NO, NO2, NH3, CO, SO2, O3, Benzene, Toluene	SVR, RFR, and CBR	RMSE: 0.1403	[34]
12	PM2.5, PM10, NO, NO2, NH3, CO, SO2, O3, Benzene, Toluene, AQI, NOx		R2: 1.0	Proposed

4 Conclusions

The study shows that the advanced ML model can predict smog levels with outstanding accuracy given the real-time AQI data. Also, the XGBR and ETR were found to be the best models, among all ML models, with very high accuracy and low error rates. These two models, in comparison to classical methods and other ML models, were able to extract the complex patterns in the data very well. These results demonstrate the potential of using advanced ML methods to strengthen environmental monitoring and public health safeguards. AQI management can be facilitated with the increased predictive accuracy of the models, which can inform expedient and efficient actions in protecting public health and reducing environmental impacts. Future work could further explore model variants, more heterogeneous data sources, and even more efficient computation to better tackle the problem of smog prediction. In turn, strengthened methodologies in smog prediction make for better environmental management and for better mitigating the impacts of AP on public health.

Author Contributions

Sahil Jatoi: Conceptualization, Methodology, Software, Investigation, Writing-original draft. **Bushra Abro**: Formal analysis, Investigation, Visualization, Data curation, Writing-review editing. **Sanam Narejo**: Validation, Supervision, Project administration. **Yaqoob Ali Baloch**: Resources, Writing- review editing, Funding acquisition. **Kehkashan Asma**: Supervision, Project administration.

Compliance with Ethical Standards

It is declared that all authors don't have any conflict of interest. Furthermore, informed consent was obtained from all individual participants included in the study.

References

[1] A. Javed et al., "The potential impact of smog spell on humans' health amid covid-19 rages," *Int. J. Environ. Res. Public Health*, vol. 18, no. 21, 2021, doi: 10.3390/ijerph182111408.

[2] M. Roser, "Data review: how many people die from air pollution?," 2021, [Online]. Available: https://ourworldindata.org/data-review-air-pollution-deaths.

[3] M. N. Azimi and M. M. Rahman, "Unveiling the health consequences of air pollution in the world's most polluted nations," *Sci. Rep.*, vol. 14, no. 1, pp. 1–25, 2024, doi: 10.1038/s41598-024-60786-0.

[4] R. Kaur and P. Pandey, "Air Pollution, Climate Change, and Human Health in Indian Cities: A Brief Review," *Front. Sustain. Cities*, vol. 3, no. August, 2021, doi: 10.3389/frsc.2021.705131.

[5] I. Manisalidis, E. Stavropoulou, A. Stavropoulos, and E. Bezirtzoglou, "Environmental and Health Impacts of Air Pollution: A Review," *Front. Public Heal.*, vol. 8, no. February, pp. 1–13, 2020, doi: 10.3389/fpubh.2020.00014.

[6] W. Raza et al., "A review on the deteriorating situation of smog and its preventive measures in Pakistan," *J. Clean. Prod.*, vol. 279, p. 123676, 2021, doi: 10.1016/j.jclepro.2020.123676.

[7] A. Grzywa-Celińska, A. Krusiński, and J. Milanowski, "'Smoging kills' – Effects of air pollution on human respiratory system," *Ann. Agric. Environ. Med.*, vol. 27, no. 1, pp. 1–5, 2020, doi: 10.26444/aaem/110477.

[8] S. A. Siddiqui, N. Fatima, and A. Ahmad, "Smart Air Pollution Monitoring System with Smog Prediction Model using Machine Learning," *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 8, pp. 401–409, 2021, doi: 10.14569/IJACSA.2021.0120846.

[9] S. Geetha and L. Prasika, "Smog prediction model using time series with long-short term memory," *Int. J. Mech. Eng. Technol.*, vol. 10, no. 1, pp. 1026–1032, 2019.

[10] J. Tian, Y. Liu, W. Zheng, and L. Yin, "Smog prediction based on the deep belief - BP neural network model (DBN-BP)," *Urban Clim.*, vol. 41, no. December 2021, p. 101078, 2022, doi: 10.1016/j.uclim.2021.101078.

[11] D. H. Shih, T. W. Wu, W. X. Liu, and P. Y. Shih, "An azure aces early warning system for air quality index deteriorating," *Int. J. Environ. Res. Public Health*, vol. 16, no. 23, 2019, doi: 10.3390/ijerph16234679.

[12] S. K. Shah, Z. Tariq, J. Lee, and Y. Lee, "Real-Time Machine Learning for Air Quality and Environmental Noise Detection," *Proc. - 2020 IEEE Int. Conf. Big Data*, Big Data 2020, pp. 3506–3515, 2020, doi: 10.1109/BigData50022.2020.9377939.

[13] Y. K. Y. & K. A. K. Nurmadiha Osman, Mohd Faizal Jamlos, Fatimah Dzaharudin, Aidil Redza Khan, "Real-Time and Predictive Analytics of Air Quality with IoT System: A Review," 2021, doi: https://doi.org/10.1007/978-981-33-4597-3_11.

[14] P. Y. Kow, I. W. Hsia, L. C. Chang, and F. J. Chang, "Real-time image-based air quality estimation by deep learning neural networks," *J. Environ. Manage.*, vol. 307, no. January 2021, p. 114560, 2022, doi: 10.1016/j.jenvman.2022.114560.

[15] C. Liu, G. Pan, D. Song, and H. Wei, "Air Quality Index Forecasting via Genetic Algorithm-Based Improved Extreme Learning Machine," *IEEE Access*, vol. 11, no. July, pp. 67086–67097, 2023, doi: 10.1109/ACCESS.2023.3291146.

[16] 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, no. May, p. 139518, 2023, doi: 10.1016/j.chemosphere.2023.139518.

[17] M. Hardini, M. H. R. Chakim, L. Magdalena, H. Kenta, A. S. Rafika, and D. Julianingsih, "Image-based Air Quality Prediction using Convolutional Neural Networks and Machine Learning," *APTISI Trans. Technopreneursh.*, vol. 5, no. 1SP, pp. 109–123, 2023, doi: 10.34306/att.v5i1Sp.337.

[18] M. Hardini, R. A. Sunarjo, M. Asfi, M. H. Riza Chakim, and Y. P. Ayu Sanjaya, "Predicting Air Quality Index using Ensemble Machine Learning," *ADI J. Recent Innov.*, vol. 5, no. 1Sp, pp. 78–86, 2023, doi: 10.34306/ajri.v5i1sp.981.

[19] T. D. Morapedi and I. C. Obagbuwa, "Air pollution particulate matter (PM2.5) prediction in South African cities using machine learning techniques," *Front. Artif. Intell.*, vol. 6, 2023, doi: 10.3389/frai.2023.1230087.

[20] Y. Chen, L. Huang, X. Xie, Z. Liu, and J. Hu, "Improved prediction of hourly PM2.5 concentrations with a long short-term memory and spatio-temporal causal convolutional network deep learning model," *Sci. Total Environ.*, vol. 912, no. November 2023, p. 168672, 2024, doi: 10.1016/j.scitotenv.2023.168672.

[21] Z. Zhang, C. Johansson, M. Engardt, M. Stafoggia, and X. Ma, "Improving 3-day deterministic air pollution forecasts using machine learning algorithms," *Atmos. Chem. Phys.*, vol. 24, no. 2, pp. 807–851, 2024, doi: 10.5194/acp-24-807-2024.

[22] N. Masseran, M. A. M. Safari, and R. R. M. Tajuddin, "Probabilistic classification of the severity classes of unhealthy air pollution events," *Environ. Monit. Assess.*, vol. 196, no. 6, 2024, doi: 10.1007/s10661-024-12700-4.

[23] C. Y. Kuo, W. W. Yang, and E. C. Y. Su, "Improving dengue fever predictions in Taiwan based on feature selection and random forests," *BMC Infect. Dis.*, vol. 24, no. Suppl 2, pp. 1–11, 2024, doi: 10.1186/s12879-024-09220-4.

[24] V. Oldenburg, J. Cardenas-Cartagena, and M. Valdenegro-Toro, "Forecasting Smog Clouds With Deep Learning," Oct. 2024, Accessed: Nov. 17, 2024. [Online]. Available: https://arxiv.org/abs/2410.02759v1.

[25] K. Kumar and B. P. Pande, "Air pollution prediction with machine learning: a case study of Indian cities," *International Journal of Environmental Science and Technology*, vol. 20, no. 5, pp. 5333–5348, May 2023, doi: 10.1007/S13762-022-04241-5/TABLES/7.

[26] D. S.-I. J. of E. R. and undefined 2021, "Implementation of machine learning algorithms for analysis and prediction of air quality," academia.eduD SanjeevInternational Journal of Engineering Research Technology (IJERT), 2021•academia.edu, Accessed: Nov. 17, 2024. [Online]. Available: https://www.academia.edu/download/66199522/implementation_of_machine_learning_algorithms_for_IJERTV10IS030323.pdf.

[27] X. B. Jin et al., "Deep Spatio-Temporal Graph Network with Self-Optimization for Air Quality Prediction," *Entropy 2023*, vol. 25, no. 2, p. 247, Jan. 2023, doi: 10.3390/E25020247.

[28] C. W. Chen, Y. S. Tseng, A. Mukundan, and H. C. Wang, "Air pollution: Sensitive detection of pm2.5 and pm10 concentration using hyperspectral imaging," *Applied Sciences (Switzerland)*, vol. 11, no. 10, p. 4543, May 2021, doi: 10.3390/APP11104543/S1.

[29] O. F. AlThuwaynee et al., "Demystifying uncertainty in PM10 susceptibility mapping using variable drop-off in extreme-gradient boosting (XGB) and random forest (RF) algorithms," *Environ Sci Pollut Res Int*, vol. 28, no. 32, pp. 43544–43566, Aug. 2021, doi: 10.1007/S11356-021-13255-4.

[30] M. L. Avila, A. M. Alonso, and D. Peña, "Modelling multiple seasonalities with ARIMA: Forecasting Madrid NO2 hourly pollution levels," Apr. 2023, doi: 10.21203/RS.3.RS-2860239/V1.

[31] S. Abirami and P. Chitra, "Probabilistic air quality forecasting using deep learning spatial–temporal neural network," *Geoinformatica*, vol. 27, no. 2, pp. 199–235, Apr. 2023, doi: 10.1007/S10707-022-00479-W/FIGURES/13.

[32] N. R. Kapoor, A. Kumar, A. Kumar, A. Kumar, and H. C. Arora, "Prediction of Indoor Air Quality Using Artificial Intelligence," *Machine Intelligence, Big Data Analytics, and IoT in Image Processing: Practical Applications*, pp. 447–469, Jan. 2023, doi: 10.1002/9781119865513.CH18.

[33] Q. Liu, B. Cui, and Z. Liu, "Air Quality Class Prediction Using Machine Learning Methods Based on Monitoring Data and Secondary Modeling," *Atmosphere 2024*, vol. 15, no. 5, p. 553, Apr. 2024, doi: 10.3390/ATMOS15050553.

[34] N. S. Gupta, Y. Mohta, K. Heda, R. Armaan, B. Valarmathi, and G. Arulkumaran, "Prediction of Air Quality Index Using Machine Learning Techniques: A Comparative Analysis," *J Environ Public Health*, vol. 2023, no. 1, p. 4916267, Jan. 2023, doi: 10.1155/2023/4916267.