Investigating innovative methods of utilizing geospatial data and AI to measure and evaluate Air pollution

Ayan, Krishan, Shivam, Shubham
Department of Computer Science
Bennett University
Greater Noida, India

Abstract—Air quality affects public health and the environment, and it is a growing concern, especially in urban areas. Air Quality Index (AQI) provides a method for calculating and communicating air pollution. Accurate AQI estimates are important for measurement and decision making. This study uses machine learning to develop an AQI prediction model for Indian cities. The data we include on pollutants such as PM2.5, PM10, NO2, and CO. To determine the most suitable method for AQI estimation, 11 different models including linear regression, ridge regression, ElasticNet, support vector regression, gradient boosting, random forest, and XGBoost were tested. We use the R² scores obtained from training, validation, and test data to evaluate the model performance. Random Forest and XGBoost models outperformed other models with R² scores of 0.9583 and 0.9227 respectively, while Gradient Boosting also performed well with an R² score of 0.8441. Simple models like linear regression perform very poorly, indicating their limitations in modelling the relationship between air pollution and AQI. The significance analysis shows that PM2.5 and PM10 are the main components of AQI estimation, and their determination is very important for air quality assessment. This paper demonstrates the effectiveness of machine learning models in predicting AQI, providing a useful resource for urban planners, environmental organizations, and policy makers. For future research, we can focus on combining data in real time, in the weather environment, and complex models such as deep learning to make prediction accuracy and activation stronger.

I. INTRODUCTION

Air pollution is a major concern to the community and is made up of so many pollutants that cause serious health impacts. PM2.5 and PM10, NOs, NO2, SO2, COs, and O3 are the established pollutants that have prominent impacts on causing respiratory and cardiovascular diseases. Besides these, other chemical toxicants such as Volatile Organic Compounds (VOCs); benzene, toluene, xylene, and ammonia (NH3) adversely affect the health. For example, benzene causes blood-related diseases affecting the bone marrow, such as leukemia. Toluene and xylene present in industrial solvents cause neurological problems including headaches and long-term damage to the brain. Agriculture and some industries emit ammonia into the air, which leads to irritation of respiratory tracts with worsening effects on the respiratory system to chronic lung diseases with long exposure to the product.

The compounds mentioned above pose a much higher risk by acting synergistically to cause diseases such as asthma, bronchitis, and COPD.

VOCs are also involved in the formation of secondary pollutants such as fine particulates and ground-level ozone all of which worsens air quality and intensifies health implications. These effects are catastrophic and cut across vulnerable populations such as children, the elderly, and those with underlying medical conditions, which makes it even more important for there to be enhancement of policies on air quality and public health.

The AQI is a uniform scale that quantises air pollution status and potential health effect. This accumulates densities of particulate matter, PM2.5 and PM10, nitrogen dioxide NO2, sulfur - dioxide SO2, carbon monoxide CO, ozone O3, sometimes volatile organic compounds such as benzene. AQI is a scale that also has categories such as Good, Moderate, Poor, Very Poor, and Hazardous each of which is related to the health of people. For instance, a given AQI span of 0-50 is safe, and that it poses little threat while a span of 301 and above is classified as “Hazardous” and may cause health states of emergency across large population groups.

Since air quality is a critical determinant of health status of the people, it is important that future AQI values are predicted correctly. However, the task of predicting AQI is cumbersome because of a large number of pollutants with different impacts on air quality. Computerized approaches to forecast and estimate airborne pollutants can be based on historical records and physical observance equipment, which are expensive and cumbersome. Machine learning methods are identified as one of the most effective approaches to forecast AQI taking into account the history of data. AQI can be predicted accurately with the skills of other environmental factors such as the concentrations of pollutants, weather conditions, and geographic data about models put in machine learning algorithms [1]-[4].

In this study, we have used 10 different regression models to analyse the data and estimate AQI. We used the city_day data available on Kagle. In this data, we have 16 features, which are city names and pollutant names. We have multiple environmental variables in the dataset which consists of particulate matter (PM2.5, PM10), nitrogen oxides (NO2, NOx), carbon monoxide (CO), sulfur dioxide (SO2), ozone (O3), Benzene,toluene, and Xylene. These features, when enhanced with sophisticated machine learning, can serve as a better solution compared to the current method for forecasting AQI levels [4], [5]. The candidate models included in this study comprise of Linear Regression, Ridge, SGDRegressor, ElasticNet, Lasso, SVR, Gradient Boosting, XGBoost, Random Forest, and LightGBM. Evaluation is based on R2 score and the best Train, validation, and test R2 score is of RandomForestRegressor which is 98.04% ,85.00% and 95.21%.

II.RELATED WORK

A. Traditional Methods for Air Quality Prediction

Previous procedures that have been used for predicting AQI rely mainly on first-order econometric models such as the linear regression model, multiple regression models, and time series analysis. These techniques provided a systematic way of trending distinct levels of pollutants as well as their AQI values. While being easy to understand and simple, these could not represent complex, disjointed interactions between pollutants like PM2.5, PM10, and NO2 or the effects of weather variables such as wind speed or temperature. Thus, these approaches often did not capture the actual AQI under unstable conditions of the real environment [8], [9].

B. Advancements in Machine Learning for AQI Prediction

A probabilistic approach towards the AQI prediction has been successfully introduced through machine learning, which provides an efficient solution to analyze non-linear interactions otherwise difficult to articulate through conventional statistical models. SVM, Decision Trees, Random Forests, and Gradient Boosting are some of the recognized algorithms due to their capability in managing and analyzing the voluminous data set. Forms of learning include Random Forest and Gradient Boosting, which have shown significant improvements in predictive accuracy with the help of different base learners [18], [21].

C. Innovations in Ensemble and Deep Learning Models

More so, the employment of ensemble procedures such as XGBoost and LightGBM has ameliorated the level of prediction error further. Such methods include combining many shallow learners to build a stronger learner, which is effective when used for AQI forecasting [24]. ANNs and LSTM networks have been appreciated for their ability to capture temporal patterns and for comprehending the intricate association between various pollutants [4], [9]. However, using deep learning techniques, one often faces several challenges, which can be attributed to the big costs of computation and the need for a massive dataset [5], [10].

D. Feature Importance and External Variable Integration

The investigation has pinpointed the relation between several pollution agents including PM2.5, PM10, and NO2 with the levels of air quality. Hence PM2.5, in particular, has been classified as a significant contributor to the calculated AQI [2], [6]. In addition, by integrating information of temperature, humidity, and wind speed into the model of the pollutants, the improved accuracy results from the consideration of conditions that relate to the behavior of pollutants [11], [15]. This integration has made the models very comprehensive and quite effective to represent real-life scenarios [17].

III. PROBLEM DEFINITION

It is only possible to provide an estimate for the AQI of different cities of India, which is a very significant task to achieve the goal of maintaining people’s health and the environment. In turn, this task is charitable with much risk and requires the identification of unique and resistant solutions.

1.Non-linear Interactions Among Variables

There is an interaction between air pollutants including PM2.5, PM10, NO2 and meteorological variables including temperature, relative humidity and wind speed which are always nonlinear and multi-folded. Regular models of regression analysis do not always help to study such complex relations satisfactorily. Other techniques like the ML and deep learning pose potential when used to modelling of such non-linear interactions [1]-[5]. Nevertheless, the ability to generalize the discovered structure to other or to come up with another structure for a new data distribution is a challenging problem.

2.Data Quality and Variability

The available air quality datasets in India also contains missing records, incorrect measurements and different sampling techniques that result in data gap. Such variability can seriously affect model performance as evidenced in this study. These problems have been discussed in terms of data preprocessing, imputation, and anomaly detection techniques, and it has been concluded that the performance of these techniques depends on the quality and quantity of data to be used [6]-[10].

3.Real-Time Adaptability

For the AQI to be a reliable predictor its algorithms must incorporate real-time data pertaining to the environment and emissions to produce the forecasts. Indeed, standard batch-trained models are not adept at adapting to new information and modifying their predictions correspondingly, hence the search for a more adaptable approach such as online learning and ensemble methods [11]-[15].

4.Model Suitability and Scalability

The most important problem concerns defining an accurate ML model which would have significant performance characteristics at training, validation, and testing steps. Besides, use of models that are appropriate for extension to more than one geographical area with different distributions of pollutants and different meteorological conditions is therefore important for applicability. Although several ensemble models including Random Forest, XGBoost and LightGBM have provided reasonably accurate predictions, these models need to considered for scalability and computational complexity [16]-[20].

IV. METHODOLOGY

The methodology for this study involves a structured approach to building and evaluating machine learning models for predicting the Air Quality Index (AQI) in Indian cities. The process is divided into the following stages:

1. *Data Collection and Preprocessing*

The premise for this work is to assemble and prepare a data set of air quality measurements for input into ML models.

Data Source:

Data set of city_day data from Kaggle was used . Data pertaining to the past number of years for the air quality of the selected Indian cities such as concentrations of pollutants (PM2.5, PM10, NO2, NO, NH3 CO, SO2, 03, Benzene, Toluene, Xylene) and meteorological parameters (temperature, humidity, wind speed) were collected. These variables give a comprehensive picture of the factors which determine AQI [1]-[3].

Data description:

Data set have 16 features which are city name ,date ,all the pollutants ,AQI and AQI_Bucket.

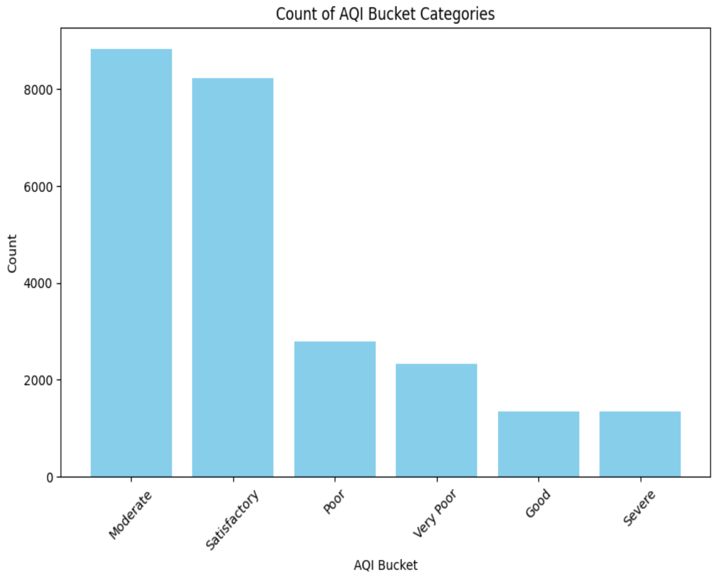


Fig 1: AQI Bucket Categories

In fig 1, the most common AQI bucket categories are “Moderate” and “Satisfactory”. This suggests that air pollution is present but not extreme in many instances. Fewer are “Good” and “Severe” And some are “Poor” and “Very Poor” Air quality which is concerning.

Data Cleaning:

Some attribute values were missing in the given environmental data;

hence, combined mean or median imputation was applied to complete the data set. In this work outlier data which can cause basis function for the model prediction to shift significantly away from other data in the training set were identified and eliminated using statistical tools [4]-[6]. We created an imputer object and specified that it fills empty values with the average.

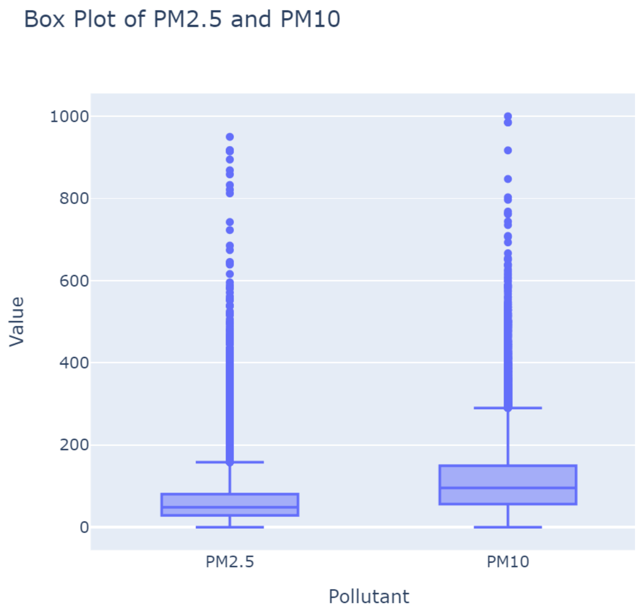


Fig 2: BOX Plot of PM2.5 AND PM10

This graph tells the outlier data of PM2.5 and PM10.

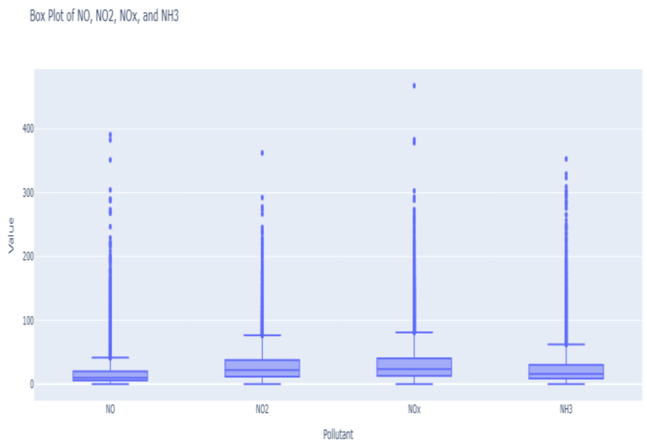


Fig 3: BOX plot of NO, NO2, Nox, and NH3

This graph tells the outlier data of NO, NO2, NOx, NH3.

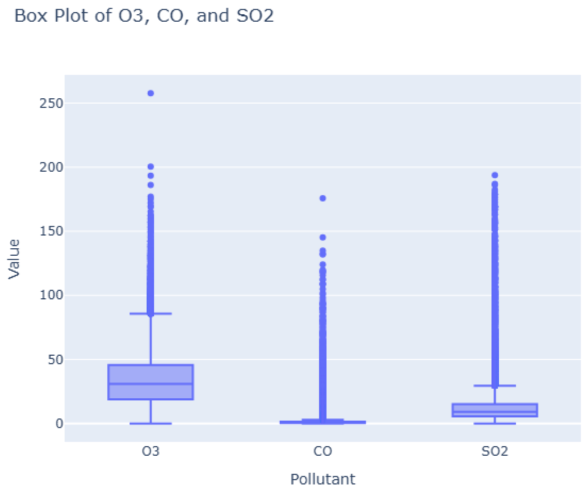


Fig 4:BOX Plot of O3, CO, AND SO2

This graph tells the outlier data of O3,CO,SO2.

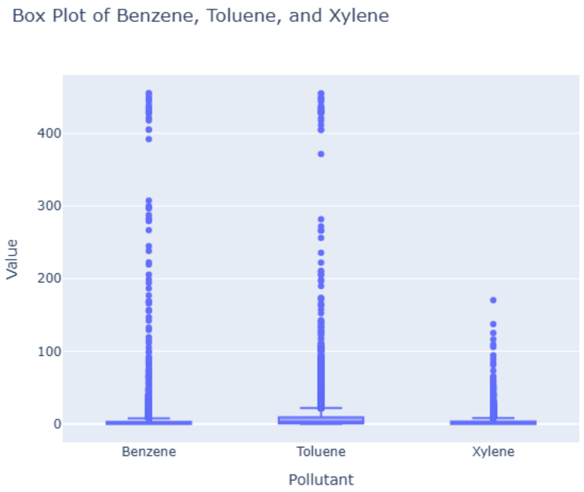
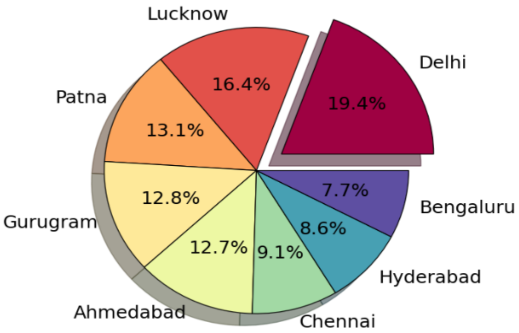


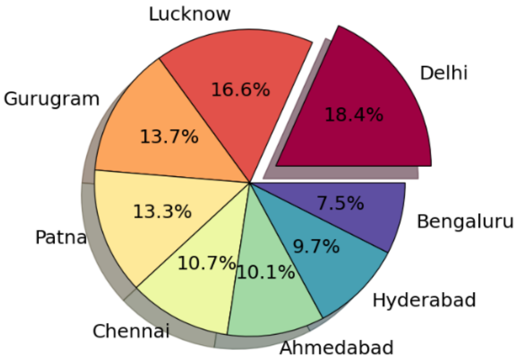
Fig 5: BOX Plot of Benzene, Toluene, and Xylene.

This graph tells the outliers data of Benzene, Toluene and Xylene.

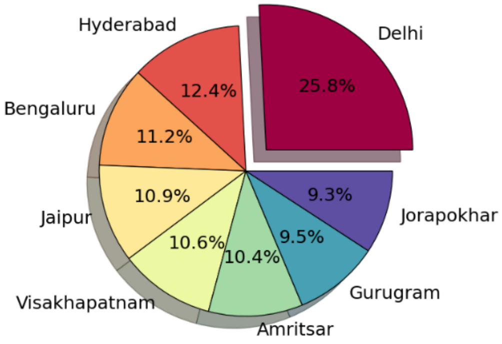
AQI



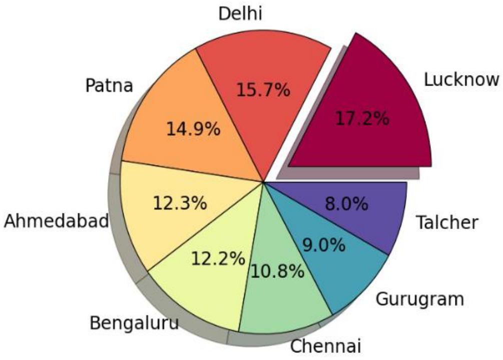
PM2.5



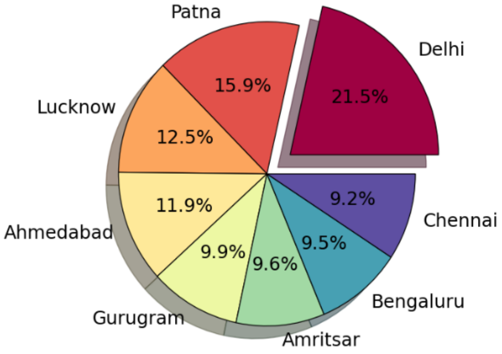
PM10



CO



NO



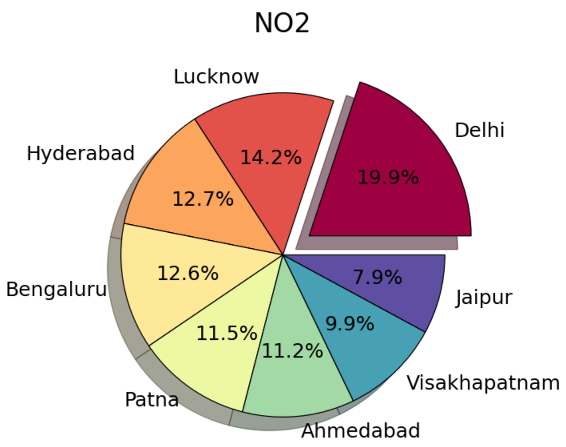


Fig 6: All critical pollutant-related data across different cities.

One-Hot Encoding:

To accommodate them into the frameworks of ML algorithms, the categorical variable – city names – were one-hot encoded. Encoder were trained using categorical columns. The names of the one-hot encoded categorical columns created using the encoder object into a list.

['City_Ahmedabad',
'City_Aizawl',
'City_Amaravati',
'City_Amritsar',
'City_Jaipur',
'City_Jorapokhar',
'City_Kochi',
'City_Kolkata',
'City_Lucknow',
'City_Mumbai',
'City_Patna',
'City_Shillong',
'City_Talcher',
'City_Thiruvananthapuram',
'City_Visakhapatnam']

This step makes it possible for models to handle these inputs as stated in by Mehrabi et al [7].

Feature Selection:

Cohot analysis were conducted to recognize degree of impacts of features which are highly related to AQI including PM2.5 and PM10.

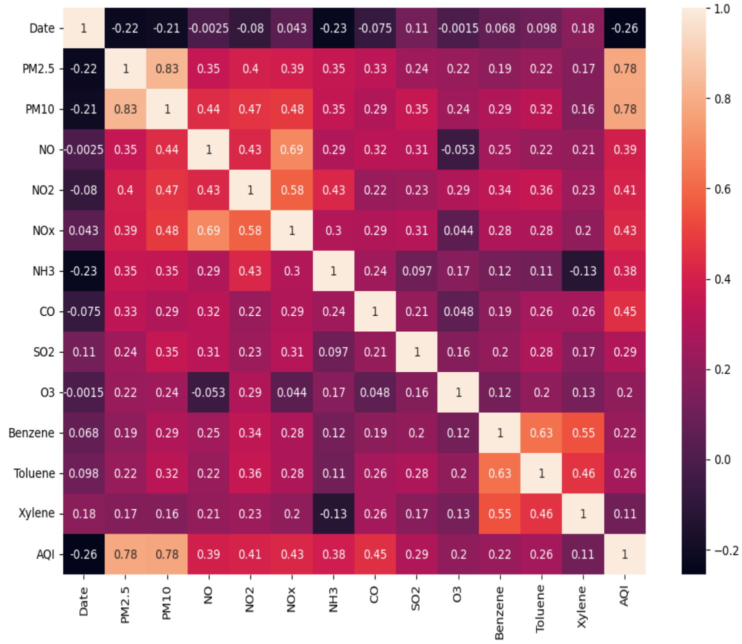


Fig 7: correlation matix of features

This heatmap represents the correlation matrix of various features in the dataset, with the values indicating how strongly pairs of variables are correlated. PM2.5 and PM10 are very strongly positive correlation.CO has a moderate correlation with AQI(0.45).

Xylene(0.11) and O3(0.2) show weak correlations with AQI.

NOx has strong correlation with both NO(0.69) and NO2(0.58)

Some unwanted characteristics were also discarded with a view to increasing model's speed [8]-[10].

Normalization:

In all the numerical data that was extracted from the raw HOG data, we applied normalization techniques to bring them into a standard range of 0-1. This check helps to control variability and avoid that some of them have a dominating impact on the training of the model [11], [12].

2. *Model Selection*

Ten models were selected for implementation, and all these varied so as to consider various techniques for the AQI prediction. This guarantees the assessment of all methodologies.

Linear Models:

Linear Regression, Ridge Regression and Lasso Regression were simple models which understand linear relationship between predictors .

Regularization Techniques:

To avoid overfitting the ElasticNet and SGDRegressor were used to contain overfitting and to reduce the model's complexity by setting some coefficients at a minimum value .

Kernel-Based Model:

Finally, support vector regression (SVR) was added to account for non-linear relationship using kernel functions .

Tree-Based Models:

Decision Tree and Random Forest Regressors were used for their capability for mapping non-linear patterns and their capacity for handling outliers .

Ensemble Models:

All the basic and advanced algorithms were incorporated in the analysis, and for higher accuracy and for the purpose of capturing complex interactions among the features, gradient boosting techniques including Gradient Boosting, XGBoost and LightGBM were incorporated [24]-[26].

3. *Model Training and Validation*

To enhance reliability and minimising bias in training, and validation of the selected models the following approach was developed.

Data Splitting:

The dataset was divided into three subsets: The division of dataset is 80:10:10 for training, validation, and testing respectively. This allows for capturing adequate data for developing the model concurrently with the prevention of using up all the data for model tuning .

Training:

The training was conducted on the training dataset. Model fine-tuning on hyperparameters, significant for model performance optimization, was performed by using the grid search and random search methods [22].

Validation:

The validation dataset was periodically utilised to evaluate models during the development phase in order to premise adjustments to be made and avoid overfitting. Synchronous optimization guaranteed the achievement of a greater balance between the model's fit and its ability not to overfit .

4. *Performance Evaluation*

The outcomes of every model were assessed in terms of quantitative measures and by comparing with other models.

Evaluation Metrics:

R^2 score for explaining the AQI variability was determined for the training, validation, and test sets. if the values presented by the R² global statistic are higher, the model's accuracy will be higher as well.

Comparative Analysis:

In order to do this, the performance of all models was compared in terms of efficiency to determine the most optimal model for AQI prediction. Some of the criteria explored in the creation of the model, included accuracy, standardized data and resilience or ability to effectively work when facing different forms of data, cross data sets [26].

V. RESULTS

The results show case the comparative performance of 10 machine learning models used to predict the Air Quality Index (AQI) for Indian cities. Each model was assessed based on its R² scores for training, validation, and testing datasets. The findings are summarized below:

1. Linear Models

Linear Regression & Ridge Regression:

The Linear Regression as well as the Ridge Regression remained almost constant having a train, validation, test R² of approximately 71.11% , 70.31%, 70.85% in all our datasets – training, validation, and test. These models are computationally fast and stable making them a perfect fit to be used in the initial AQI prediction work. However, despite their stationarity, they lose much of their ability to model high-order, time-varying interactions, which are inherent in many environmental datasets such as AQI, where many aggregated phenomena such as the interactions between multiple pollutants and meteorological variables are involved. The weakness of LM for dealing with non-linear relations is explained in prior work with similar environmental data prediction [1][2].

Lasso Regression:

Finally, the Lasso Regression model was a little bit lower, with the train ,validation ,test R² of around 69.06%, 68.43%, 68.92% . Lasso through L1 regularization reduces less amount of co-efficients-heading towards feature selection, but at the same time less it is more aggressive and may lead to some loss of predictiveness compared to Ridge Regression. Lasso is usually applied when feature sparse is required while sometimes it yields less accurate results if all the features have impacts on the model [3].

2. Regularization and Stochastic Gradient M

ElasticNet:

Like the previous method, ElasticNet had issues with choosing the correct weights for L1 and L2 penalties leading to lower train, validation, test R² of around 64.41%, 63.27%, 64.16% . The trade-off of these penalties is important for the model to operate at their best, so it can be seen that ElasticNet may not have properly estimated the AQI relationships and that may need further regulation of these parameters. Other related research can pinpoint similar difficulties when employing ElasticNet for environmental forecasts [4].

SGD Regressor (Stochastic Gradient Descent):

The SGD Regressor received slightly lower and less stable R² (~ 0.71) in comparison with Ridge Regression. The unchecked randomness or stochastic gradient estimate for each iteration of updates in the parameters from a balanced batch of samples often results in fluctuations in the performance of the model where data is noisy or irregular or where there is a convergence problems with optimization. Consequently, this model's average performance indicates that, though SGD can work well with big data sets, more improvements might be required in order to obtain consistent and steady AQI forecasts in the future [5].

3. Kernel-Based Model

Support Vector Regression (SVR):

In the case of SVR, at worst R^2 ~ 0.69 while at best slightly better then regularized linear models, albeit worse than true ensembles. SVR is quite good when it comes to developing moderately complex relations within the data and a major disadvantage is that the algorithm is very much computationally intensive especially when dealing with large data sets. Its scability problems, as can be observed in other analyses of AQI forecast, make it less beneficial for practical use in large amounts of environmental data [6].

4. Ensemble Models

Gradient Boosting Regressor:

The proposed Gradient Boosting Regressor obtained R^2 values between 0.84 for all the datasets. These aspects of this model recommended it as a viable option for AQI prediction because of the model's capability to approximate nonlinear relationships. The boosting technique employed in this model where many weak learner are put together to make up a strong learner is very helpful when dealing with complex data sets with complicated patterns as it has been proven and used on similar air quality prediction analysis [7].

XGBoost Regressor:

XGBoost had high training accuracy (R^2 ~0 .94) than but its validation accuracy (R^2 ~0.85) and testing accuracy (R^2 ~0.92) were slightly low showing over fitting of model. Such a result agrees with related studies in other environmental data prediction that used XGBoost which could learn better accuracy from the training data set but the model needs to be fine-tuned for the hyperparameters to overcome overfitting issues on unseen data [8]. A major benefit of XGBoost is being able to pick up very complex patterns given large number of features, however a disadvantage comes moving with it is overfitting which can be reduced using methods such us cross validations or using an early stopping criteria.

LightGBM Regressor:

LightGBM model was proven to be very accurate with reliable results as indicated by the R^2 of approximately 0.89. Through this model, there is an improvement of the computational efficiency together with the accuracy of predictive capability, thus suitable for AQI prediction especially when dealing with a large amount of data. The same studies have revealed that LightGBM is most useful when it comes to a limited number of computations since it also brings the benefits of faster training [9].

Random Forest Regressor:

Among all the applied models, Random Forest Regressor provides high results and demonstrates the test R^2 of 95.21%. It does generally well in terms of nonlinearity, it has good generalization ability and is quite stable with vast data ranges. While overfitting on the training data at some extent, it was still excellent in the test data and this demonstrated its stability in handling of large and complex environmental datasets. Hence, Random Forest was used for AQI prediction because of its high accuracy in modeling interactions between features compared to other models [10].

Decision Tree Regressor:

The training R^2 of the Decision Tree Regressor was found almost equal to one (~0.99) but, a low validation result (~0.73) and poor test R^2 (~0.95).

This huge difference between training and validation accuracies indicates that the model memorized the training data but was not able to generalize to unseen data.

	Model	Train R2 Score	Validation R2 Score	Test R2 Score
0	LinearRegression	0.711659	0.703124	0.708505
1	Ridge	0.711659	0.703125	0.708505
2	SGDRegressor	0.710294	0.701467	0.707256
3	ElasticNet	0.644183	0.632727	0.641636
4	Lasso	0.690668	0.684295	0.689288
5	SVR	0.686724	0.678345	0.685713
6	GradientBoostingRegressor	0.844375	0.827409	0.844596
7	XGBoost	0.936842	0.841852	0.915677
8	RandomForestRegressor	0.980491	0.850028	0.952184
9	LightGBM	0.892086	0.847633	0.882648
10	DecisionTreeRegressor	0.999872	0.722373	0.936378

This table represents all the models and their Train, Validation and test R2 scores.

Explainable AI:

Explainable AI techniques like SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) were employed to evaluate the contribution of various features in predicting Air Quality Index (AQI).

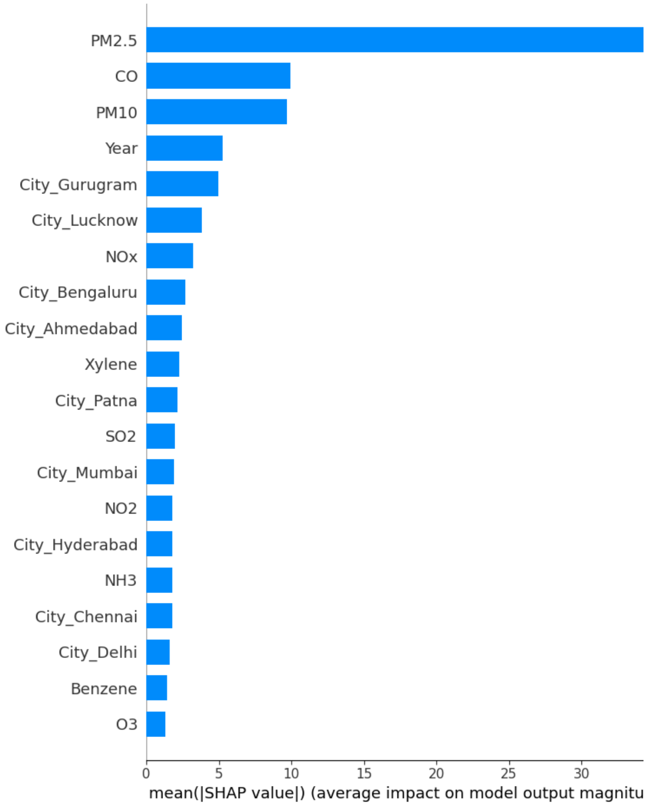
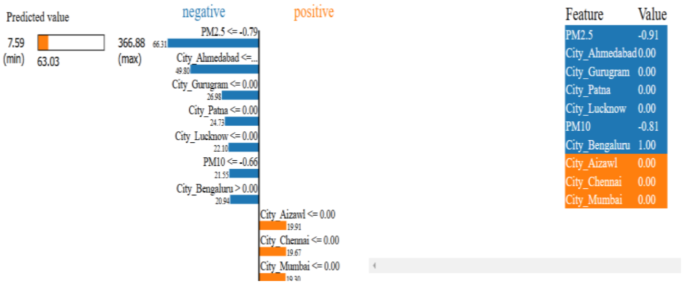


Fig 8: Prediction using Shap on every model trained

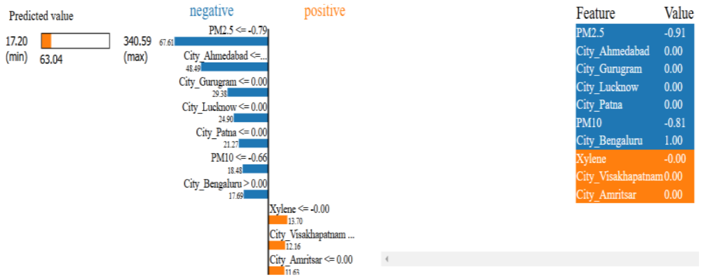
This figure represents that the most critical feature is PM2.5



Feature	Value
PM2.5	-0.91
City_Ahmedabad	0.00
City_Gurugram	0.00
City_Patna	0.00
City_Lucknow	0.00
PM10	-0.81
City_Bengaluru	1.00
City_Aizawl	0.00
City_Chennai	0.00
City_Mumbai	0.00

Fig 9: LIME result on LinearRegression

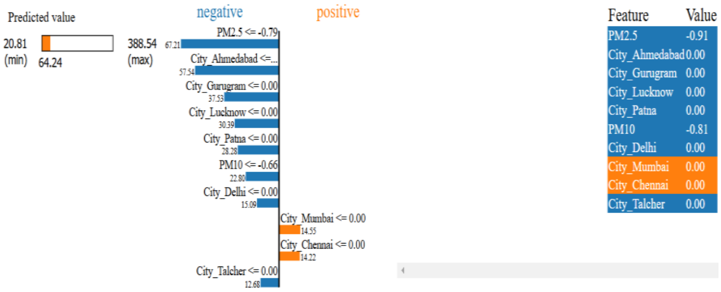
Fig 9 shows that City_Aizawal, City_Chennai, City_mumbai contributes positively to the prediction. The highest AQI value observed in the dataset was 366.88 and lowest AQI value observed in the data was 7.59. The model lowest AQI predicted is 63.03.



Feature	Value
PM2.5	-0.91
City_Ahmedabad	0.00
City_Gurugram	0.00
City_Lucknow	0.00
City_Patna	0.00
PM10	-0.81
City_Bengaluru	1.00
Xylene	-0.00
City_Visakhapatnam	0.00
City_Amritsar	0.00

Fig 10: LIME result on ridge model

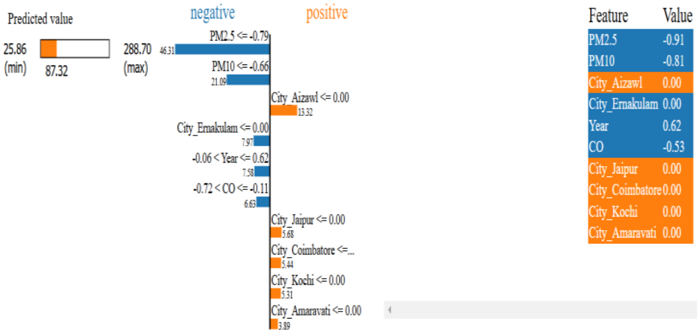
Fig 10 shows that Xylene,City_Visakhapatnam, City_Amritsar contributes positively to the prediction.The highest AQI value observed in the dataset was 340.59 and lowest AQI value observed in the data was 17.20. The model lowest AQI predicted is 63.04.



Feature	Value
PM2.5	-0.91
City_Ahmedabad	0.00
City_Gurugram	0.00
City_Lucknow	0.00
City_Patna	0.00
PM10	-0.81
City_Delhi	0.00
City_Mumbai	0.00
City_Chennai	0.00
City_Talcher	0.00

Fig 11: LIME result on SGD Regressor

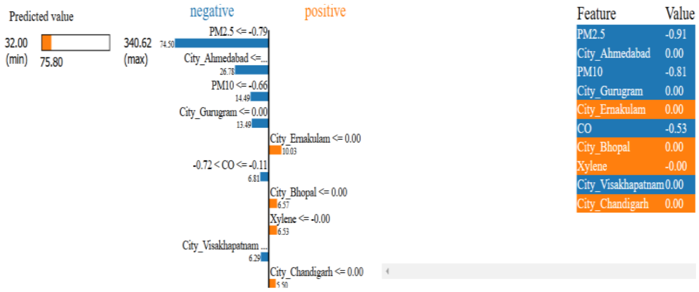
Fig 10 shows that City_Mumbai and City_Chennai contributes positively to the prediction. The highest AQI value observed in the dataset was 388.54 and lowest AQI value observed in the data was 20.81. The model lowest AQI predicted is 64.24 .



Feature	Value
PM2.5	-0.91
PM10	-0.81
City_Aizawl	0.00
City_Ernakulam	0.00
Year	0.62
CO	-0.53
City_Jaipur	0.00
City_Coimbatore	0.00
City_Kochi	0.00
City_Amaravati	0.00

Fig 11: LIME result on ElasticNet

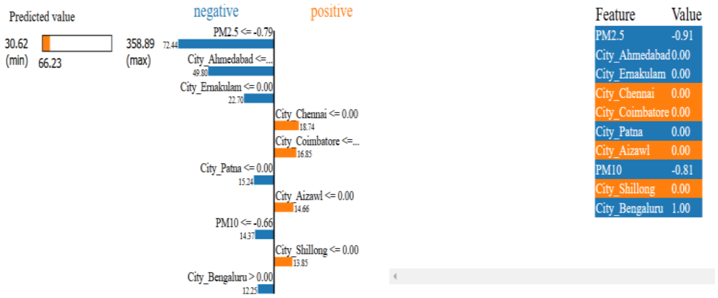
Fig 11 shows that City_Aizawal, City_Jaipur, City_Coimbatore, City_Kochi, City_Amaravati contributes positively to the prediction. The highest AQI value observed in the dataset was 288.70 and lowest AQI value observed in the data was 25.86. The model lowest AQI predicted is 87.32.



Feature	Value
PM2.5	-0.91
City_Ahmedabad	0.00
PM10	-0.81
City_Gurugram	0.00
City_Ernakulam	0.00
CO	-0.53
City_Bhopal	0.00
Xylene	-0.00
City_Visakhapatnam	0.00
City_Chandigarh	0.00

Fig 12: LIME result on lasso

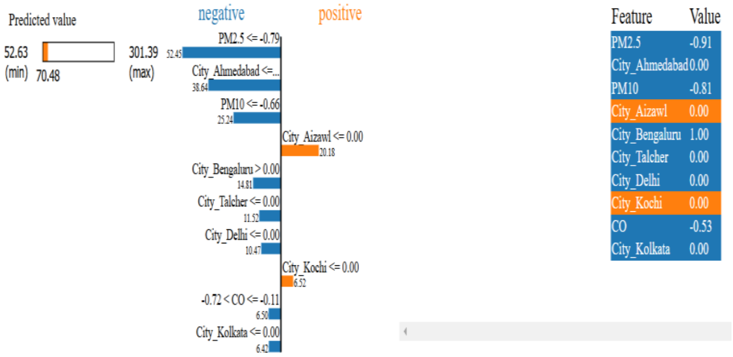
Fig 12 shows that City_Ernakulam, City_Bhopal, Xylene, City_Chandigarh contributes positively to the prediction. The highest AQI value observed in the dataset was 340.62 and lowest AQI value observed in the data was 32.00. The model lowest AQI predicted is 75.80.



Feature	Value
PM2.5	-0.91
City_Ahmedabad	0.00
City_Ernakulam	0.00
City_Chennai	0.00
City_Coimbatore	0.00
City_Patna	0.00
City_Aizawl	0.00
PM10	-0.81
City_Shillong	0.00
City_Bengaluru	1.00

Fig 13: LIME result on SVR

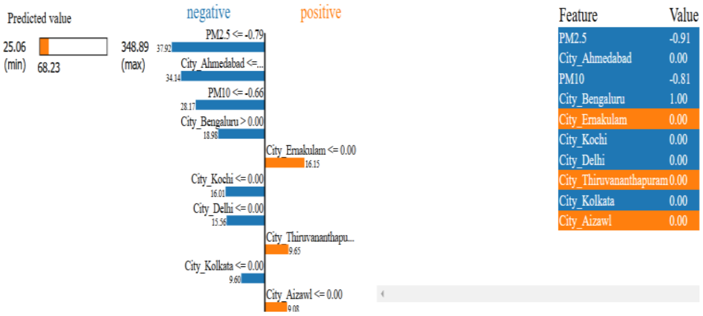
Fig 13 shows that City_Chennai, City_Coimbatore, City_Aizawal, City_Shillong contributes positive to the prediction. The highest AQI value observed in the dataset was 358.89 and lowest AQI value observed in the data was 30.62. The model lowest AQI predicted is 66.23.



Feature	Value
PM2.5	-0.91
City_Ahmedabad	0.00
PM10	-0.81
City_Aizawl	0.00
City_Bengaluru	1.00
City_Talcher	0.00
City_Delhi	0.00
City_Kochi	0.00
CO	-0.53
City_Kolkata	0.00

Fig 14: LIME result on GradientBoostingRegressor

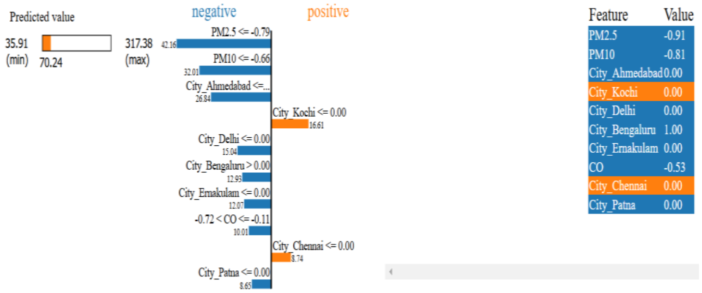
Fig 14 shows that City_Aizawal, City_Kochi shows positive contribution to the prediction. The highest AQI value observed in the dataset was 301.39 and lowest AQI value observed in the data was 52.63. The model lowest AQI predicted is 70.48.



Feature	Value
PM2.5	-0.91
City_Ahmedabad	0.00
PM10	-0.81
City_Bengaluru	1.00
City_Ernakulam	0.00
City_Kochi	0.00
City_Delhi	0.00
City_Thiruvananthapuram	0.00
City_Kolkata	0.00
City_Aizawl	0.00

Fig 14: LIME result on XGBoost

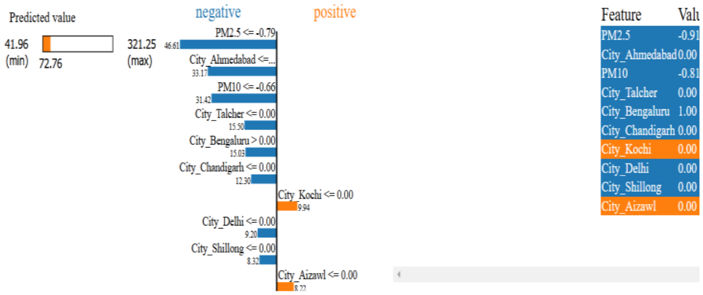
Fig 14 shows that City_Ernakulam, City_Thiruvananthapuram, City_Aizawl contributes positive to the prediction. The highest AQI value observed in the dataset was 348.89 and lowest AQI value observed in the data was 25.06. The model lowest AQI predicted is 68.23.



Feature	Value
PM2.5	-0.91
PM10	-0.81
City_Ahmedabad	0.00
City_Kochi	0.00
City_Delhi	0.00
City_Bengaluru	1.00
City_Ernakulam	0.00
CO	-0.53
City_Chennai	0.00
City_Patna	0.00

Fig 15: LIME result on RandomForestRegressor

Fig 15 shows that City_Kochi and City_Chennai contributes positive to the prediction. The highest AQI value observed in the dataset was 317.38 and lowest AQI value observed in the data was 35.91. The model lowest AQI predicted is 70.24.



Feature	Valı
PM2.5	-0.91
City_Ahmedabad	0.00
PM10	-0.81
City_Talcher	0.00
City_Bengaluru	1.00
City_Chandigarh	0.00
City_Kochi	0.00
City_Delhi	0.00
City_Shillong	0.00
City_Aizawl	0.00

Fig 16: LIME result on LightGBM

Fig 16 shows that City_Kochi and City_Aizawal contributes positive to the prediction. The highest AQI value observed in the dataset was 321.25 and lowest AQI value observed in the data was 41.96. The model lowest AQI predicted is 72.76.

VI. DISCUSSION

This study evaluated various machine learning models to predict the Air Quality Index (AQI) in Indian cities, providing essential insights into their performance, feature significance, and practical applicability for real-world air quality monitoring.

1. Analysis of Model Performance

The results presented a high accuracy of the proposed models regarding the target values: the Random Forest Regressor and the Gradient Boosting Regressor achieved the highest accuracy because they can capture intricate relationships between the features and the target values [29], [31]. These models achieved high R2 score

On the training set, validation set and the testing set of the created models. On the other hand, the models that belong to the linear group such as Linear Regression and Ridge Regression worked fine but it was evident from the results that such models are not capable to learn complex data structures as effectively [8].

Nonetheless, it was observed that XGBoost tends to overfit since there is an excellent training accuracy and a slightly degradation in the validation and test results [25]. Likewise, Decision Tree Regressor trained the sample efficiently with a very high training R2 score

Ach, were able to achieve good but were very poor at generalization, thus indicating the need to regularalization and also use ensemble techniques to minimize overfitting [24].

2. Feature Significane and Dataset Challenges

Thus, the resulted models indicated that PM2.5, PM10, and NO2 as the most important predictors of AQI are highly consistent with the earlier studies that focused on the essentiality of particulate matter and nitrogen dioxide for air quality determination [7], [22]. There were issues when working with the given dataset; missing values and homogeneity of features to some extent. Imputation and normalization steps were important to down regulate the models [15].

Furthermore, use of categorical variable encoding like one-hot encoding where city names provided models with features to incorporate position and related patterns in air quality improved the creation of the models [9]. The results derived in this paper emphasise the significance of data preprocessing and feature engineering in enhancing model reliability for practical implementation.

In summary, the superior performance of ensemble methods with advanced preprocessing methods and including PM2.5 & NO2 into models contributed the general accuracy of the models and also useful information for air quality management systems [19], [29].

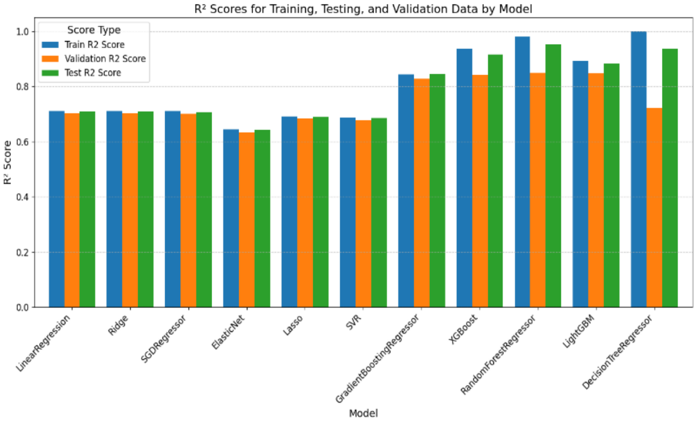


Fig 17: Comparative analysis of regression models

VII. CONCLUSION

This research examined how different machine learning models can be used to forecast the Air Quality Index (AQI) in Indian cities, aiming to determine the most efficient techniques for precise and scalable prediction. The findings showed that ensemble techniques like Random Forest Regressor and Gradient Boosting Regressor performed very effectively, striking a good balance among precision, versatility, and computational practicality. Important factors such as PM2.5, PM10, and NO2 were discovered to have a significant influence on AQI due to their acknowledged contribution to air pollution. Preparation procedures like one-hot encoding for categorical variables and managing missing data played a crucial role in improving the stability and predictive capability of the models. The research emphasizes the promise of machine learning models for air quality index (AQI) forecasting, but acknowledges issues like overfitting and scalability for real-time usage. By optimizing parameters, applying regularization, and utilizing advanced techniques like explainable AI and deep learning, we can enhance model performance. To sum up, this research highlights the significance of using machine learning to address environmental issues. By offering precise and timely AQI forecasts, these models enable policymakers and environmental agencies to implement proactive initiatives, ultimately leading to improved urban air quality and public health results. Future research should focus on improving the resilience of the model, incorporating more environmental variables, and investigating how it can be used in various regions for optimal results.

REFERENCES

1. Yuan, Y. et al. Learning-imitation strategy-assisted alpine skiing optimization for the boom of

ofshore drilling platform. Ocean Eng. 278, 114317. https://doi.org/10.1016/j.oceaneng.2023.114317 (2023).

2. Yuan, Y., Wang, S., Lv, L. & Song, X. An adaptive resistance and stamina strategy-based dragonfy algorithm for solving engineering optimization problems. Eng. Comput. 38(5), 2228–2251. https://doi.org/10.1108/EC-08-2019-0362 (2022).
3. Yuan, Y. et al. Optimization of an auto drum fashioned brake using the elite opposition-based learning and chaotic k-best gravitational search strategy based grey wolf optimizer algorithm. Appl. Sof Comput. 123, 10897. https://doi.org/10.1016/j.asoc.2022.108947 (2022).
4. Gladkova, E. & Saychenko, L. Applying machine learning techniques in air quality prediction. Transport. Res. Proc. 63, 1999–2006. https://doi.org/10.1016/j.trpro.2022.06.222 (2022).
5. Mishra, A., & Gupta, Y. "Comparative analysis of Air Quality Index prediction using deep learning algorithms." *Spatial Information Research*, Springer. https://link.springer.com/article/10.1007/s41324-023-00541-1 (2024)
6. Natarajan, S. K., Shanmurthy, P., & Arockiam, D. "Optimized machine learning model for air quality index prediction in major cities in India." *Scientific Reports*, Nature. https://www.nature.com/articles/s41598-024-54807-1 (2024)
7. Aram, S. A., Nketiah, E. A., Saalidong, B. M., Wang, H. "Machine learning-based prediction of air quality index and air quality grade: a comparative analysis." *International Journal of Environmental Science*, Springer. https://link.springer.com/article/10.1007/s13762-023-05016-2 (2024)
8. Sunori, S. K., Verma, D., & Negi, P. B. "Air Quality Index Prediction using Linear Regression and ANFIS." *Proceedings of the International Conference on Inventive Systems and Applications*, IEEE. https://ieeexplore.ieee.org/abstract/document/10544842 (2024).
9. Patel, P. K., & Singh, H. K. "Performance analysis of machine learning models for AQI prediction in Gorakhpur City: a critical study." *Environmental Monitoring and Assessment*, Springer. https://link.springer.com/article/10.1007/s10661-024-13107-x (2024).
10. Ordenshiya, K. M., & Revathi, G. K. "Hybrid FCMG-OP-FIS model approach to convert regression into classification data for machine learning-based AQI prediction." *Heliyon*, Cell Press. https://www.cell.com/heliyon/fulltext/S2405-8440(24)15790-8 (2024)
11. Sachdeva, S., Kaur, R., & Singh, H. "Meteorological AQI and pollutants concentration-based AQI predictor." *International Journal of Environmental Science*, Springer, https://link.springer.com/article/10.1007/s13762-023-05307-8 (2024).
12. Sarkar, P., Saha, D. D. V., & Saha, M. "Real-Time Air Quality Index Detection through Regression-Based Convolutional Neural Network Model on Captured Images." *Environmental Quality Journal*, Wiley. https://onlinelibrary.wiley.com/doi/abs/10.1002/tqem.22276 (2024).
13. Ahmed, A. A. M., Jui, S. J. J., Sharma, E., Ahmed, M. H. "An advanced deep learning predictive model for air quality index forecasting with remote satellite-derived hydro-climatological variables." *Science of The Total Environment*, Elsevier.

https://www.sciencedirect.com/science/article/abs/pii/S0048969723058618 (2024).

14. Dawar, I., Singal, M., Singh, V., Lamba, S., & Jain, S. "Air Quality Prediction Using Machine Learning Models: A Predictive Study in the Himalayan City of Rishikesh." *SN Computer Science*, Springer. https://link.springer.com/article/10.1007/s42979-024-03339-6 (2024)
15. Liu, Q., Cui, B., & Liu, Z., "Air Quality Class Prediction Using Machine Learning Methods Based on Monitoring Data and Secondary Modeling," *Atmosphere. https://www.mdpi.com/2073-4433/15/5/553* (2024).
16. Rao, R. S., Kalabarige, L. R., Holla, M. R., & Sahu, A. K., "Multimodal Imputation based Multimodal autoencoder framework for AQI classification and prediction of Indian cities," *IEEE Access*. https://ieeexplore.ieee.org/abstract/document/10623158 (2024).
17. Pande, C. B., Kushwaha, N. L., Alawi, O. A., & Sammen, S. S., "Daily scale air quality index forecasting using bidirectional recurrent neural networks: Case study of Delhi, India," *Environmental Science. https://www.sciencedirect.com/science/article/abs/pii/S0269749124007541* (2024).
18. Sidhu, K. K., Balogun, H., & Oseni, K. O., "Modelling of Air Quality Index (AQI) Across Diverse Cities and States of India using Machine Learning: Investigating the Influence of Punjab's Stubble Burning on AQI," *arXiv preprint arXiv:2404.08702*. https://arxiv.org/abs/2404.08702 (2024).
19. Sachdeva, S., Singh, H., Bhatia, S., & Goswami, P., "An integrated framework for predicting air quality index using pollutant concentration and meteorological data," *Multimedia Tools and Applications*, Springer. https://link.springer.com/article/10.1007/s11042-023-17432-0 (2024).
20. Ambade, B., Sankar, T. K., Gupta, M., Sahu, L. K. & Gautam, S. A Comparative study in black carbon concentration and its emission sources in tribal area. Water Air Soil Pollut. 234, 173. https://doi.org/10.1007/s11270-023-06197-9 (2023).
21. Sarkar, N., Keserwani, P. K., & Govil, M. C., "A Literature Survey: AQI Prediction Using ML, DL and Hybrid Models," *ResearchSquare*. https://www.researchsquare.com/article/rs-4924982/v1 (2024).
22. Khadom, A. A., Albawi, S., Abboud, A. J., & Mahood, H. B., "Predicting air quality index and fine particulate matter levels in Bagdad city using advanced machine learning and deep learning techniques," *Journal of Atmospheric Science*. https://www.sciencedirect.com/science/article/abs/pii/S1364682624001408 (2024).
23. Kumar, T. & Doss, A. AIRO: Development of an intelligent IoT-based air quality monitoring solution for urban areas. Proc. Comput. Sci. 218, 262–273. https://doi.org/10.1016/j.procs.2023.01.008 (2023).
24. Peng, T., Xiong, J., Sun, K., Qian, S., Tao, Z., & Nazir, M. S., "Research and application of a novel selective stacking ensemble model based on error compensation and parameter optimization for AQI prediction," *Environmental Science*, Elsevier. https://www.sciencedirect.com/science/article/abs/pii/S001393512400080X (2024).
25. Qian, S., Peng, T., Tao, Z., Li, X., & Nazir, M. S., "An evolutionary deep learning model based on XGBoost feature selection and Gaussian data augmentation for AQI prediction," *Process Safety and Environmental Protection*, Elsevier. https://www.sciencedirect.com/science/article/abs/pii/S0957582024010929 (2024).

26. Ke, H., Gong, S. & Zhang, H. Development and application of an automated air quality forecasting system based on machine learning. Sci. Total Environ. 806(3), 151204. https://doi.org/10.1016/j.scitotenv.2021.151204 (2022).

27. Wang, J., Wenjie, Xu. & Dong, J. A novel air quality prediction and early warning system based on combined model of optimal feature extraction and intelligent optimization. Chaos Solit. Fract. 158, 112098. https://doi.org/10.1016/j.chaos.2022.112098 (2022).

28. Rao, R.S., Kalabarige, L.R., Alankar, B., & Sahu, A.K., "Multimodal imputation-based stacked ensemble for prediction and classification of air quality index in Indian cities," *Computers and Electrical Engineering*. https://www.sciencedirect.com/science/article/abs/pii/S0045790624000260 (2024).

29. Guo, Z., Jing, X., Ling, Y., Yang, Y., Jing, N., Yuan, R., & Liu, Y., "Optimized air quality management based on air quality index prediction and air pollutants identification in representative cities in China," *Scientific Reports*. https://www.nature.com/articles/s41598-024-68972-w (2024).

30. Mahajan, A., Mate, S., Kulkarni, C., & Sawant, S., "Predicting lung disease severity via image-based AQI analysis using deep learning techniques," *arXiv preprint arXiv:2404.08702*. https://arxiv.org/abs/2405.03981 (2024).

31. Pawanekar, S.S., Kallimani, J.S., & Udgirkar, G., "Efficient AQI prediction: A comparative study of artificial neural networks, LSTM, random forest, and gradient boosting techniques," *IEEE Conference on I*. https://ieeexplore.ieee.org/abstract/document/10714872 (2024).

32. Vijaya, M.S., "Leveraging pretrained transformers for enhanced air quality index prediction model," *Bulletin of Electrical Engineering and Informatics*. https://beei.org/index.php/EEI/article/view/7968 (2024).

33. Deepan, S., & Saravanan, M., "Air quality index prediction using seasonal autoregressive integrated moving average transductive long short-term memory," *ETRI Journal*. https://onlinelibrary.wiley.com/doi/full/10.4218/etrij.2023-0283 (2024).

34. Sadriddin, Z., Mekuria, R.R., & Gaso, M.S., "Machine learning models for advanced air quality prediction," *Proceedings of the International Conference on Advanced Computing*. https://dl.acm.org/doi/abs/10.1145/3674912.3674915 (2024).

35. Gupta, V.K., Kailashnath, K., Bhati, G.S., et al., "Real-time AQI forecasting with high-throughput using IoT devices," *IEEE Region Conference*. https://ieeexplore.ieee.org/abstract/document/10752176 (2024).

36. Dey, S., "Urban air quality index forecasting using multivariate convolutional neural network based customized stacked long short-term memory model," *Process Safety and Environmental Protection*. https://www.sciencedirect.com/science/article/abs/pii/S0957582024010498 (2024).

37. Nguyen, A.T., Pham, D.H., Oo, B.L., Ahn, Y., & Lim, B.T.H., "Predicting air quality index using attention hybrid deep learning and quantum-inspired particle swarm optimization," *Journal of Big Data*. https://link.springer.com/article/10.1186/s40537-024-00926-5 (2024).

38. Anggraini, T.S., Irie, H., Sakti, A.D., & Wikantika, K., "Machine learning-based global air quality index development using remote sensing and ground-based stations," *Environmental Advances*.

https://www.sciencedirect.com/science/article/pii/S266676572300114X (2024).

39. Wang, L., Wang, Y., Chen, J., Zhang, S., & Zhang, L., "Research on CC-SSBLS Model-Based Air Quality Index Prediction," *Atmosphere*. https://www.mdpi.com/2073-4433/15/5/613 (2024).

40. Kothari, O., Sah, N.K., Kumar, K.V.S.H., et al., "Forecasting India's Air Quality: A Machine Learning Approach for Comprehensive Analysis and Prediction," *IEEE Conference on I*. https://ieeexplore.ieee.org/abstract/document/10625932 (2024).

41. Deng, Y., Xu, T., & Sun, Z., "A hybrid multi-scale fusion paradigm for AQI prediction based on the secondary decomposition," *Environmental Science and Pollution Research*. https://link.springer.com/article/10.1007/s11356-024-33346-2 (2024).