Air Quality Index Prediction Using Various Machine Learning Algorithms

Mann Bajpai, Tarun Jain, Aditya Bhardwaj, Horesh Kumar, and Rakesh Sharma

Abstract One of the most critical factors for human survival is air. The quality of air inhaled by humans affects their health and lives significantly. The continuously rising air pollution is a significant concern as it threatens human health and is an environmental issue in many Indian cities. A proper AQI prediction system will help tackle the problem of air pollution more efficiently and mitigate the health risks it causes. Government agencies use the Air Quality Index, a number to indicate the pollution level of the air to the public. It qualitatively illustrates the current state of the air. Aggregate values of PM2.5, PM10, CO_2, NO_2, and SO_2 have been taken to forecast the AQI for Pune city using the dataset collected by Pune Smart City Development Corporation Limited and IISc in 2019. This study aims to find the machine learning method which forecasts the most accurate AQI and its analysis.

Keywords Air quality index · Prediction · Machine learning · Linear regression · Random forest · KNN

1 Introduction

Over the past decade, the continuous rise in the air pollutants level in the atmosphere is one of the major emerging issues faced by the world. This problem has been fuelled by factors such as urbanization, industrialization, rapid growth in India's population and vehicles in the country, etc. The air pollutants mainly include nitrogen dioxide, carbon monoxide, ozone and sulfur dioxide, etc. the Air quality

M. Bajpai · T. Jain (✉) · R. Sharma
Manipal University Jaipur, Jaipur, Rajasthan, India

A. Bhardwaj
Bennett University, Greater Noida, Uttar Pradesh, India
e-mail: aditya.bhardwaj@bennett.edu.in

H. Kumar
Greater Noida Institute of Technology, Greater Noida, Uttar Pradesh, India

M. Kumar et al. (eds.), *6G Enabled Fog Computing in IoT*,
https://doi.org/10.1007/978-3-031-30101-8_4

Fig. 1 AQI index range

index (AQI) has been devised to get an idea of the concentration of pollutants in the air and their quality. It has been divided into 6 categories with specific colors and health hazard levels [1]. AQI ranges from 0–500, and a higher AQI value indicates greater levels of air pollution. From Fig. 1**,** it can be observed that for the range 0–50, the air quality is satisfactory; between 51 and 100, it's acceptable but might be a risk for some people; in the range of 101–150 effects on health can be experienced by sensitive groups' members. The range of AQI values 151–200 is considered unhealthy and can adversely affect human health. A content of 201–300 is considered very harmful, and AQI values higher than 300 are considered hazardous [2].

Pune has witnessed over 733 deaths per million people because of cardiovascular diseases developed due to exposure to (PM10 and SO_2) due to air pollution. People already suffering from lung diseases like pneumonia and asthma are more susceptible to lung and heart diseases on exposed to polluted air. The inhalation of air contaminated with PM2.5 and PM10 makes self-purifying the human immune system very difficult. Results from the emission inventory for PM2.5 of Pune city showed that half of the primary emissions come from the transport sector. Air pollution was also directly emitted from sources like industrial operations, resuspended dust, solid fuel combustion, etc. [3]. This paper aims to find the best air quality index prediction method for Pune city by implementing and evaluating various machine learning algorithms like regression, support vector machine, k-nearest neighbor algorithm, and random forest. In this project, firstly, the selection of significant and relevant features has been made, following which there has been the implementation of a different machine learning model for the prediction of the pollutants to yield highly accurate and error-free results for the AQI estimation.

The remainder of this paper is organized as follows. Section 2 presents a state-of-the-art existing related work. The working methodology of the proposed work is discussed in Sect. 3. Section 4 presents the performance evaluation parameters. Finally, results, followed by concluding remarks, are presented in Sects. 5 and 6.

2 Literature Review

Several studies and research have been done to predict the air quality index of different cities. These studies mainly focus on accurate estimation of the air quality index of cities for the formulation of better plans and preventive measures for controlling the air pollution levels in developing smart cities as well as the existing ones. The data collected by the sensors and actuators help understand the correlation between the different pollutants, the major pollutants causing severe damage to the human respiratory system, and the pattern in the emission of these pollutants.

In [4], Moolchand Sharma et al. presented a model to predict the AQI with an emphasis on performance and accuracy. Its robustness and accuracy were validated after testing six different machine learning classifiers. Different combinations of classifiers were tested to check which one gave the most accurate results. An accuracy of 99.7% was achieved using a Decision Tree, which increased by 0.02% when the Random tree classifier was applied. The study aimed to show the possibility of improvement in the AQI forecast using nonlinear machine learning algorithms.

In [5], Mehzabeen Mannan et al. have reviewed studies from different countries regarding the progress made in IAQ research, examining parameters like volatile matters, PM, carbon dioxides, and monoxides. Most works are focused on VOCs using gas chromatography-mass spectrometry for their analysis. Significant contributors to VOC concentrations are building structure and materials; for PM levels, it's the construction process and human movement and for indoor NH_3 it's concrete additives as per the research that has been reviewed in this work.

The pattern and air pollution trends have been analyzed for Delhi, Chennai, and Kolkata in [6] by Shrabanti Dutta et al. The air quality index has been developed using four major pollutants for 3 years.PM10 is the major pollutant affecting the air in all three cities. The climatic conditions are a significant factor in a place's air pollution along with the pollutant's seasonal distribution. The only pollutant accomplishing the NAAQ standard is SO_2. The rest of the pollutants have emissions much higher than the NAAQ standards.

In [7], C. Amruthadevi et al. have compared different machine learning algorithms like Statistical multilevel regression, Neuro-Fuzzy, Deep Learning Long-Short-Term memory(DL-LTSM) and Non- Linear Artificial Neural Networks (ANN). Results show that the DL-LSTM is the most suitable algorithm for analyzing and forecasting pollutants in the air. Parameters used to compare the results include RMSE, MAPE and R^2. In R^2, a deviation in the range of 0.71–0.89 is there during the prediction of the pollutants' contamination level.

In [8] to forecast Wuhan city's air quality index, Al-Qaness et al. have proposed an adaptive neuro-fuzzy inference system. It has been named PSOSMA as it uses a slime mold algorithm (SMA), modified meta-heuristic algorithm (MH) and Particle Swarm Optimizer has been used to improve its performance (PSO). The data has been trained to predict the air pollutants like PM 2.5, SO_2, CO_2 and NO_2. The

performance of this proposed modified ANFIS, which uses PSOSMA is better than its counterparts.

R. Senthil Kumar et al. in [9], have proposed a method for analyzing and visualizing Bengaluru's AQI. Attribute selection methods like correlation matrix and decision tree have been used to analyze the important pollutants which are selected. The J48 decision tree has been used to select features with maximum gain ratio. Input data's similar features have been removed using Correlation matrix analysis. Data analysis and the calculation of results have been done using Expectation Maximization (EM) Clustering.

In [10], RM Fernando et al. predicted the concentration of PM2.5 in Columbo using the concentrations of air pollutants. The training and evaluation of the prediction model were done using machine learning algorithms like SVM, KNN, Random forest and Multiple Linear-Regression. The Random forest model had over 85% accuracy.

In [11], Ditsuhi Iskandaryan et al. studied the research works related to air quality prediction. The main observations were that most of the datasets being used, over 94.6%, were meteorological. At the same time, the rests were spatial and temporal data, and a large majority of the studies used open datasets too. To supplement the data gathered using air quality sensors, about 26 datasets have been used, which include 'Temporal', 'MET', 'Social media and 'Spatial' etc. The parameter of the analysis of the papers includes prediction target, type of dataset, data rate, algorithm, case studies, time granularity, etc. The authors found Random Forest, Support vector machine, and LSTM to be the most widely used methods for predicting particulate matter.

In [12] Kadir Diler Alemdar et al. have proposed a geographic information system-based approach for redesigning mitigation strategies in accordance with the risk classification and assumed scenario. The study aimed to demonstrate the changes in the mobility of traffic and the improvement in air quality due to the restrictions applied during the pandemic. It was observed that the level of air pollutants like PM10, CO, SO_2, NO_2 etc. decreased significantly and the speed of traffic improved.

In [13], Laura Gladson et al. have developed an air quality index that shows the health risks caused by outdoor pollution in children. The creation of indices evaluated the impact of air pollutants like fine matter, ozone, nitrogen dioxide, etc. The indices presented normal distributions of locally scaled index values after adjustment and use values of the daily index of air pollutants. The author has provided the resources and steps for applying the final adjusted indices.

In [14], Xiali Sun et al. have proposed an IPSO – BP forecasting model which optimizes BP neural network's threshold and particle swarm weights. It's based on an improvised PSO-BP algorithm. The model improved prediction accuracy in comparison with BP and GA-BP. The particles search the optimal initial Value and BP's threshold value to create an IPSO-BP model for forecasting. This enhances the prediction accuracy and reduces the MAE too.

In [15], Narathep Phruksahiran et al. have proposed a geographically weighted predictor method for the hourly prediction of variables. The methodology combines GWP techniques and machine learning algorithms for the prediction of pollutants in the air on an hourly basis. It has better and more accurate forecast in all horizons compared to the existing prediction methods, improving the AQI prediction accuracy.

In [16], Manmeet Singh et al. analyzed the air quality across the globe using merged products of air pollutants and spatiotemporal resolution satellites during the COVID-19 lockdowns and found significant reductions in the concentrations of Nitrogen Dioxide, PM2.5, and aerosol optical depth.

In [17], Subhashini Penetiet al. introduced blockchain-defined networks and a grey wolf-optimized modular neural network approach to managing intelligent environment security. User authentication-based blocks are designed for security in the construction, translation, and application layers. In IoT-enabled innovative applications, the maintenance of latency and computational resource utilization is done by applying optimized neural networks. In the results, the system ensures higher security and lower latency as compared to deep learning networks and multi-layer perceptron.

In [18], Dmitry Kochetkov et al. studied and discussed the implementation and development of 5G-based technologies for an urban environment. In the selected areas, a scientometric analysis of the field and a study of patent landscapes was conducted for the analysis of new technologies. The study of citation patterns was the object of scientometric analysis.

In [19], Ting Li et al. proposed a novel method for data collection from multiple sensor devices by partnering vehicles and unmanned aerial vehicles (UAV) in IoT. Using a genetic algorithm, vehicle collectors are selected for data collection through sensors following which collection routes of UAVs are planned using a novel deep reinforcement learning(DLR) based- route policy. Experiments conducted demonstrated that the proposed scheme reduces collection costs, and improves the coverage ratio of data collections for future 6G networks.

In [20], Aparna Kumari et al. presented a review of IoT and blockchain technology's functionality for smart cities. A blockchain-based decentralized architecture for IoT smart cities has been proposed which covers different application perspectives like Intelligent Transportation Systems, smart grids, and underlying 6G communication networks, giving directions to efficiently integrate blockchain into IoT-envisioned smart cities.

In [21], Metehan Guzel et al. have reviewed AQI prediction alliteration from an algorithmic view and have introduced a new air quality framework that processes large quantities of data in real-time using Complex Event Processing. Scalability and extendability are achieved using fog computing, and the manageability is enhanced using a software-defined network.

3 Methodology

Let's take a brief look at the description of the data. After the screening and analysis of the data, it is split into two parts; one is used for training and the other for testing. We'll be using different machine learning algorithms to maximize the accuracy of AQI prediction. The machine learning algorithms used in this project include SVM, Linear regression, Random Forest, Decision tree, XGBoost, and RNN.

3.1 Dataset

Descriptions of list of variables used is shown in Table 1.

PM2 has been used to represent the concentration of particulate matter with a size less than 2.5 microns and PM10 for matter with a size less than 10 microns. Similarly, SO_2, NO_2, CO, and O_3 have been used to represent sulfur dioxide, nitrogen dioxide, carbon monoxide and ozone respectively.

The data that we have worked with in this project is Smart City Testbed's subset which IISc Bangalore and Pune Smart City Development Corporation Limited collected in 2019 while using smart city testbed to solve simple to complex use cases. Provided the parameters and data for a particular region in Pune, this data can be used to predict the Air quality index of that particular area, for example, airports, IT hubs, Residential areas, Railway stations etc. The analysis of the Air quality index based on the data attributes present in this dataset like the percentage of different pollutants in the air, light, sound, etc., can significantly help improve the city's living conditions.

3.2 Data Collection and Pre-processing

The data set used in the study has been taken from Kaggle [22]. However, in this study, we have done a lot of data cleaning and preprocessing to remove the outliers and null values, etc. We have considered the averages of the maximum and

Table 1 Data variables and description

Variable	Description
PM2	Average of particulates <2.5 microns maximum and minimum
PM10	Average of particulates <10 microns max
SO_2	Average of Sulphur dioxide maximum and minimum
NO_2	Average of nitrogen dioxide maximum and minimum
CO	Average of carbon monoxide maximum and minimum
O_3	Average of maximum and minimum

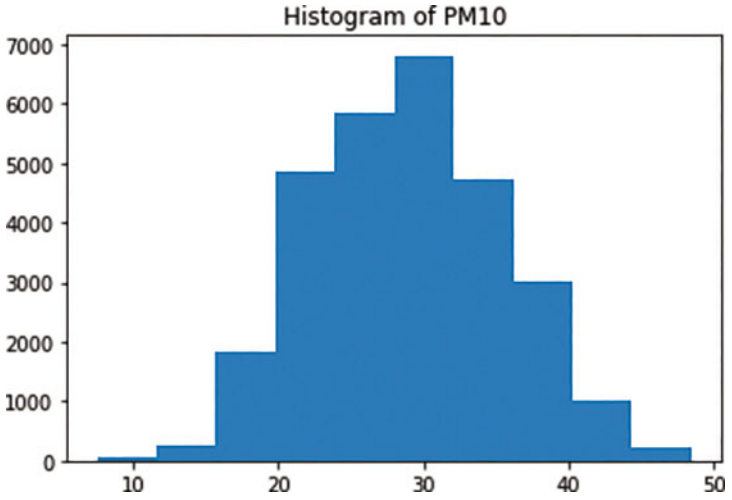

Fig. 2 PM10 histogram

minimum values of air pollutants like Ozone (O_3), Particulates <2.5 microns (PM2), Particulates <10 microns (PM10), Sulphur dioxide (SO_2), and Nitrogen Monoxide (NO).

3.2.1 Independent Variable Analysis

Histograms have been used to represent the relationship between independent variables and their frequency.

As shown in Fig. 2, we have a histogram for PM10. On the y-axis, we have the frequency and on the x-axis, we have the pollutant.

Figure 3 is the histogram for PM2. On the y-axis, we have the frequency; on the x-axis we have PM2.

Figure 4 is a histogram for NO_2. On the y-axis we have the frequency; on the x-axis we have the pollutant.

In Fig. 5 we have a histogram for O_3. On the y-axis, we have O_3's frequency and on the x-axis, we have the pollutant.

Figure 6, we show a histogram for SO_2. On the y-axis, we have the frequency and on the x-axis we have SO_2.

Figure 7 is the histogram for CO. On the y-axis, we have CO's frequency and, on the x-axis, we have the pollutant.

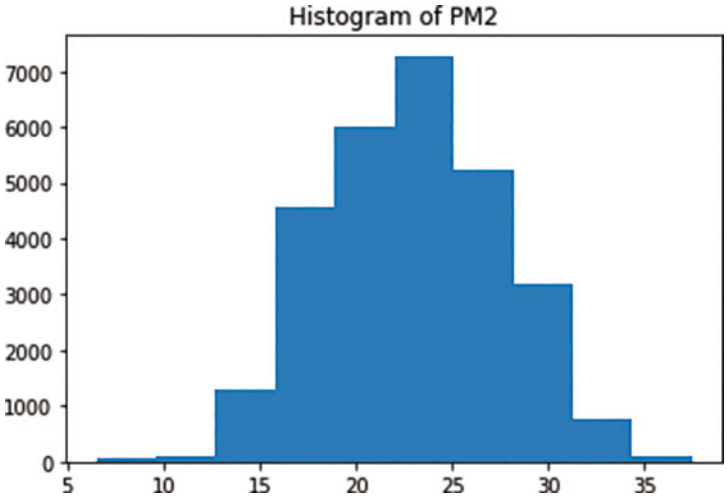

Fig. 3 PM2 histogram

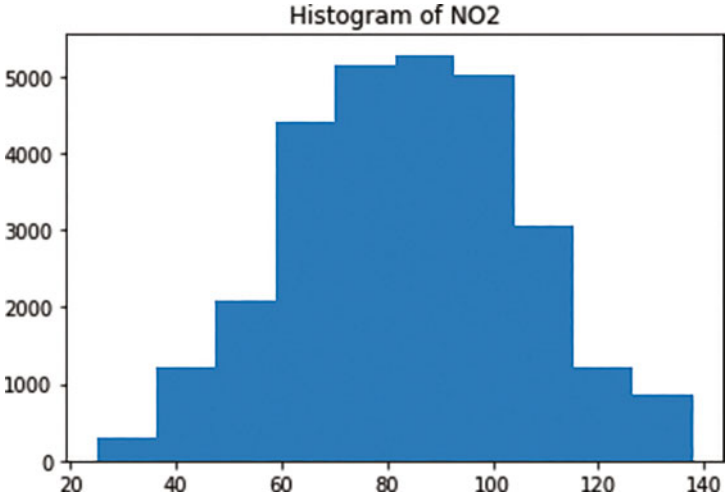

Fig. 4 NO_2 histogram

3.3 Data Analysis

Correlation matrix and distribution charts are used to determine the correlations among air pollution variables and the dataset's distribution and nature. For the analysis of data, Google collab notebooks have been used. In a dataset, correlations between all possible pairs of features are depicted using a Correlation matrix as shown in Table 2. The same has been used for the identification of features that

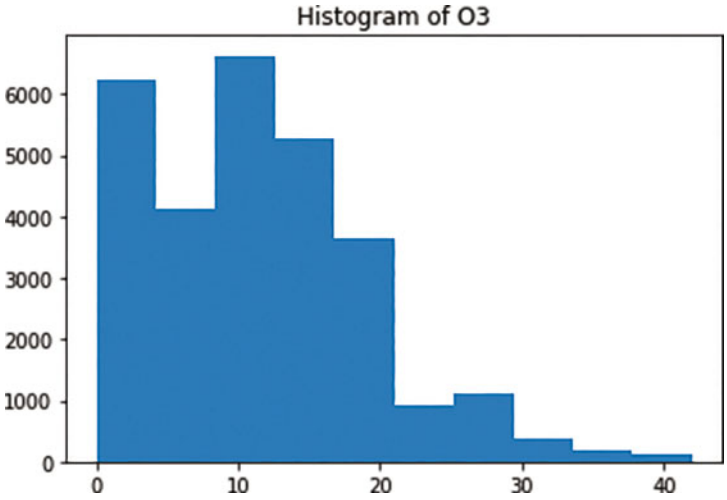

Fig. 5 O_3 histogram

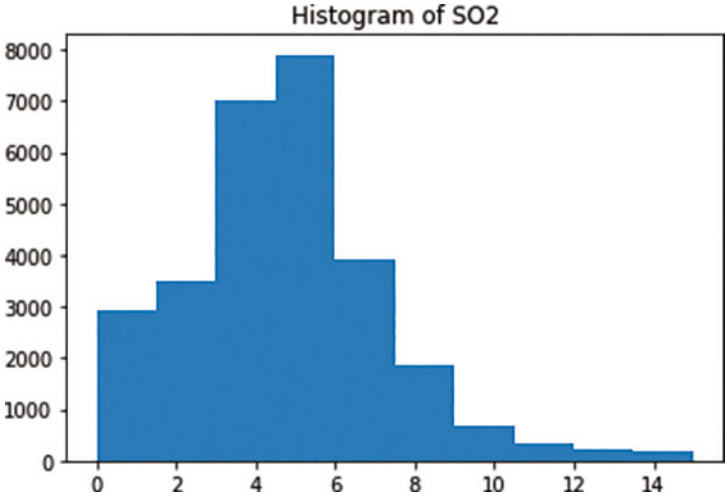

Fig. 6 SO_2 histogram

are most and least affected by PM2 to make the identification and visualization of patterns in the dataset easy and summarize it conveniently.

The above correlation matrix displays the correlation coefficient between pollutants such as PM10, PM2.5, SO_2, O_3, NO_2, and CO with each cell correlating the pollutants corresponding to the respective row and column. A 2D correlation matrix between two dimensions can be pictorially represented using a correlation heatmap, representing data with coloured cells from a monochromatic scale. Rows of the table are formed using values of the first dimension and the columns consist of values

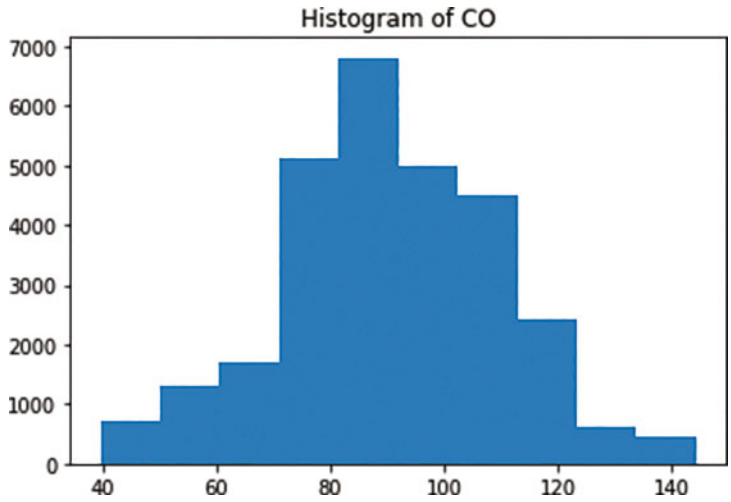

Fig. 7 CO histogram

Table 2 Correlation matrix

	Name	**PM10**	**NO_2**	**O_3**	**PM2**	**SO_2**	**CO**
Name	1	−0.19626	0.222522	0.09315	−0.19547	0.121699	−0.4072
PM10	−0.19626	1	0.199159	0.090565	0.965509	−0.02489	0.408033
NO_2	0.222522	0.199159	1	−0.19696	0.206393	0.195102	0.277064
O_3	0.09315	0.090565	−0.19696	1	0.044566	0.208589	−0.13103
PM2	−0.19547	0.965509	0.206393	0.044566	1	−0.04153	0.468013
SO_2	0.121699	−0.02489	0.195102	0.208589	−0.04153	1	0.000685
CO	−0.4072	0.408033	0.277064	−0.13103	0.468013	0.000685	v1

from the second dimension. The cell color is proportional to the measurements' number which matches the dimensional Value and is assisted by a color bar to make it understandable. It highlights the variation and differences in data making the patterns readable.

3.4 Machine Learning Algorithms

In this study, we are dealing with a regression problem where we are supposed to investigate the relationship between the independent feature variables and the dependent target variable to be able to make the prediction of the target attribute using the selected feature attributes. Training datasets are used to train the regression models with the values of the target variable and provided the feature attributes, the model learns to forecast the target variable [23, 24].

3.4.1 Support Vector Machine

It is one of the most commonly used machine learning algorithms and can be used for regression and classification's segregates n-dimensional space into classes using decision boundaries to simplify the categorization of new features. Hyperplanes are nothing but these best boundary lines. Hence datasets are divided into classes by SVM for finding a maximum marginal hyperplane. It chooses extreme data points called vectors for the creation of hyperplanes.

3.4.2 Random Forest Model

Random forest falls under supervised machine learning algorithms and is mainly used for Regression and Classification problems. For classification problems, it can tackle datasets with categorical variables and can deal with continuous variables too for problems related to regression. Decision trees are built by a Random forest algorithm taking their average for regression and majority vote in case of classification.

3.4.3 Linear Regression

It is a model used to depict the relationship between one dependent variable and one or multiple independent variables. In Simple Linear Regression, just a single dependent or explanatory variable is present. In this model, the summation of the distance between the predicted and actual Value of data is calculated and a line is chosen where this sum is minimum.

3.4.4 LSTM

It stands for Long Short-Term memory network. It is a type of recurrent neural network where the input of the current step is the output of the previous step; hence it can learn order dependencies in problems related to sequence prediction. Apart from single data points, like images, it can process the whole sequence of data as it has feedback connections.

3.4.5 Decision Tree

A decision tree is one of the most used methods for supervised learning. Decision trees split data sets on the basis of different conditions. It is used for both regression and classification tasks. Tree representation is used by the decision tree algorithm for problem-solving where leaf nodes represent class labels and internal nodes represent attributes Using decision trees, boolean functions can be represented on discrete attributes.

3.4.6 XGBoost

It is a machine learning algorithm based on gradient-boosted decision trees and has become very popular for structured or tabular data. It has been designed for speed and performance by using the dfs approach for tree pruning and parallelized tree building.

4 Evaluation Parameters and Implementation

For each of the pollutants, a sub-index is calculated based on their concentrations, health impacts, and their standards. The Value of the overall AQI is calculated and reflected by the worst sub-index. By using the help of medical experts, health impacts caused by these pollutants for various AQI categories have been suggested. For the pollutants that we have considered in the study, the AQI values are as follows:

Table 3 show the details about the category in which a pollutant lies for a certain concentration of its particles in the air. In this study, we have taken PM10 as the target variable on the basis of which we will be predicting the AQI for Pune City. Depending on the category in which the estimated Value of PM10 lies, the AQI for that specific area can be predicted.We have used machine learning models like SVM, Linear Regression, Random Forest, LSTM, Decision tree and XGBoost for the prediction of PM10's Value for the AQI calculation.

5 Results and Discussion

All the machine learning models used in the study, including the Support Vector Machine, Random forest model, Linear Regression, Decision Tree, LSTM and XGBoost, gave accuracies of over 90%.The Accuracies of the models are as follows. The exactness of the selected regression models is shown in Table 4**.**

Table 3 AQI values

AQI category (range)	PM2.5	PM10	NO_2	O_3	SO_2	CO
Good (0–50)	0–30	0–50	0–40	0–50	0–40	0–1.0
Satisfactory (51–100)	31–60	51–100	42–80	51–100	41–80	1.1–2.0
Moderate (101–200)	61–90	101–250	81–180	101–168	81–380	2.1–10
Poor (201–300)	91–120	251–350	181–280	169–208	381–800	10–17
Severe (301–400)	121–250	351–430	281–400	209–748	801–1600	17–34
Hazardous (401+)	250+	430+	400+	748+	1600+	34+

Table 4 Models' accuracy

Model	Accuracy
Support vector machine	0.92307
Random forest	0.99964
Linear regression	0.93657
LSTM	0.991627
Decision tree	0.99958
XGBoost	0.97000

The accuracy of various different models has been displayed in the table above. Both the Random forest and Decision tree model had accuracy above 99.9%, but the Random forest beats the Decision tree's accuracy by 0.0001% hence for the given dataset Random Forest model is the best model. Moving on to the model evaluation metrics, we'll be calculating the Mean Absolute error, Mean Square Error, Root Mean Squared Error, and R-Squared Score for the models that have been selected for this study.

5.1 Model Evaluation Metrics

5.1.1 Mean Absolute Error (MAE)

It is an evaluation metric that measures the mean of the absolute difference between the actual and predicted values. Basically, it calculates the average of residuals in the dataset.

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N} \mid yi - \hat{\text{y}}i \mid \tag{1}$$

Where N = Number of data samples,
ŷi = Predicted Value of y,
y = Actual Value of y.

5.1.2 Mean Square Error (MSE)

This evaluation metric calculates the mean of the squared difference between original and predicted values. It's used for the calculation of the variance of residuals in the dataset.

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N} \left(yi - \hat{\text{y}}i\right)^2 \tag{2}$$

Where N = Number of data samples,
ŷi = Predicted Value of y,
y = Actual Value of y.

5.1.3 Root Mean Squared Error (RMSE)

It is an evaluation metric that calculates the square root value of MSE. It calculates the standard deviation of residuals in the dataset.

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(yi - \hat{y}i\right)^2} \tag{3}$$

Where, N = Number of data samples,
ŷi = Predicted Value of y,
y = Actual Value of y.

5.1.4 R-Squared Score (R^2)

The R^2 score or the Coefficient of determination is an evaluation metric used for the evaluation of a regression model. It's used for the calculation of variance in the predicted values of the dataset. The Value of R squared will always be less than one irrespective of the values.

$$\text{R}^2 = 1 - \frac{SS\ Regression}{SS\ Total} = 1 - \frac{\sum_{i=1}^{N}\left(yi - \hat{y}i\right)^2}{\sum_{i=1}^{N}\left(yi - \bar{y}\right)^2} \tag{4}$$

Where, N = Number of data samples,
ŷi = Predicted Value of y,
y = Actual Value of y.

As per the evaluation metrics, we prefer lower values for MAES, MSE, and RMSE for relatively better performance. For the R^2 score, we prefer having larger values for better performance. Its Value usually lies between 0 and 1. A negative value for R^2 suggests that the chosen model doesn't follow the pattern and trend of data.

5.2 Model Analysis

The model having lower MAE, MSE, and RMSE values is considered to have a better performance as per the evaluation metrics and a model with higher R^2 scores is preferred over the ones with lower R^2 scores. The R^2 Value usually lies between 0 and 1. From Tables 5, 6, 7, and 8 we get the R^2, MSE, MAE, and RMSE scores respectively.

From Table 5, it can be observed that the Random forest regressor, Decision tree regressor, and LSTM regressor have the best performance with a score of 0.99965, 0.99961, and 0.99117 respectively. The Linear regressor and XGBoost also show a great performance with a score of .9365 and .9700 respectively. SVM's performance falls short in comparison to the other regressors but is still very good with a score of 0.87219.

Coming to the MSE scores, the Decision tree regressor and the Random Forest regressor have the lowest values of 0.0534 and 0.04503 respectively hence having the best performance. SVM has an MSE score of 16.523 and hence is the least suitable regressor as per the MSE metric. LSTM, XGBoost and Linear Regressor have MSE scores of 1.1407, 3.8779, and 8.1996 respectively.

Table 5 R^2 scores

Model	R^2 score
Linear regression	0.93657
Decision tree	0.99961
Random forest	0.99965
SVM	0.87219
XGBoost	0.97000
LSTM	0.99117

Table 6 MSE scores

Model	MSE score
Linear regression	8.19964
Decision tree	0.05341
Random forest	0.04503
SVM	16.5235
XGBoost	3.8779
LSTM	1.1407

Table 7 MAE scores

Model	MAE score
Linear regression	1.76394
Decision tree	0.05521
Random forest	0.07546
SVM	3.04468
XGBoost	1.16371
LSTM	0.66770

Table 8 RMSE scores

Model	RMSE score
Linear regression	2.86350
Decision tree	0.23111
Random forest	0.21220
SVM	4.06491
XGBoost	1.96925
LSTM	1.06804

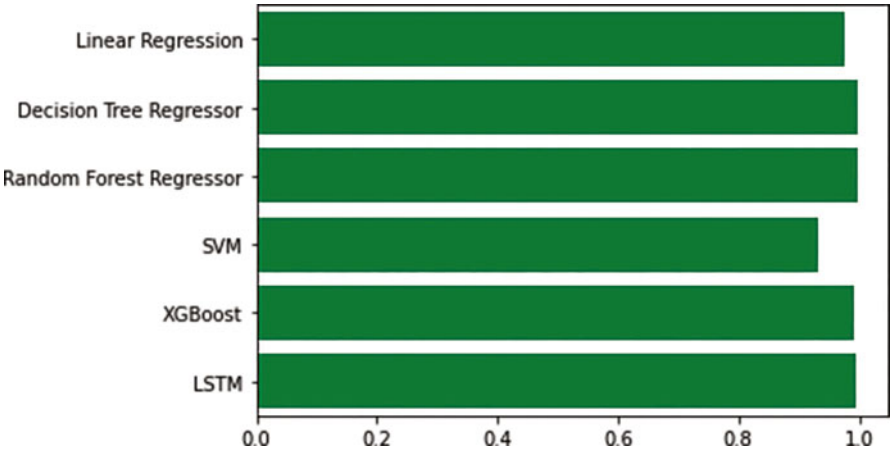

Fig. 8 R^2 scores

For the MAE values Decision Tree and Random forest Regressor once again emerge as the most suitable options with MAE scores of 0.05 and 0.075 respectively. They are followed by LSTM, XGBoost, and Linear regressor and have MAE scores of 0.667, 1.1637, and 1.7639 respectively. SVM has the highest MAE score with a value of 3.044.

Following the trends of MSE scores, the RMSE scores of the Decision Tree and Random forest regressor have the best values of 0.2311 and 0.21220 respectively, followed by LSTM and XGBoost have a value of 1.068 and 1.969 respectively. Linear regressor and SVM have the highest RMSE values at 2.863 and 4.064 respectively and hence are the least suitable regressors.

5.2.1 Bar Plot of R^2

Figure 8 displays the R^2 scores of the different models used in the study in the form of a bar graph. On the x-axis, we have the scores and on the y-axis, we have the different models.

Table 5 displays the R^2 scores of the different models used in the study in the form of table.

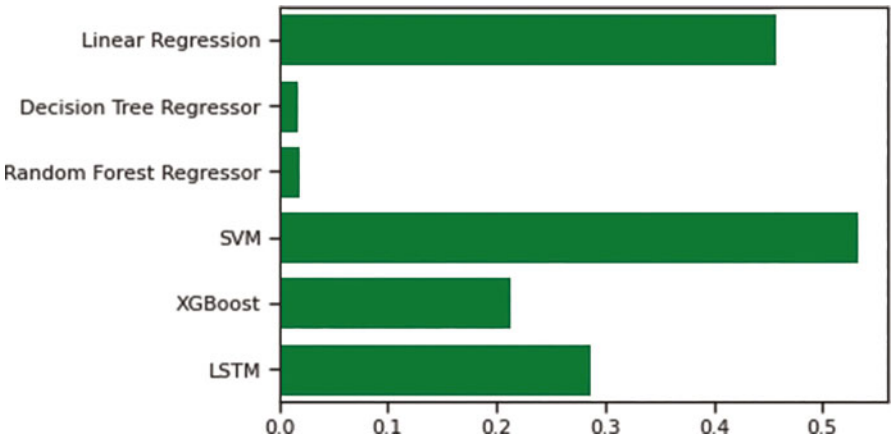

Fig. 9 MSE scores

5.2.2 Bar Plot of MSE

Figure 9 displays the MSE scores of the different models used in the study in the form of a bar graph. On the x-axis, we have the scores and on the y-axis, we have the different models used in the study.

MSE Scores
Table 6 displays the MSE scores of the different models used in the study in tabular format.

5.2.3 Bar Plot of MAE

Figure 10 displays the MAE scores of the different models used in the study in the form of a bar graph. We have the different models on the y-axis and the scores on the x-axis.

Table 7 displays the MAE scores of the different models used in the study in the form of a table.

5.2.4 Bar Plot of RMSE

Figure 11 displays the RMSE scores of the different models used in the study in the form of a bar graph. On the x-axis, we have the scores and on the y-axis, we have the different models which have been used in the study.

Table 8 displays the RMSE scores of the different models used in the study in tabular format.

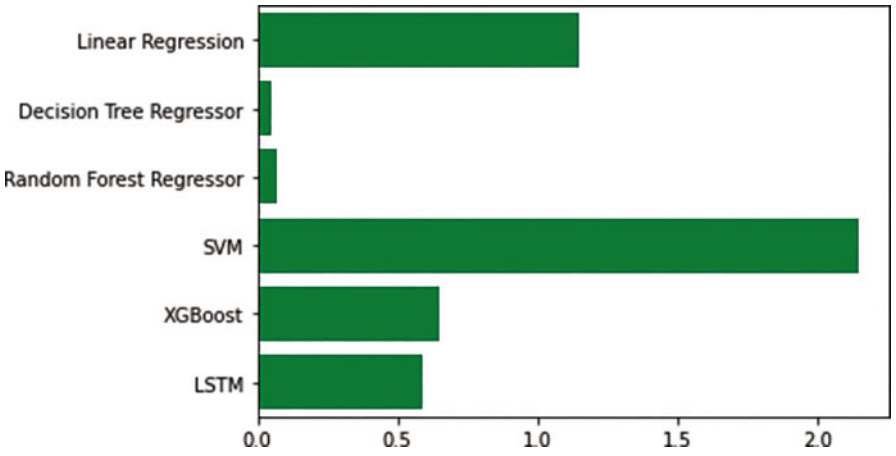

Fig. 10 MAE scores

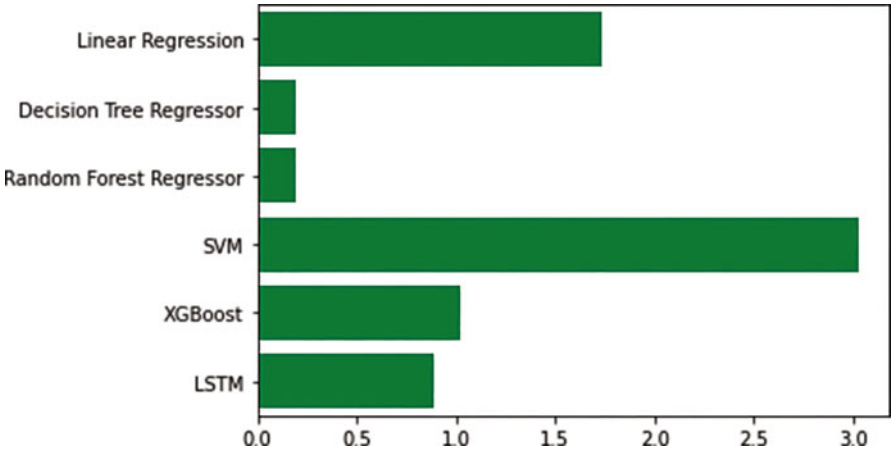

Fig. 11 RMSE scores

6 Conclusion and Future Scope

Air is a crucial element for human survival. Air Quality Index (AQI) value is a numerical representation of the current air quality. A correlation analysis was performed in this investigation to identify the contaminants influencing the air quality index. Pune's concentration of PM2.5 is anticipated using a rigorous correlation analysis. The current study assessed the accuracy of various deep learning and machine learning classification models on the dataset provided to estimate Pune's Air Quality Index (AQI). The dataset was preprocessed and cross-validated to increase prediction accuracy, with 70% of the data used for

model training and 30% for model testing. Support Vector Machine, Random Forest, Linear Regression, Decision Tree, LSTM, and XGBoost are included in the study's model. SVM obtained 92.307% accuracy, Random Forest 99.96%, Linear Regression 93.96%, LSTM 99.16%, Decision Tree 99.95%, and XGBoost model 97.0%. After evaluating all models with the most accurate predictions, the random forest emerged as the top model. Nonetheless, the Decision Tree model was also almost 99% more accurate.

Upon computing evaluation measures such as the R^2 score, MAE, MSE, and RMSE, it was determined that the Random Forest regressor had the greatest R^2 Value and the lowest MAE, MSE, and RMSE values, followed by the Decision Tree Regressor. The LSTM and XGBoost regressors also performed well on the dataset. The linear regressor also performed admirably. However, the SVM regressor had the lowest R^2 score and the highest MAE, MSE, and RMSE values for the dataset, indicating that it was the least acceptable model among the six investigated in this study. Future research could study the development of a prediction model based on deep learning that can determine the AQI of a given city or district.

References

1. WHO air pollution report. Available at: https://www.who.int/health-topics/air-pollution. Accessed 20 Sept 2022.
2. Air quality data statistics. Available at: https://www.airnow.gov/. Accessed 1 Oct 2022.
3. Central pollution control board. Available at: https://cpcb.nic.in. Accessed 5 Oct 2022.
4. Sharma, M., Jain, S., Mittal, S., & Sheikh, T. H. (2021). Forecasting and prediction of air pollutants concentrate using machine learning techniques: The case of India. In *IOP conference series: Materials science and engineering* (Vol. 1022, No. 1, p. 012123). IOP Publishing.
5. Mannan, M., & Al-Ghamdi, S. G. (2021). Indoor air quality in buildings: A comprehensive review on the factors influencing air pollution in a residential and commercial structure. *International Journal of Environmental Research and Public Health, 18*(6), 3276.
6. Dutta, S., Ghosh, S., & Dinda, S. (2021). Urban air-quality assessment and inferring the association between different factors: A comparative study among Delhi, Kolkata and Chennai megacity of India. *Aerosol Science and Engineering, 5*(1), 93–111.
7. Amuthadevi, C., Vijayan, D. S., & Ramachandran, V. (2021). Development of air quality monitoring (AQM) models using different machine learning approaches. *Journal of Ambient Intelligence and Humanized Computing*, 1–13.
8. Al-Qaness, M. A., Fan, H., Ewees, A. A., Yousri, D., & Abd Elaziz, M. (2021). Improved ANFIS model for forecasting Wuhan City air quality and analysis COVID-19 lockdown impacts on air quality. *Environmental Research, 194*, 110607.
9. Kumar, R. S., Arulanandham, A., & Arumugam, S. (2021, October). Air quality index analysis of Bengaluru city air pollutants using expectation maximization clustering. In *2021 international conference on advancements in electrical, electronics, communication, computing and automation (ICAECA)* (pp. 1–4). IEEE.
10. Fernando, R. M., Ilmini, W. M. K. S., & Vidanagama, D. U. (2022). Prediction of air quality index in Colombo.
11. Iskandaryan, D., Ramos, F., & Trilles, S. (2021). Features exploration from datasets vision in the air quality prediction domain. *Atmosphere, 12*(3), 312.
12. Alemdar, K. D., Kaya, Ö., Canale, A., Çodur, M. Y., & Campisi, T. (2021). Evaluation of air quality index by spatial analysis depending on vehicle traffic during the COVID-19 outbreak in Turkey. *Energies, 14*(18), 5729.

13. Gladson, L. A., Cromar, K. R., Ghazipura, M., Knowland, K. E., Keller, C. A., & Duncan, B. (2022). Communicating respiratory health risk among children using a global air quality index. *Environment International, 159*, 107023.
14. Sun, X., Li, S., Chen, X., & Wang, K. (2021, March). Air quality index prediction based on improved PSO-BP. In *IOP conference series: Earth and environmental science* (Vol. 692, No. 3, p. 032069). IOP Publishing.
15. Phruksahiran, N. (2021). Improvement of air quality index prediction using geographically weighted predictor methodology. *Urban Climate, 38*, 100890.
16. Singh, M., Singh, B. B., Singh, R., Upendra, B., Kaur, R., Gill, S. S., & Biswas, M. S. (2021). Quantifying COVID-19-enforced global changes in atmospheric pollutants using cloud computing-based remote sensing. *Remote Sensing Applications: Society and Environment, 22*, 100489.
17. Peneti, S., et al. (2021). BDN-GWMNN: Internet of things (IoT) enabled secure smart city applications. *Wireless Personal Communications, 119*(3), 2469–2485.
18. Kochetkov, D., Vuković, D., Sadekov, N., & Levkiv, H. (2019). Smart cities and 5G networks: An emerging technological area? *Journal of the Geographical Institute "Jovan Cvijić" SASA, 69*(3), 289–295.
19. Li, T., et al. (2021). DRLR: A deep-reinforcement-learning-based recruitment scheme for massive data collections in 6G-based IoT networks. *IEEE Internet of Things Journal, 9*(16), 14595–14609.
20. Kumari, A., Gupta, R., & Tanwar, S. (2021). Amalgamation of blockchain and IoT for smart cities underlying 6G communication: A comprehensive review. *Computer Communications, 172*, 102–118.
21. Guzel, M., & Ozdemir, S. (2019). A new CEP-based air quality prediction framework for fog based IoT. *2019 International Symposium on Networks, Computers and Communications (ISNCC)*. IEEE.
22. Pune smart city dataset. Available at: https://www.kaggle.com/datasets/akshman/pune-smartcity-test-dataset. Accessed 1 Oct 2022.
23. Soni, K. M., Gupta, A., & Jain, T. (2021). Supervised machine learning approaches for breast cancer classification and a high-performance recurrent neural network. In *2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA)*, pp. 1–7. https://doi.org/10.1109/ICIRCA51532.2021.9544630aset easy and summariz
24. Jain, T., Verma, V. K., Agarwal, M., Yadav, A. & Jain, A. (2020). Supervised machine learning approach for the prediction of breast cancer. In *2020 International Conference on System, Computation, Automation and Networking (ICSCAN)*, pp. 1–6. https://doi.org/10.1109/ICSCAN49426.2020.9262403