International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 4, 2025
https://theaspd.com/index.php

Predictive Analytics For Air Quality Classification: A Multi-City Study In India

Mr.Nikhil Vilasrao Deshmukh[1], Dr.S.Barani[2], Dr.S.Poornapushpakala[3], Dr.M.Subramoniam[4]
[1]School of Electrical and Electronics, Research Scholar, Sathyabama Institute of Science and Technology Chennai, India, n.deshmukh83@gmail.com
[2]School of Electrical and Electronics, Research Scholar, Sathyabama Institute of Science and Technology Chennai, India, baraniselvaraj77@gmail.com
[3]School of Electrical and Electronics, Research Scholar, Sathyabama Institute of Science and Technology Chennai, India, poornapushpakalas@gmail.com
[4]School of Electrical and Electronics, Research Scholar, Sathyabama Institute of Science and Technology Chennai, India,subramoniam.viru@gmail.com

Abstract
Pollution is one of the major causes for human health diseases. In highly populated countries like India, to meet out the job requirements of the people major cities are expanded with tremendous industrial and population growth. Predictive analysis of air quality is essential as it would caution people about the air they inhale and proper remediations to reduce the pollution could be taken. In India, including four metropolitan cities Chennai, Kolkata, Mumbai, Delhi, Bangalore is also more urbanized. This urbanization effects leads to lot of industrial hubs in those regions causing more pollution especially water and air. In recent days the air pollution is tremendously increases and lot of breathing disorders are being reported. Hence the proposed work attempts to study the air pollution trends for these cities for a decade from 2013 to 2023. A prediction and classification model is developed to predict the quality of air and classify the level of pollution. The techniques implemented are Long Short-Term Memory (LSTM), XGBoost and SARIMAX regression for prediction and XGBoost classifier for classification of various levels. XGBoost regression and classifier model outperforms with 0.9519 R^2 and 0.9583 classification accuracy.
Keywords: *AQI, $PM_{2.5}$, PM_{10}, SARIMAX, XGBoost, LSTM, Regression, Classifier.*

INTRODUCTION

Urbanization of cities have paved way for tremendous growth of industries all over the globe. This leads to increase in various pollution such as water, air and noise etc which creates more impact on environment. High pollution necessitates an effective pollution monitoring [1]. Environmental conditions are becoming worst further as an impact of urbanization. Studies has been carried out for analyzing the trend of $PM_{2.5}$ particle and urbanization [2]. This article also discussed about the socioeconomic factors, clean energy etc. Urbanization is a process of integrating socioeconomic elements, such as energy use, building development, transportation, and industrial output, all of which are closely linked to $PM_{2.5}$ emission sources [3][4]. The air quality of many countries exceeds greater than 10 times the level of $PM_{2.5}$ recommended by World Health Organization WHO which is less than 5 μg/m³. India ranks in the 6th position bearing $PM_{2.5}$ value as 50.6 μg/m³in the year 2024. Though the pollution in India has decreased from 54.4 μg/m³ to 50.6 μg/m³, air pollution is a big threat in the country. In rural areas like Punjab and Haryana the residues of agricultural produce are burnt which contributes in significant air pollution in winter. The capital city Delhi is consistently suffering from air pollution in recent years. The major contribution to air pollution in India is due to the pollution in major cities where industrial growth is tremendous.

International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 4, 2025
https://theaspd.com/index.php

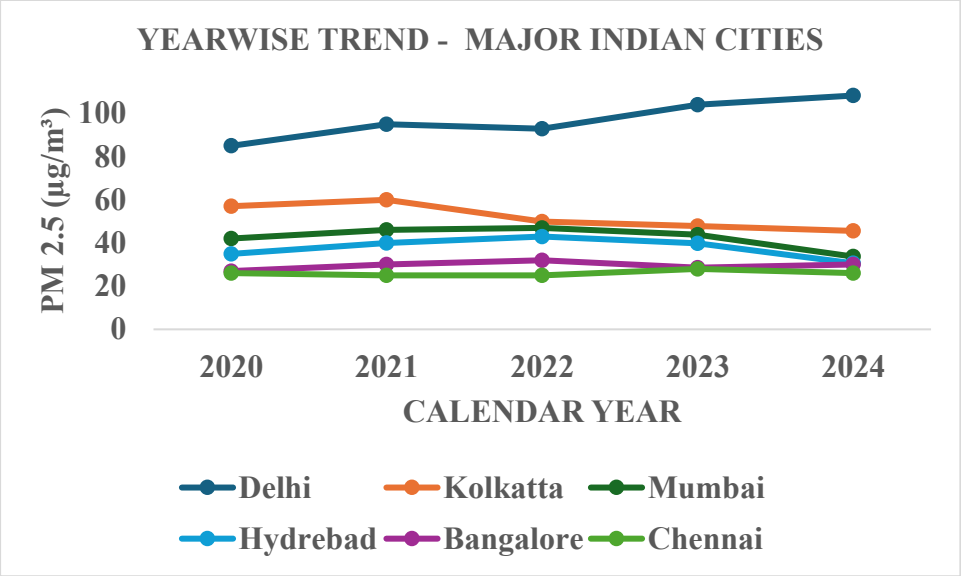


Figure 1 Air Pollution in Major Indian Cities

The pollution in India could be due to varied reasons such as vehicle pollution, industrial exhaust, adulteration of fuels and heavy traffic which increases the level of $PM_{2.5}$ heavily [5]. According to the data book released by IQAIR [6] in the year 2024 the air pollution due to $PM_{2.5}$ has been recorded for the past five years and is given in Figure 1. Air pollution in Delhi is potentially high which is greater than 20% of the WHO recommended level. This necessitates the study on air pollution analysis over a decade of years in major cities like Delhi, Kolkata, Chennai, Bangalore and Mumbai in India. The quality of air during festival seasons varies tremendously in India particularly during Diwali, since lot of crackers have been burnt by children and adults which is customary throughout the country. Form Figure 1 it is evident that in all major cities in India the finest particle $PM_{2.5}$ is gradually increasing from the year 2020. In cities like Delhi, in spite of taking actions to reduce air pollution it gradually increases. Analysis of air pollution and recommendations based on this is effective only when the measurements are recorded frequently for all days and all seasons throughout the year and also multiple such readings are required for the same geographical region. Redundancy of information confirms the pollution level at the measured monitoring station and city.

Research is being carried out for prediction of Air Quality Index. In Taiwan, the datasets for a decade from 2008 to 2018 were consolidated and trained for forecasting. Random Forest Regression, KNN, SVM Regression, Stacked Ensemble techniques are implemented for forecasting the AQI for next 8 hours. the features considered for training are $PM_{2.5}$, PM_{10}, AQI and CO [7]. In India, for a coastal city Vishakhapatnam in Kerala, AI based prediction of AQI is implemented. The evaluation metrics estimated to evaluate the efficiency of algorithm are Root Mean Square Error (RMSE), Mean Square Error (MSE) and R^2. The input parameters are $PM_{2.5}$, PM_{10}, CO, NO, NO_2, NH_3, NO_X, SO_2, O_3 and AQI. The implemented methodology are Random Forest, XGBoost, AdaBoost and Light BGM etc [8,12].

Also, analysis for made for major cities in India. But the dataset collected is very less for training. the maximum accuracy achieved are 88.98 %, 91.49 %, 94.48 %, 92.66 %, 95.22 % for New Delhi, Bangalore, Kolkata, Hyderabad and Chennai respectively. The algorithm implemented is decision tree. The data collected is enhanced by SMOTE technique. The results are analyzed for both imbalanced and balanced dataset [9]. Similar research is carried out using Gradient Boost Technique and Root Mean Square Error (RMSE) and R^2 were used as evaluation indicators [10]. Research is also focused for web based real time prediction using machine learning algorithms. The data form various sites that are open-source data is acquired directly from sites and trained categorization. Totally 23,463 data is being used in training with the input parameters as $PM_{2.5}$, O, NO_2, O_3. The AQI is classified in to Good, Moderate, Unhealthy for Sensitive Group, Very Unhealthy, Hazardous [11].

Form 2 sites Marenplatz and AmNeckartor in Germany the data is collected for pollutants $PM_{2.5}$, PM_{10}, NO_2. The data for a period of four years from 2018 – 2022 is taken for training. The techniques adopted

are ridge regression, SVM, Random Forest Regression, Gradient Boosting and extra tree regression [13]. Research is carried out to predict NO_2 and SO_2. The entire model is implemented in three different ways. Model 1 uses only features related to weather for training whereas model 2 uses quantities of fossil fuel consumed by plants which is an indirect measure of emission rate for training. Model 3 integrates both model 1 and 2 parameters as input for training [14]. In Kolkata, readings from 7 monitoring stations Rabindra Bharath University, Ballygunge, Rabindra Sarovar, Jadavpur, Victoria, Fort William and Bindhanagar were collected and trained using KNN and ridge regression techniques for prediction [15]. Similar kind of AI based prediction of air quality is implemented using machine learning algorithms using regression, classifier and LSTM network. The proposed model is developed for classification of level of air quality in major cities Chennai, Bangalore, Kolkata, Delhi and Mumbai.

METHODS AND MATERIALS

To estimate the Air Quality Index (AQI) of major metropolitan cities in India, three supervised learning models such as Random Forest, Artificial Neural Network (ANN), and XGBoost were used in this study. Each model uses distinct mechanisms and mathematical foundations for classification. The description of those is given below.

Data Exploration

The data collected is from Central Pollution Control Board of India which is an open-source data for a decade from the Calander year 2013 to 2023. The data is taken for major cities of India like Chennai, Mumbai, Kolkata, Delhi and Bangalore and is represented in Figure 2. The input features are Year, City, $PM_{2.5}$, PM_{10}, NO_2 and SO_2 and AQI [18].

Data distribution analysis is done to study the data. Violin plot is an integrated approach of kernel density and box plot. Figure 3 indicates the violin plot of $PM_{2.5}$ which is the finest particle of air pollution. It is the range of Particulate Matter 2.5 that decides the quality of air since these finest particles having less than 2.5 micrometer diameter protrude in to lungs and blood vessels easily causing hazardous health issues. These Particulate Matter is combination of metals, carbon, sulphate and nitrates

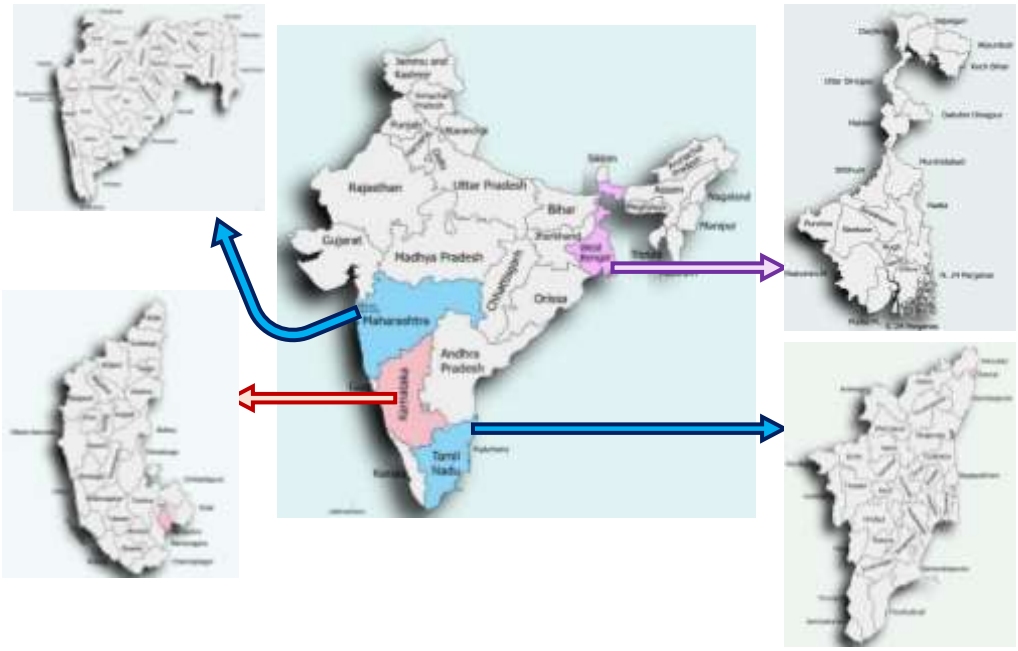


Figure 2 Geographical Map of locations for data collection

International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 4, 2025
https://theaspd.com/index.php

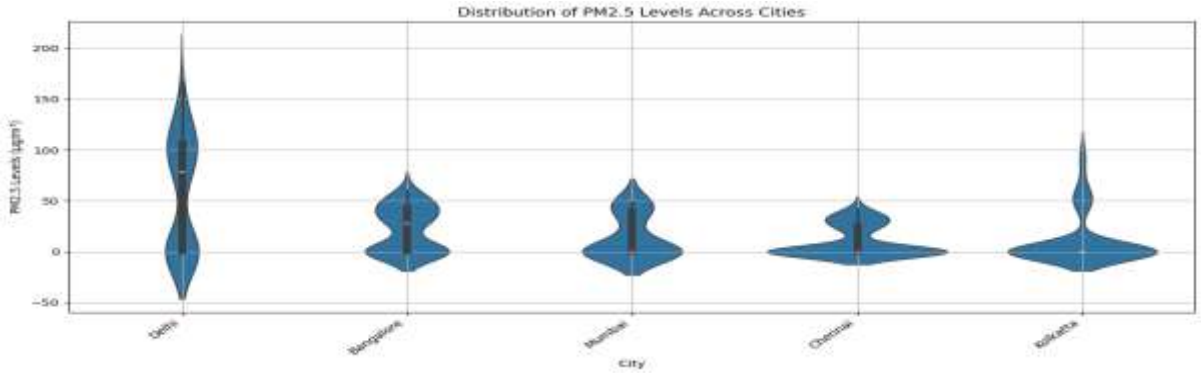


Figure 3 Distribution of PM 2.5 for Major Cities in India

Form Figure 3 it is inferred that the distribution of $PM_{2.5}$ in Delhi is wide and Peak in shape indicating the value of $PM_{2.5}$ in Delhi is increasing. The level of $PM_{2.5}$ in Delhi reaches the value of 200 µg/m^3 which is an indication of severe pollution. Chennai and Kolkata have similar shapes but the peak in Chennai is 50 whereas Kolkata is greater than 100. Bangalore and Mumbai is greater than 50 but less than 100. When $PM_{2.5}$ is less than 35 µg/m^3 the region is moderate and is slightly unhealthy for sensitive humans who already suffers with respiratory issues. All the major cities in India have Particulate Matter greater than 35 µg/m^3. Chennai, Bangalore and Mumbai is less polluted when compared to other cities. Though the peak value of these cities is less, the plot is wider indicating possibility of variations in value. In Delhi the pollution level is very and consistent for years. Kolkata also causes severe air pollution next to Delhi. Hence it is a threat to these cities. From Figure 4 it is inferred that the $PM_{2.5}$ increases in Delhi over the years and is considerably reduced from 2019 - 2020 due to lockdowns. Chennai and Bangalore are maintaining consistent values from 2021 onwards and has moderate value of $PM_{2.5}$. Whereas Kolkata and Mumbai has increasing trend.

Figure 5 shows the distribution of air pollutants SO_2, NO_2, PM_{10} and $PM_{2.5}$ over a period of years in all major cities in India. It is evident from the figure that $PM_{2.5}$ increases gradually increases after the year 2020 and PM_{10} is high over a period of years. This increasing trend in air pollution in all major cities of India necessitates to develop a model The techniques implemented for training is LSTM, XGBoost regression and SARIMAX regression model. to predict the quality of air and its level of pollution. The proposed work develops a model to predict the quality of air through machine learning techniques. With the trained model as input the air quality is categorized in to five different levels good, satisfactory, moderate, poor and very poor. Hence a stacked machine learning model is used for prediction.

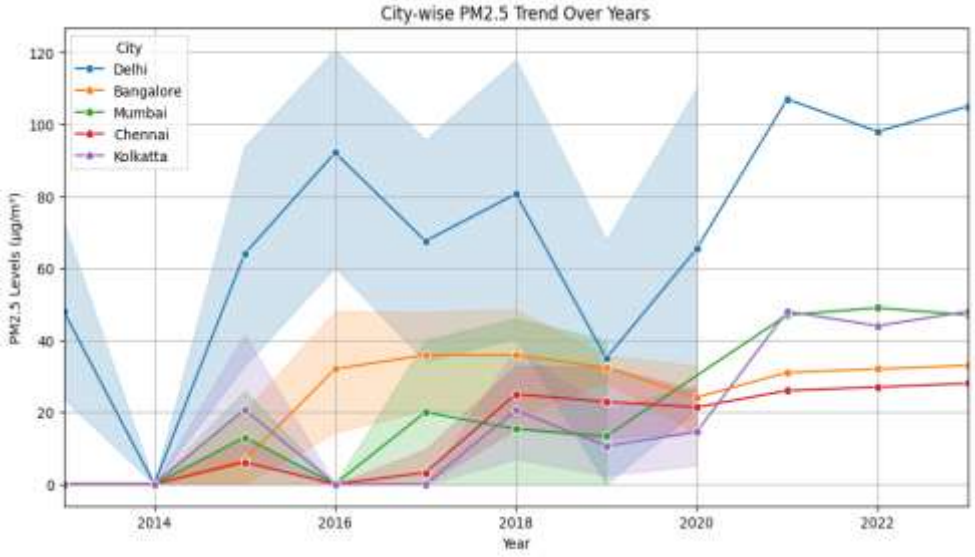


Figure 4 City-wise $PM_{2.5}$ Trend over Years

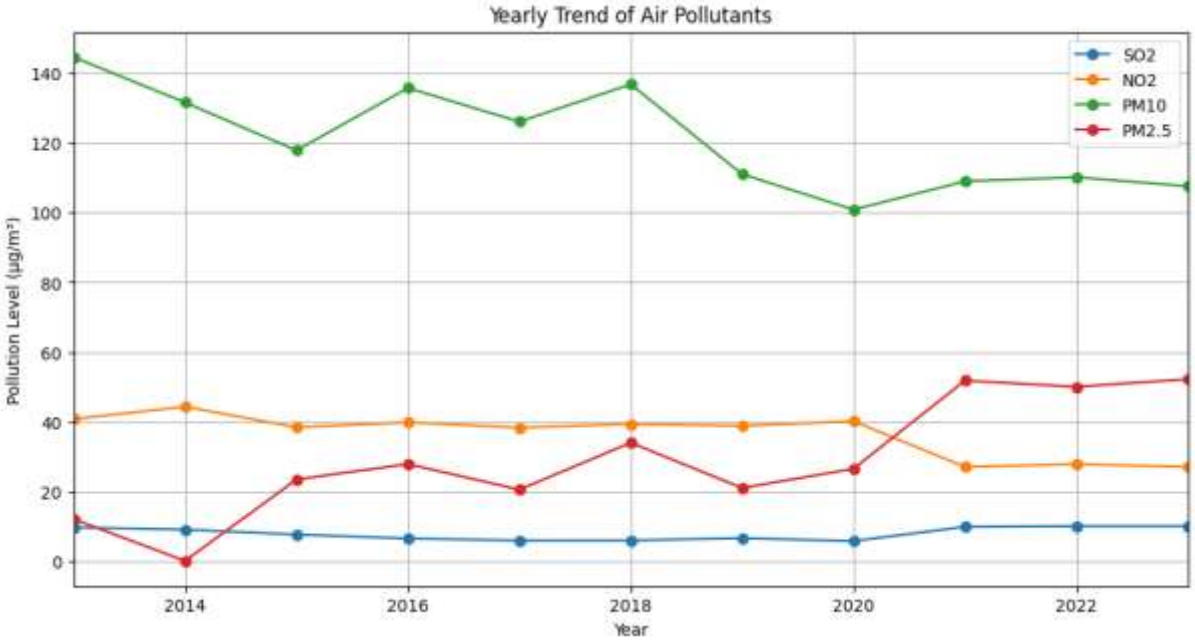


Figure 5 Year wise Air Pollution Level

Proposed Model

The model proposed for prediction is stacked machine learning model. Initially regression model and LSTM is adopted for training the network with year, city, SO_2, NO_2, PM_{10}, $PM_{2.5}$, $PM_{2.5}$_AQI, PM_{10}_AQI and AQI as features. After training with regression model the Air Quality Index is categorized in to five various classes and tested for classification using XGBoost classifier. Figure 6 depicts the functional block diagram of the proposed work. The various category of Air Quality Index and its class for training and testing is given in Table 1 and the algorithm of the proposed model is given in Table 2.

Table 1 AQI Scale for Air Quality [19]

Class	AQI Values	Nature of Air Quality
0	0-50	Good
1	51-100	Satisfactory
2	101 -200	Moderate
3	201-300	Poor
4	300 and above	Very Poor

Table 2 Algorithm for the proposed model

Algorithm - Proposed Model
BEGIN
 # Step 1: Data Collection
1.1 Collect data containing air pollutants, year and cities.
1.2 Verify the collected data for cleaning
 # Step 2: Development of Regression Model
2.1 Develop a regression model with the air pollutants, year and location as Features and train the model.
 # Step 3: Validate the prediction model
3.1 Compare and Validate the developed regression model for error parameters

and its performance and identify the best fit model.

 # Step 4: Apply Classifier on Regression output

4.1 Apply classifier for the regression output

4.2 Classify the level of air quality for the predicted output.

 # Step 5: Evaluate the performance

5. 1Evaluate the performance of Regression and Classifiers and find best fit model for the application.

 # Step 6: End Process

6.1 Stop Execution

END

Long Short-Term Memory

Long Short-Term Memory (LSTM) is a special type of recurrent neural network (RNN). It is designed to learn long-term dependencies in sequential data. The traditional RNN models generally suffer from vanishing and exploding gradient problems. This problem is eliminated in LSTM. This is done by introducing memory cells. These memory cells will retain information over extended time steps. Each of the LSTM cell comprises of three gates such as input gate, forget gate, and output gate. These gates will regulate the flow of information.

The Mathematical expression for LSTM network is as follows.

Let input vector xtx_txt, previous hidden state ht−1h_{t-1}ht−1, and previous cell state Ct−1C_{t-1}Ct−1, for a given time step t, the gates are computed as follows:

Forget Gate: $f_t = \sigma(W_f.[h_{t-1}, x_t] + b_f)$ (1)

Input Gate: $i_t = \sigma(W_f.[h_{t-1}, x_t] + b_i)$ (2)

$\hat{C}_t = \tan h\,(W_c.[h_{t-1}, x_t] + b_c)$ (3)

Cell state update: $C_t = f_t * C_{t-1} + i_i * \hat{C}_t)$ (4)

Output Gate: $o_t = \sigma(W_o.[h_{t-1}, x_t] + b_o)$ (5)

$h_t = o_t * \tan h\,(C_t)$ (6)

Where,

σ- denotes the sigmoid activation function, tanh is the hyperbolic tangent function, w and b are learnable weights and biases respectively.

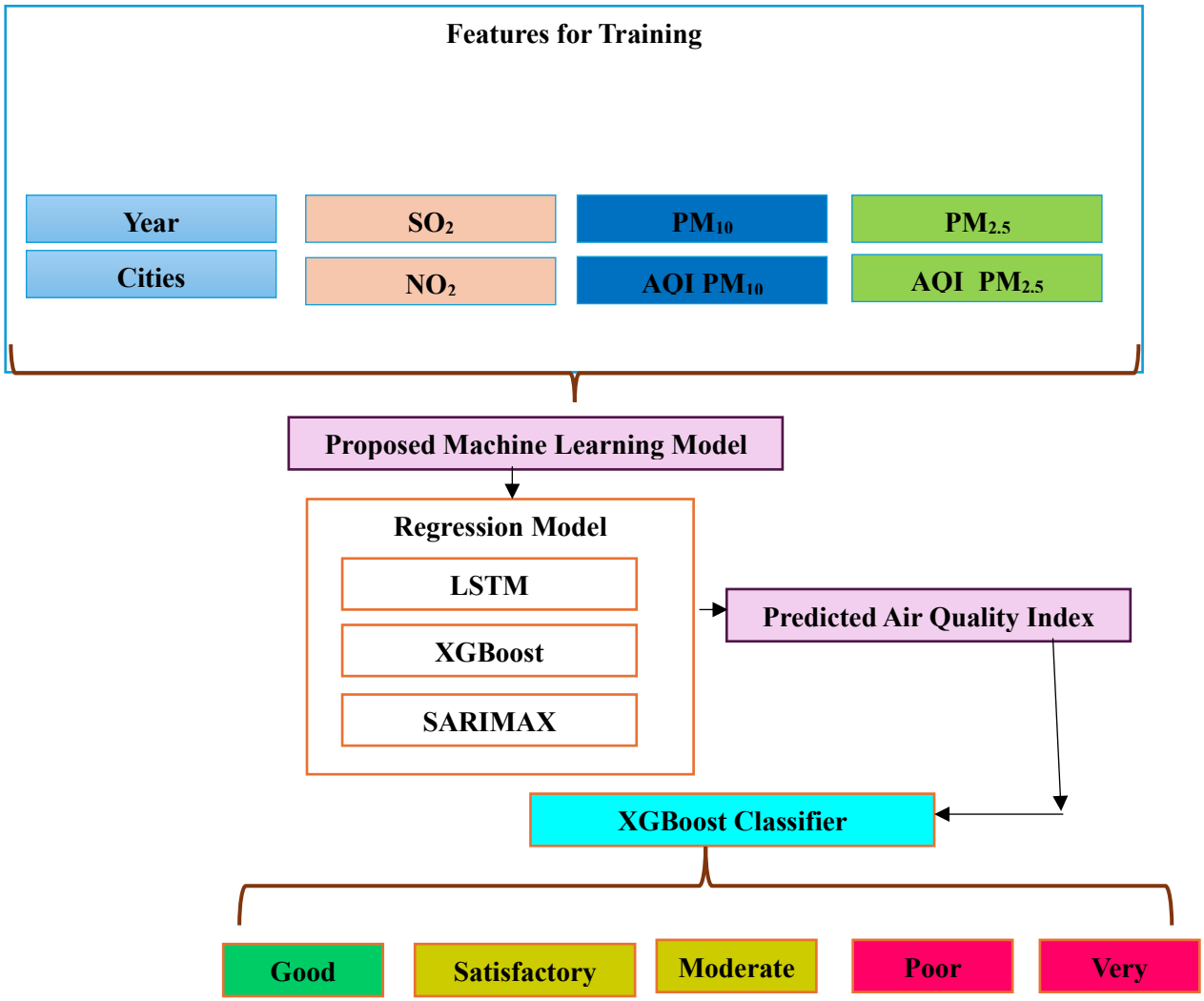
LSTM models take the sequence of feature vectors as input, passes through a fully connected dense layer with a softmax or sigmoid activation function and produces the output class label. The parameters such as number of LSTM units, number of layers, sequence length, batch size, dropout rate, and learning rate are the main parameters influencing LSTM classification. The model performance can be optimized by tuning these parameters.

Figure 6 Functional Block Diagram of Proposed Model

XGBoost Regression

International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 4, 2025
https://theaspd.com/index.php

XGBoost Regression (Extreme Gradient Boosting) is an advanced ensemble learning algorithm. It is based on gradient-boosted decision trees, in which the model is optimized for speed and performance. Due to its regularization mechanisms and ability to handle missing values, XG Boost is much effective for structured and tabular data in regression tasks. It builds an additive model in forward step manner by minimizing the regularized objective function and combining weak learners iteratively to improve



the prediction accuracy. The properties such as early stopping, parallelization, and handling of missing values internally, makes XGBoost robust and scalable for regression applications. it robust and scalable for regression applications.

At the t-th iteration, the prediction is updated as:

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(x_i), f_t \in F \quad (7)$$

Where F is the space of regression trees, $\hat{y}_i^{(t)}$ is the predicted value of sample I, and f_t is the newly added tree. The objective function to be minimized is

$$L^{(-t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t)}\right) + \sum_{k=1}^{t} \Omega(f_k) \quad (8)$$

Where l is the differentiable convex loss function (eg. Squared Error) and

$\Omega(f) = \gamma T + \frac{1}{2}\lambda \|\omega\|^2$ is the regularization term that penalizes model complexity, with T denoting the number of trees in the leaves in the tree and ω the leaf weights.

To optimize this objective, XGBoost employs a second order Taylor approximation:

$$L^{(-t)} \approx L^{(-t)} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)\right] + \Omega(f_t) \quad (9)$$

Where $g_i = \partial_{\hat{y}_{(t-1)}} l(y_i, \hat{y}_i{}^{(t-1)}$ and $h_i = \partial^2{}_{\hat{y}_{(t-1)}} l(y_i, \hat{y}_i{}^{(t-1)}$ are the first and second derivatives of the loss function.

SARIMAX Regression

SARIMAX Regression is an extension of the ARIMA model. It incorporates both seasonal effects and external variables, making it a powerful tool for time series forecasting and regression tasks. This model is particularly useful when the target variable is affected by both its own past values and additional explanatory time-dependent variables.

Mathematically a SARIMAX is represented as

$$y_t = C + \emptyset_1 y_{t-1} + \cdots + \emptyset_p y_{t-p} + \theta_1 \in_{t-1} + \cdots + \theta_q \in_{t-q} + \beta^T x_t + \in_t \quad (10)$$

Where y_t – dependent variable at time t.

$\emptyset_1$ – autoregressive coefficients.

θ_j – moving average coefficients.

$\in_t$ – white noise

x_t – exogenous variables at time t and β regression coefficients for exogenous variables.

The exogenous part in the equation allows the model to learn from additional predictors. This is done by minimizing the error between actual and predicted values using a maximum likelihood estimation (MLE) approach. The performance of the model can be improved by tuning the key parameters such as AR/MA orders, seasonal orders, differencing terms, and lag structures of the exogenous regressors. For time series forecasting SARIMAX provides a statistically grounded framework to model temporal autocorrelation, trend, seasonality, and exogenous influence in a unified regression model.

1.1 XGBoost Classifier

XG boost or extreme gradient boosting is an optimized gradient boosting algorithm. It builds the decision trees sequentially in which the errors in a particular tree will be corrected or minimized in the next tree. This is done by fitting the residuals or errors of the previous trees into the new tree. It also uses L1 and L2 regularization method to prevent overfitting. The mathematical description of this model is given by Given a prediction

$$\hat{y}^{(t)} = \sum_{i=1}^{n} l\left(y_i, y_i{}^{-(t-1)} + f_t(x_i)\right) + \Omega(f_t) \quad (11)$$

$$L^t = \sum_{i=1}^{n} l\left(y_i, \hat{y_i}{}^{-(t-1)} + f_t(x_i)\right) + \Omega(f_t) \quad (12)$$

XGBoost efficiently finds optimal splits by using second-order Taylor expansion of the loss and greedy tree growing. The prediction done by this model will be fast and much accurate. The confusion matrix produced by these models and its discussion with the air quality dataset is described in section

RESULTS AND DISCUSSION

The data collected from pollution control board comprises of year, location and pollutants. Since the collected information is year wise, it is a time series data. LSTM network is best suited for time series data hence the initial training is started with LSTM technique. The features of training are year, city, SO_2, NO_2, PM_{10}, $PM_{2.5}$, $PM_{2.5}$_AQI, PM_{10}_AQI and the target is AQI. The LSTM model is initially built with 64 neurons where randomly 20% of the neurons are dropped during training. Again, a layer with 32 neurons is built with same 20% dropouts with fully connected one output neuron. The r^2 obtained is -0.225 where the errors are greater than 50. Hence the network is tuned for enhancing the performance. the training and validation loss is given in Figure 7. The predicted and actual value of AQI is given in Figure 8.

International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 4, 2025
https://theaspd.com/index.php

Figure 7 Training and Validation Loss of LSTM before Tuning

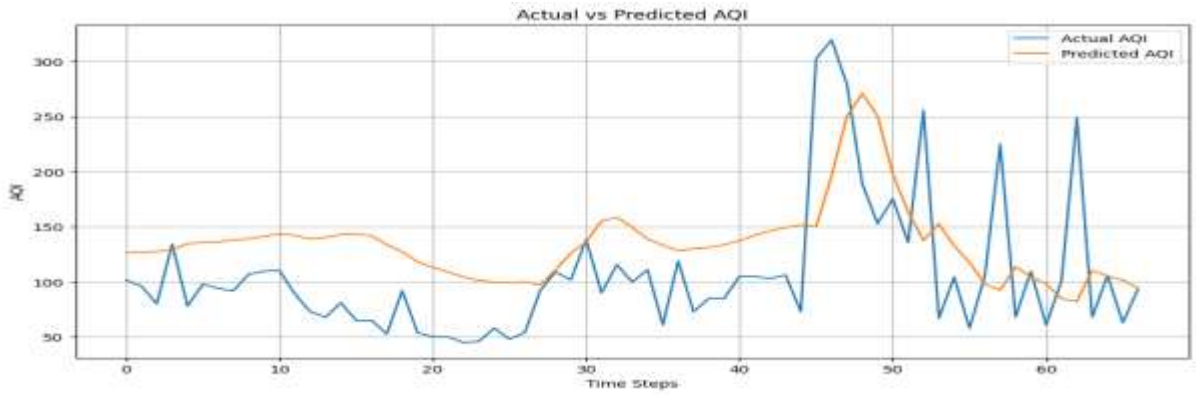


Figure 8 Actual Vs Predicted plot of AQI before Tuning

Figure 8 clearly depicts the deviation between predicted and actual AQI is widened. Form Figure 7 it is inferred that the training loss decreases indicating the network is learning with training. Whereas for validation till 20 epochs the loss decreases after which the loss starts increasing. There by to eradicate this and enhance the performance tuning of learning rate is essential and early stopping of training would resolve this. Hence the network is trained with early stopping and optimized with Adam. Figure 9,10 gives Training and Validation Loss of LSTM after Tuning and prediction plot. From the figure it is evident that the training and validation loss decreases with increasing epochs and the predicted and actual AQI values are close there by reducing the errors. But still the errors are high and the r2 value is only 0.2694. Table 3 gives the error metrics and r2 values of training before and after tuning.

Table 3 Measuring metrics of LSTM technique before and after tuning

Metrics	Before Tuning	After Tuning
Mean Absolute Error (MAE)	50.13	34.94
Mean Square Error (MSE)	3678.78	2628.52
Root Mean Square Error (RMSE)	60.65	51.27
Mean Absolute Percentage Error (MAPE)	56.67%	32.18%
R^2	-0.0225	0.2694

International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 4, 2025
https://theaspd.com/index.php

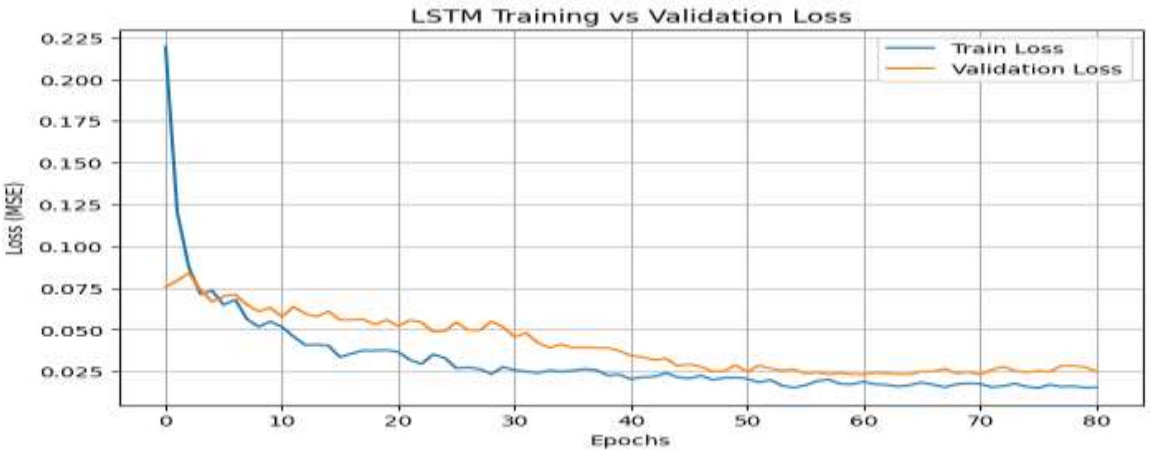


Figure 9Training and Validation Loss of LSTM after Tuning

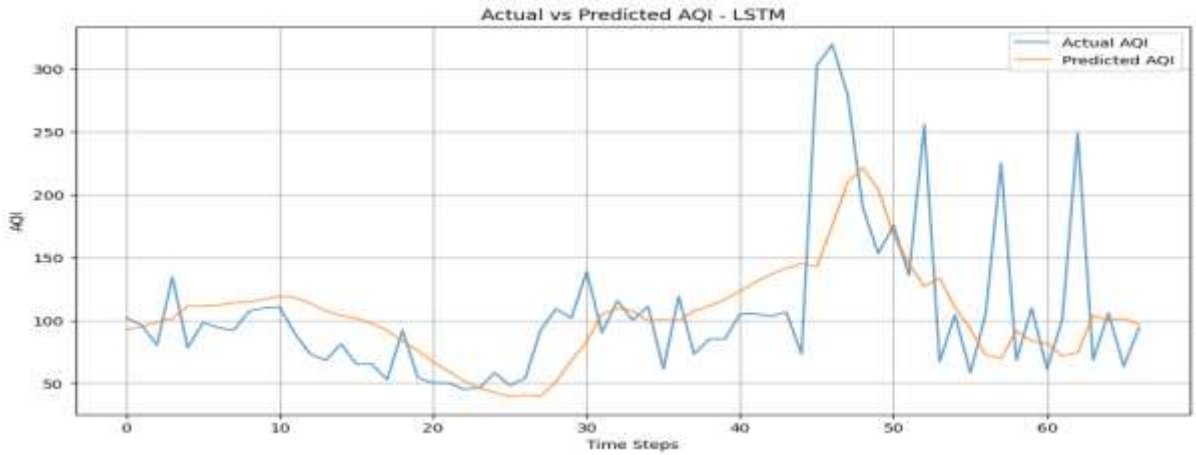


Figure 10 Actual Vs Predicted plot of AQI after Tuning

Hence from the training and validation it is identified that this time series data contains only year which is not sufficient for training in LSTM. Hence the regression XGBoost and SARIMAX regression has been adopted with the same features and target.

Figure 11 shows the SARIMAX actual and predicted values from which it is evident that the predicted and actual values are close with reduced deviation.

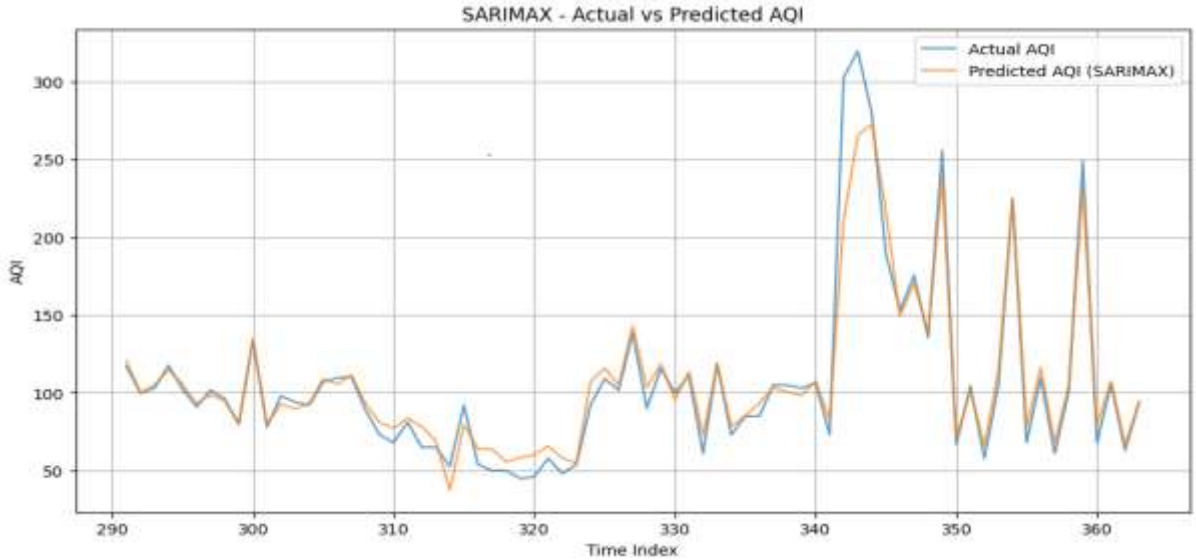
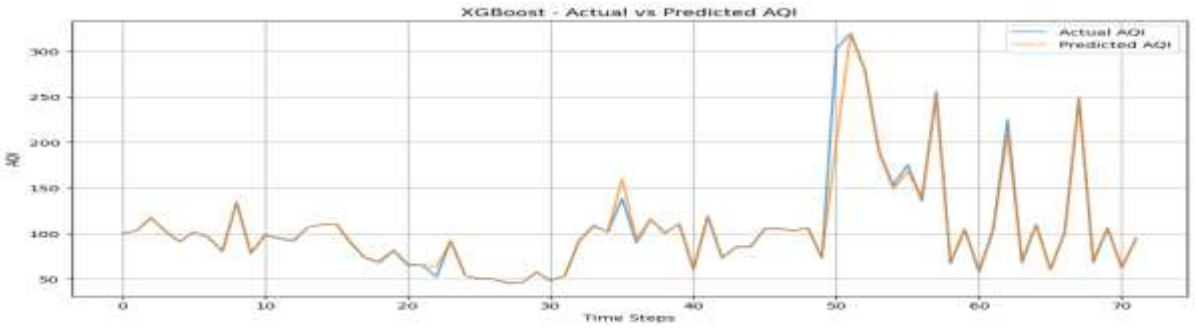


Figure 11 Actual Vs Predicted – SARIMAX

Figure 12 shows the prediction plot for XGBoost regression where the deviation is still reduced and values are closer.

Figure 12 Actual Vs Predicted - XGBoost



The measuring metrics for SARIMAX and Regression are given in Table 4. From table 4 it is evident that XGBoost outperforms with r^2 of 0.9519.

Table 4 Measuring Metrics of XGBoost and SARIMAX

Metrics	SARIMAX	XGBoost
Mean Absolute Error (MAE)	7.95	3.26
Mean Square Error (MSE)	217.08	161.33
Root Mean Square Error (RMSE)	14.73	12.70
R^2	0.9344	0.9519

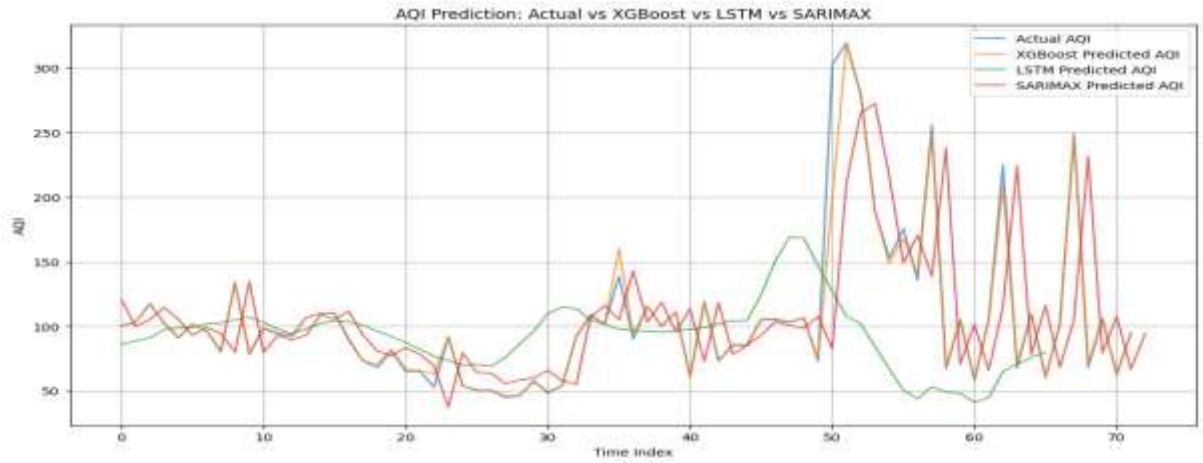


Figure 13 shows the comparison of performance of three techniques implemented for training.

Figure 13 Comparison of Actual Vs Predicted between LSTM, SARIMAX and XGBoost

In second phase of work with the regressed output the model is tested for classification. The confusion matrix for XGBoost and SARIMAX is shown in Figure 14. The classification accuracy of for XGBoost is 0.9583. The classification accuracy of SARIMAX model is 0.7945.

International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 4, 2025
https://theaspd.com/index.php

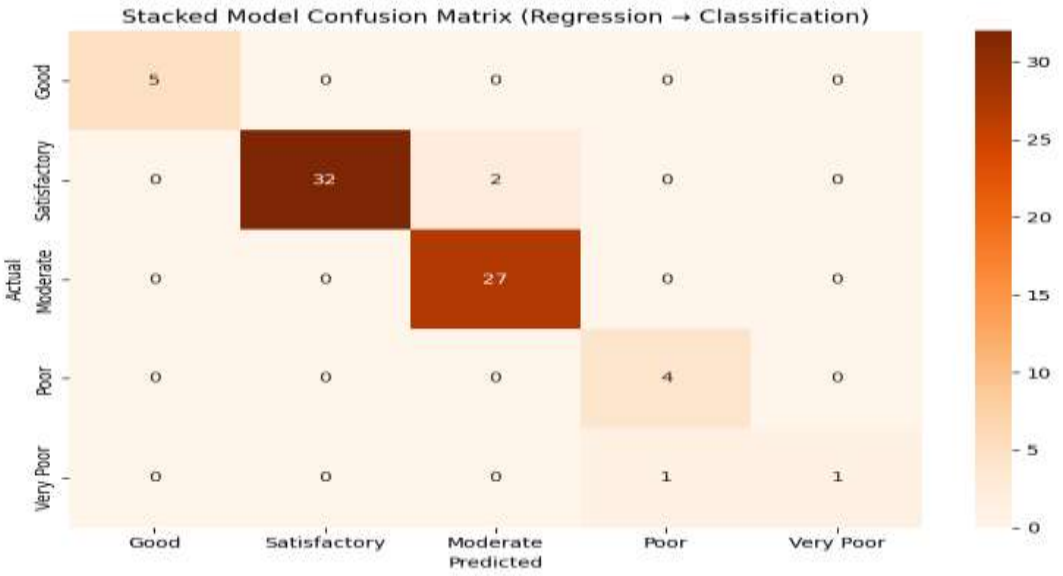


Figure 14 -a Confusion Matrix – XGBoost

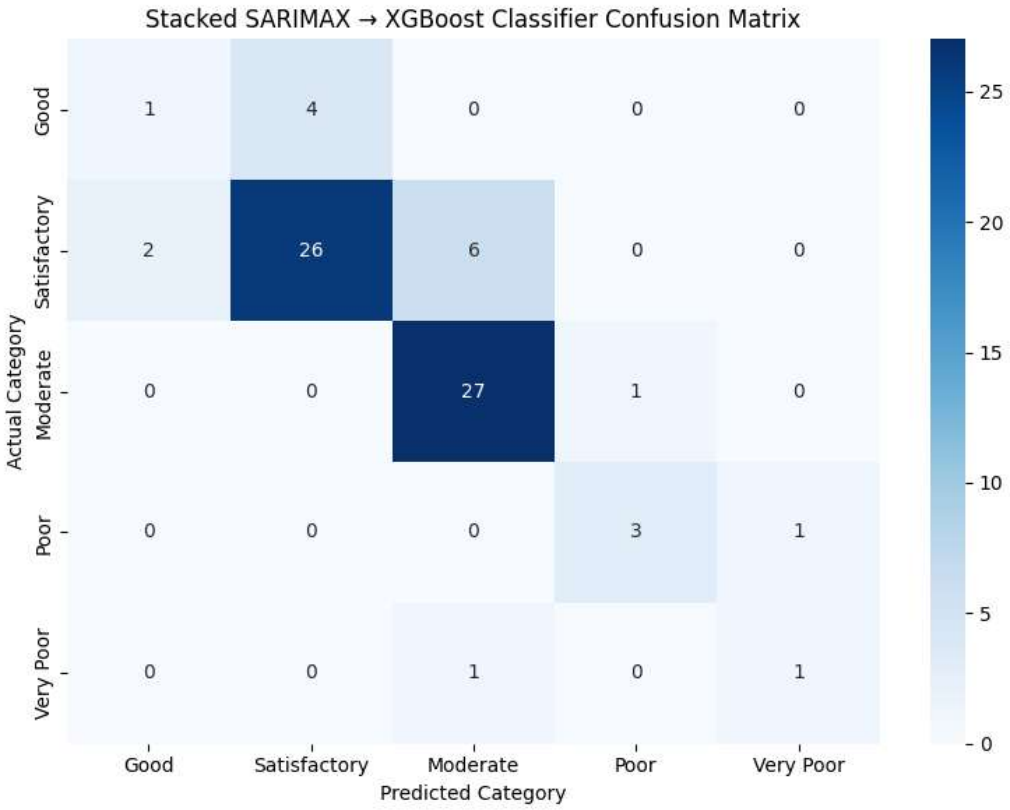


Figure 14 - b Confusion Matrix – SARIMAX

CONCLUSION

The proposed work implements a machine learning model to predict and classify the quality of air for major cities Chennai, Bangalore, Kolkata, Mumbai and Delhi in India. The implemented technique is stacked technique in which initially the model is trained with the dataset containing pollutants, city, year and AQI as features. As a first phase the training is done with LSTM and other regression models. The XGBoost regression performs better producing R^2 as 0.9519. The LSTM produces more error as the time series in dataset is not sufficient to learn. Over the regression model classification model is stacked to classify various levels of air quality. The stacked performance of XGBoost regression and Classification

performs better with classification accuracy of 0.9583. SARIMAX regression though trained and validated successfully the performance of stacked operation is poor with 0.7945 accuracy in classification. Hence the proposed work with XGBoost techniques is best fit for prediction of air quality.

REFERENCES

1. Biao SUN, Chuanglin Fang, Xia Liao, Xiaomin Guo, Zhitao Liu,The relationship between urbanization and air pollution affected by intercity factor mobility: A case of the Yangtze River Delta region, Environmental Impact Assessment Review, Volume 100, 2023, 107092, ISSN 0195-9255, https://doi.org/10.1016/j.eiar.2023.107092.
2. Xiaomin Wang, Guanghui Tian, Dongyang Yang, Wenxin Zhang, Debin Lu, Zhongmei Liu,Responses of PM2.5 pollution to urbanization in China, Energy Policy,Volume 123,2018,Pages 602-610,ISSN 0301-4215,https://doi.org/10.1016/j.enpol.2018.09.001.
3. Fangyuan Wang, Xiao Han, Huan Xie, Yi Gao, Xu Guan, Meigen Zhang,Investigating trends and causes of simultaneous high pollution from PM2.5 and ozone in China, 2015–2023,Atmospheric Pollution Research, Volume 16, Issue 1,2025,102351,ISSN 1309-1042,https://doi.org/10.1016/j.apr.2024.102351.
4. Yichen Wang, ChenGuang Liu, Qiyuan Wang, Quande Qin, Honghao Ren, Junji Cao,Impacts of natural and socioeconomic factors on PM2.5 from 2014 to 2017,Journal of Environmental Management,Volume 284,2021,112071,ISSN 0301-4797,https://doi.org/10.1016/j.jenvman.2021.112071.
5. Gurjar BR. Air pollution in India: Major issues and challenges. The Energy and Resources Institute. Published April 5, 2021. https://www.teriin.org/article/air-pollution-india-majorissues-and-challenges
6. https://www.iqair.com/in-en/world-most-polluted-countries
7. Liang, Y.-C.; Maimury, Y.; Chen, A.H.-L.; Juarez, J.R.C. Machine Learning-Based Prediction of Air Quality. *Appl. Sci.* **2020**, *10*, 9151. https://doi.org/10.3390/app10249151
8. Gokulan Ravindiran, Gasim Hayder, Karthick Kanagarathinam, Avinash Alagumalai, Christian Sonne,Air quality prediction by machine learning models: A predictive study on the indian coastal city of Visakhapatnam,Chemosphere,Volume 338,2023,139518,ISSN 0045-6535,https://doi.org/10.1016/j.chemosphere.2023.139518.
9. Natarajan, S.K., Shanmurthy, P., Arockiam, D. *et al.* Optimized machine learning model for air quality index prediction in major cities in India. *Sci Rep* **14**, 6795 (2024). https://doi.org/10.1038/s41598-024-54807-1
10. C. Li, Y. Li and Y. Bao, "Research on Air Quality Prediction Based on Machine Learning," *2021 2nd International Conference on Intelligent Computing and Human-Computer Interaction (ICHCI)*, Shenyang, China, 2021, pp. 77-81, doi: 10.1109/ICHCI54629.2021.00022.
11. Rahman, M.M., Nayeem, .E.H., Ahmed, .S. *et al.* AirNet: predictive machine learning model for air quality forecasting using web interface. *Environ Syst Res* **13**, 44 (2024). https://doi.org/10.1186/s40068-024-00378-z
12. Aram, Simon & Nketiah, Edward & Saalidong, Benjamin & Wang, · & Afitiri, Abdul-Rahaman & Akoto, Akwasi & Osei Lartey, Patrick. (2023). Machine learning-based prediction of air quality index and air quality grade: a comparative analysis. International journal of Environmental Science and Technology. 10.1007/s13762-023-05016-2.
13. A. Samad, S. Garuda, U. Vogt, B. Yang,Air pollution prediction using machine learning techniques – An approach to replace existing monitoring stations with virtual monitoring stations,Atmospheric Environment,Volume 310,2023,119987,ISSN 1352-2310,https://doi.org/10.1016/j.atmosenv.2023.119987.
14. Mohammad A. Alolayan, Abdullah Almutairi, Suad M. Aladwani, Shiekha Alkhamees,Investigating major sources of air pollution and improving spatiotemporal forecast accuracy using supervised machine learning and a proxy,Journal of Engineering Research,Volume 11, Issue 3,2023,Pages 87-93,ISSN 2307-1877,https://doi.org/10.1016/j.jer.2023.100126.
15. Sabyasachi Mondal, Abisa Sinha Adhikary, Ambar Dutta, Ramakant Bhardwaj, Sharadia Dey,Utilizing Machine Learning for air pollution prediction, comprehensive impact assessment, and effective solutions in Kolkata, India,Results in Earth Sciences,Volume 2,2024,100030,ISSN 2211-7148,https://doi.org/10.1016/j.rines.2024.100030.
16. Hai Tao, Ali H. Jawad, A.H. Shather, Zainab Al-Khafaji, Tarik A. Rashid, Mumtaz Ali, Nadhir Al-Ansari, Haydar Abdulameer Marhoon, hamsuddin Shahid, Zaher Mundher Yaseen,Machine learning algorithms for high-resolution prediction of spatiotemporal distribution of air pollution from meteorological and soil parameters,Environment International,Volume 175,2023,107931,ISSN 0160-4120,https://doi.org/10.1016/j.envint.2023.107931.
17. Yunzhe Li, Zhipeng Sha, Aohan Tang, Keith Goulding, Xuejun Liu,The application of machine learning to air pollution research: A bibliometric analysis,Ecotoxicology and Environmental Safety,Volume 257,2023,114911,ISSN 0147-6513,https://doi.org/10.1016/j.ecoenv.2023.114911.
18. https://cpcb.nic.in/
19. Suresh Kumar Natarajan, Prakash Shanmurthy, Daniel Arockiam, Balamurugan Balusamy & Shitharth Selvarajan, Optimized machine learning model for air quality index prediction in major cities in India, Scientific Reports | (2024) 14:6795.