International Journal of Applied Science
ISSN (Online): 2208-2182
Volume 11 Issue 04 December 2025
DOI: https://doi.org/10.69980/zjd1mp66

MACHINE LEARNING–BASED PREDICTION OF AIR QUALITY LEVELS ACROSS INDIAN CITIES USING ENVIRONMENTAL MONITORING DATA

Dr. Arjun Mehta[1*], Prof. Priya S. Nair[2], Dr. Rohan Kulkarni[3], Dr. Neha Kapoor[4],
[1]Department of Computer Science and Engineering, Indian Institute of Technology Delhi, New Delhi India Email: arjun.mehta@iitd.ac.in*
[2]Department of Environmental Engineering, Indian Institute of Technology Madras, Chennai, Tamil Nadu, India Email: priyanair@iitm.ac.in
[3]School of Artificial Intelligence and Data Science, Savitribai Phule Pune University, Pune, Maharashtra, India Email: rohan.kulkarni@unipune.ac.in
[4]Centre for Atmospheric Sciences, Indian Institute of Technology Kanpur, Kanpur, Uttar Pradesh, India, Email: neha.kapoor@iitk.ac.in

Abstract—*Air pollution is a key environmental and public health problem in rapidly urbanizing Indian cities. The present study aimed to design and assess the performance of machine learning models for the prediction of the next day's air-quality categories, based on the environmental monitoring data available at large-scale. The data set includes 235,785 observations from 291 cities in 32 states and union territories in India for the period of April 2022 to April 2025. Following the calculation of lagged AQI values and moving-average indicators, 219,888 AQI model ready observations were retained. The predictors included temporal and geographical variables, pollutants, and monitoring variables, and the train-test split was a chronological split to avoid information leakage. Models evaluated were multinomial logistic regression, decision tree, random forest and histogram gradient boosting models with respect to performance metrics including accuracy, balanced accuracy, precision, recall, macro F1 score, weighted F1 score and confusion matrices. The highest overall accuracy was obtained by histogram gradient boosting, while the highest macro F1 score was obtained by the random forest algorithm and the performance was most balanced across the six AQI categories. Random forest model gave the best prediction for Good, Satisfactory and Moderate conditions, but reduced prediction accuracy for Very Poor and Severe conditions due to significant class imbalance. The majority of errors in the AQI range were between neighbouring categories. The results prove that machine learning is very useful for the air-quality forecasting in multiple cities, early-warning systems and evidence-based environmental management in India.*

Keywords: *Air Quality Index; machine learning; air pollution forecasting; Indian cities; environmental monitoring*

I. INTRODUCTION

In rapidly urbanizing areas, air pollution is now a serious environmental and public health issue, especially when the increase in traffic, industrial activity, building, energy use and climatic changes of the season are combined. However, in recent years, the monitoring and forecasting of atmospheric pollution is significantly improved by innovations in the field of Internet of Things infrastructure, big data analytics and machine learning (Gangwar et al., 2023). In smart city, sensor-based systems are becoming a popular solution as they allow continuous measurement and automate the prediction of changing AQ conditions (Iskandaryan et al., 2020). The potential of using pollutant sensing coupled with predictive analytics for pollution management during operations has also been shown through the use of IoT based environmental platforms (Hussain et al., 2020). Machine learning is thus a valuable asset to complement real-time monitoring systems and enhance early warnings and public-health decision-making (Karnati, 2023).

The prediction of the Air Quality Index has been widely used the machine-learning techniques that can model the nonlinear relationships between pollutants, time and geographical characteristics. Studies have illustrated that the supervised algorithms can reliably classify the AQI using the environmental monitoring variables in comparative studies (Gupta et al., 2023). Deep learning models have also shown to be successful for the forecasting of AQI with complex temporal dependence extracted from historical observations in the metropolitan level (Janarthanan et al., 2021). Incorporating complementary feature extraction and sequence learning mechanisms have achieved further accuracy gains in prediction using hybrid deep-learning architectures (Sarkar et al., 2022). The spatiotemporal deep-learning-based regional forecasting shows that the model incorporating both location and time can improve the estimation of air quality in interconnected regions (Abirami & Chitra, 2021). Pollution levels in India vary significantly from city to city, season to season and region to region making it a particularly interesting setting for air-quality research that employs a predictive approach.

The city-level pollution data proved to be very useful in predicting the pollution pattern in multiple Indian cities with the help of machine-learning models, which proved to be very useful in predicting the pollution pattern in multiple Indian cities (Kumar & Pande, 2023). For key cities in India, optimized models have been validated to show that historical AQI patterns and context predictors can be used to accurately classify and predict (Natarajan et al., 2024). Other AI-related solutions have been suggested as potential practical solutions for pollution management and enhanced regulation in India (Rautela & Goyal, 2024). Forecasting of monthly AQI has additionally confirmed that for new urban environments, cross validation and model comparison is vital to determine robust forecasting methods (Pande et al., 2025).

There are several city-specific studies giving additional evidence. In Hyderabad, spatiotemporal modelling was shown to be effective using artificial-intelligence techniques to predict PM2.5 concentrations (Gokul et al., 2023). A study in Hyderabad found that machine-learning algorithms can be used to understand pollutant variability in a complex urban environment (Mathew et al., 2023). In Ahmedabad, multiple machine-learning models were tested, indicating good prediction performance of the AQI (Maltare & Vahora, 2023). Predictive modelling was able to represent the variation of air-quality in coastal urban environment successfully in Visakhapatnam (Ravindiran et al., 2023a). A machine learning-based model has been found to accurately estimate PM2.5 levels in the atmosphere, as demonstrated in studies conducted in Delhi (Kumar et al., 2020).

Additionally, particle-matter prediction models based on data and deep-learning techniques have been found to be improved in severe pollution conditions (Masood & Ahmad, 2023). Other meteorological and urban information can be useful for environmental forecasting. Meteorological data has helped to model air quality in urban areas by accounting for dispersion and seasonality (Mampitiya et al., 2023). Pollution pattern has been mapped with the land-surface temperature and urban-form indicators in Bengaluru (Suthar et al., 2024). The machine learning (ML)-based forecasting of pollution has emerged as a promising tool for intelligent environmental management (Kothandaraman et al., 2022). Sustainable monitoring and interpretation of air quality have also been facilitated by statistical learning models (Imam et al., 2024).

Air-quality classifications can be determined using machine-learning methods based on monitoring and secondary-model data (Liu et al., 2024), and systematic evidence has been established that these methods can be used to accurately predict air quality and make sustainable urban planning decisions (Essamlali et al., 2024). The more comprehensive modelling has also been connected with the pattern of pollutants and the change in the AQI with the change in climate (Ravindiran et al., 2023a). The application of time-series machine learning models has also been employed to predict the emission trends in India, highlighting the broader capabilities of data-driven environmental prediction (Kumari & Singh, 2023).

Hence, the current study is focused on the development and comparison of machine learning models that can predict air-quality level in cities in India. It includes daily environmental monitoring observations, temporal characteristics, geographical variables, prominent pollutants, and historical indicators of AQI. The goals are to describe the spatial and temporal air-quality patterns, assess several classifiers, compare their predictive ability using the balanced metrics and select the most successful in predicting common and severe air-quality classifications.

2. Methodology

2.1 Research Design

The data consisted of 235,785 readings of air quality for each day, for 291 cities in India and 32 states and UTs. The study period was April 2022 – April 2025 (Udayana, 2025). The variables were date, state, city or area, number of monitoring stations, prominent pollutant, Air Quality Index value, measurement unit and air-quality category . The variable for air quality was coded as good, satisfactory, moderate, poor, very poor, and severe. Geographical location, prominent pollutant, number of monitoring stations, and date-derived characteristics were used as predictor variables.

2.2 Data Description

The data was checked for missing data, duplicate data, inconsistent category labels and invalid AQI values. The non-informative note field has been deleted because it was not used to develop the model. Duplicate city–date records were identified and inconsistent AQI classifications were confirmed by reference to the standard AQI classification thresholds. Date variable was structured as a date. To capture the effects due to seasonality and time, temporal variables such as year, month, quarter, season and day of week were created. State, city and prominent pollutant was recoded to numerical values with appropriate coding methods. The AQI value of the same day was not used as an independent variable to predict the air-quality category, for fear of data leakage. Rather, historical AQI indicators were developed such as previous day AQI, three day moving average, seven day moving average and lagged air quality status. The models were able to predict future air-quality conditions using these variables, which were available prior to the monitoring.

2.3 Data Preprocessing

Descriptive statistics were calculated for AQI values, monitoring stations, cities, states, pollutants and air-quality categories. The frequency distributions were analyzed to determine the incidence of each level of air quality. The monthly and seasonal variations were examined to determine the time periods of high pollution. Geographical variations in air quality were also done cities-wise and state-wise. Because the categories of Severe and Very Poor had less observations than the Good, Satisfactory, and Moderate categories, this factor was explored. Thus, the class-weighting techniques were taken into account for the training of models.

2.4 Exploratory Data Analysis

Descriptive statistics were calculated for AQI values, monitoring stations, cities, states, pollutants, and air-quality categories. Frequency distributions were used to examine the prevalence of each air-quality level. Monthly and seasonal trends were assessed to identify periods of elevated pollution. City-wise and state-wise comparisons were also conducted to evaluate geographical variations in air quality.

Class distribution was examined because the Severe and Very Poor categories contained fewer observations than the Good, Satisfactory, and Moderate categories. Class-weighting techniques were therefore considered during model training.

2.5 Machine-Learning Model Development

The data were split into subsets of training and testing data along these lines to ensure that data that occurred before testing were used to predict data that occurred afterwards. The reason for this approach was that random splitting would have meant that information from latter dates might have been included in the training data. A variety of classifiers were created and evaluated, such as multinomial logistic regression, decision tree, random forest, gradient boosting and extreme gradient boosting. Hyperparameters were optimized during the training process by means of cross validation. Where needed, class weights were used to minimize prediction bias for more common classes.

2.6 Model Evaluation

Performance of the model was assessed based on accuracy, balanced accuracy, precision, recall, macro-averaged F1 score and confusion matrix. Due to the imbalance of the category frequencies, both the balanced accuracy and macro-F1 score were deemed more informative than the overall accuracy. The model with the highest accuracy in the classification of both common and rare air quality levels was chosen. The geographical, temporal and pollutant-related factors that are most strongly related to air-quality prediction were then identified using feature-importance analysis.

3.Result

3.1 Dataset Characteristics and Analytical Sample

A total of 235,785 daily air-quality data from 291 cities/monitoring areas in 32 Indian states/union territories were analysed. Observations began on 1st April 2022 and will continue through to 30th April 2025. The mean and standard deviation for the overall mean Air Quality Index were 111.13 and 71.45, respectively. The range of AQI values was 3 to 500, which showed a significant variation in air-quality conditions across locations and time. Successive city-day measurements were only kept for next-day air-quality forecasting. The number of model-ready observations was 219,888 after creating the lagged AQI values and 3-day and 7-day moving averages. A data split was made according to time (chronological), and 169,111 observations were used for training and 50,777 for testing. Observations prior to 15 September 2024 were used to develop the models, and out-of-sample observations were used at a later time. The key features of the data and analytical sample are shown in Table 1: There are 235,785 observations in this dataset, 219,888 of which are suitable for analysis, 291 cities, 32 states and union territories and an AQI range of 3-500.

Table 1. Characteristics of the air-quality dataset and analytical sample

Indicator	Result

Total raw observations	235,785
Model-ready consecutive city-day pairs	219,888
States and union territories	32
Cities or monitoring areas	291
Study period	1 April 2022–30 April 2025
Training observations	169,111
Testing observations	50,777
Chronological testing cutoff	15 September 2024
Mean AQI	111.13
Standard deviation of AQI	71.45
Minimum AQI	3
Maximum AQI	500

3.2 Distribution of Air-Quality Levels

Satisfactory air quality was the biggest category in the model-ready sample (37.59% of the sample). Moderate air quality accounted for 33.21% and Good air quality accounted for 17.47%. The percentage of observations during the Poor, Very Poor and Severe air-quality conditions were combined and were about 11.73%. This Severe category had a total of 511 observations of which 0.23% of the analytical sample was found. This distribution showed that there was a significant class imbalance especially in the cases of Very Poor and Severe air-quality conditions. As can be seen in Table 2, the observations at Satisfactory and Moderate air-quality levels were much more prevalent, with Very Poor and Severe conditions making up only a small percent of the observations for analysis. As seen in figure 1, the air-quality categories in Indian cities were more or less distributed among them, with Satisfactory and Moderate being most common, and Severe being least common.

Table 2. Distribution of air-quality categories in the model-ready sample

Air-quality level	**Frequency**	**Percentage**
Good	38,413	17.47%
Satisfactory	82,661	37.59%
Moderate	73,018	33.21%
Poor	19,939	9.07%
Very Poor	5,346	2.43%
Severe	511	0.23%
Total	219,888	100.00%

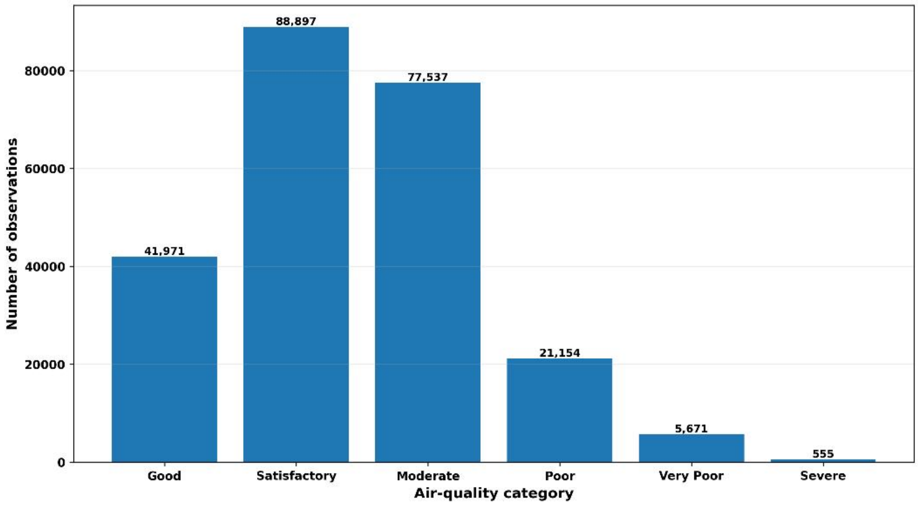

Figure 1. Distribution of recorded air-quality levels across Indian cities

3.3 Temporal Variation in Air Quality

The monthly mean AQI showed significant temporal variability in the study period. The mean air quality Index (AQI) data exhibited repeated rising trend in the post monsoon and winter months while comparatively low AQI was recorded during the monsoon months. The differences imply that there were seasonal tendencies to air quality conditions in Indian cities. The mean AQI varies during the months in Indian cities as shown in Figure 2, with higher pollution levels in the post-monsoon and winter months, and comparatively lower levels in monsoon months.

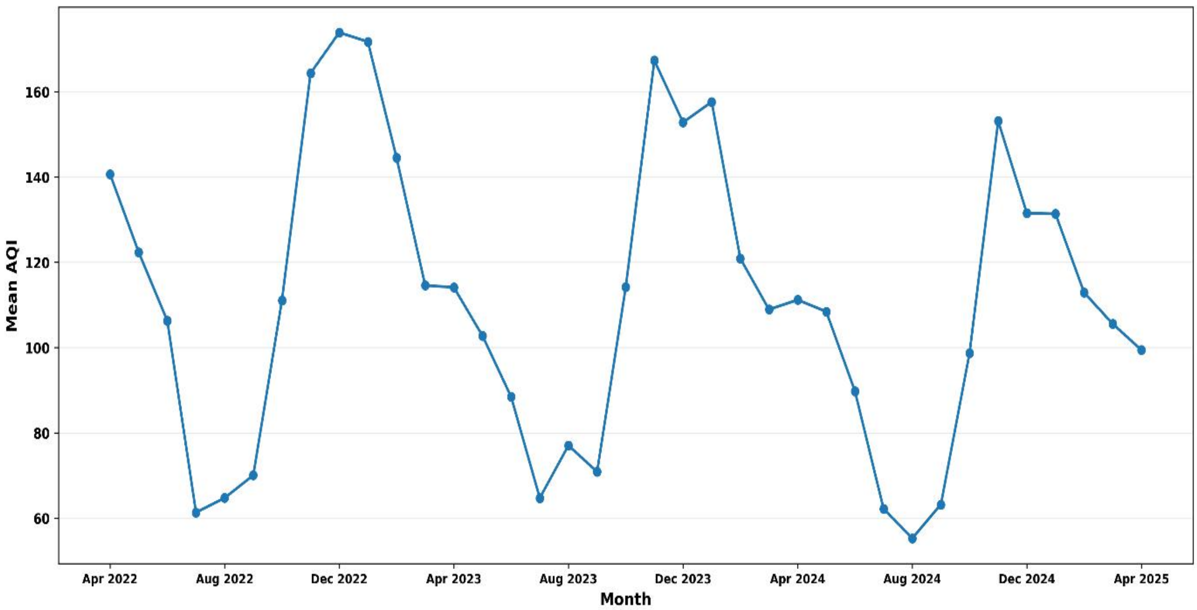

Figure 2. Monthly mean Air Quality Index across Indian cities

3.4 Machine-Learning Model Performance

Four machine learning algorithms were designed: multinomial logistic regression, decision tree, random forest and histogram gradient boosting, which were then evaluated. They were evaluated based on accuracy, balanced accuracy, macro precision, macro recall, macro F1-score and weighted F1-score. The highest overall accuracy (72.16%) was obtained by histogram gradient boosting. However, it performed poorly in minority classes with a balanced accuracy of 53.31%. Random forest had an accuracy of 69.89%, a balanced accuracy of 62.14% and the best macro F1 score of 58.73%. Hence, it was deemed as the most balanced model for forecasting all six types of air-quality.

The highest balanced accuracy was achieved by Multinomial logistic regression (64.24%) while lower than random forest, precision for some categories was lower. The overall predictive ability of the decision tree model was the worst. Table 3 provides the comparative performances of the four machine-learning models, with histogram gradient boosting having the highest accuracy and random forest, the highest macro F1 score. The differences in predictive performance between the machine-learning models are presented in Figure 3, where the overall predictive performance of the histogram gradient boosting model is very good and the predictive performance of the multiclass random forest model is relatively well balanced.

Table 3. Comparative performance of the machine-learning models

Model	Accuracy	Balanced accuracy	Macro precision	Macro recall	Macro F1	Weighted F1
Random forest	0.699	0.621	0.564	0.621	0.587	0.702
Histogram gradient boosting	0.722	0.533	0.626	0.533	0.561	0.720
Multinomial logistic regression	0.673	0.642	0.509	0.642	0.549	0.679
Decision tree	0.643	0.603	0.461	0.603	0.488	0.654

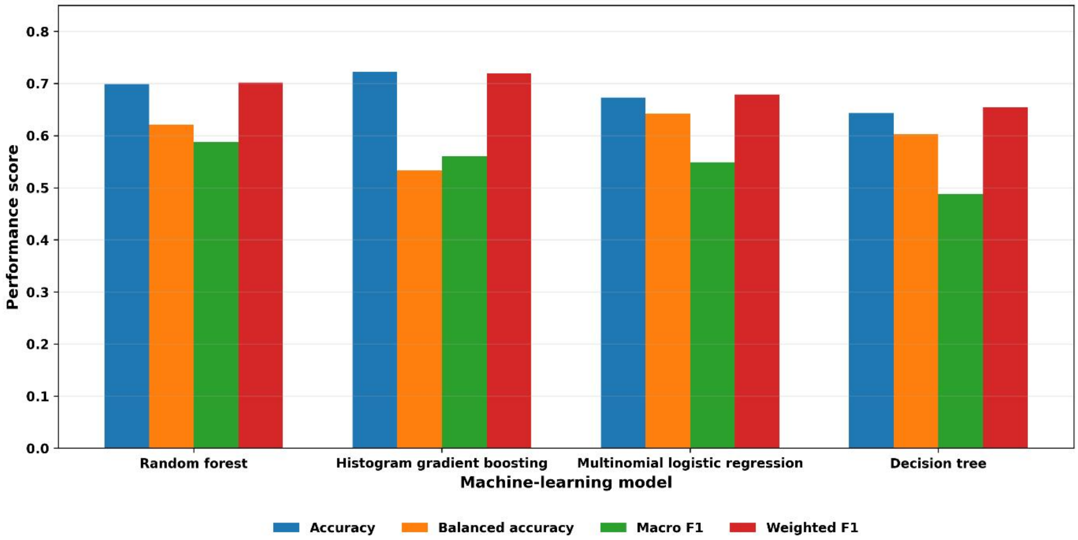

Figure 3. Comparative performance of the machine-learning models

3.5 Category-Wise Performance of the Random Forest Model

The random forest model showed the best results for Good, Satisfactory and Moderate air-quality categories. F1-scores for these categories were 0.739, 0.710 and 0.729, respectively. The model also had a recall rate of 0.661 for the Poor category, meaning that the model correctly detected about 66.1% of the actual Poor observations. The accuracy of predictions was poorer for less common categories. The F1-score for Very Poor air quality was 0.421 and the F1-score for Severe air quality was 0.365. Despite the limited number of Severe observations in the testing sample, the model's performance of about 42.9% (Severe-category recall) shows that the model was able to detect around 42.9% of the Severe observations. The results of random forest model is presented in Table 4, which indicates that the model achieved better precision, recall and F1 score for good, satisfactory and moderate categories than for very poor and severe.

Table 4. Category-wise performance of the random forest model

Air-quality level	Precision	Recall	F1-score	Testing observations
Good	0.722	0.757	0.739	6,226
Satisfactory	0.752	0.673	0.710	18,126
Moderate	0.736	0.721	0.729	20,710
Poor	0.486	0.661	0.560	4,978
Very Poor	0.370	0.488	0.421	688
Severe	0.318	0.429	0.365	49

3.6 Confusion-Matrix Analysis

The random forest model correctly classified a significant percentage of observations from the Good, Satisfactory and Moderate classes, suggesting a good level of accuracy. The majority of the forecast errors were between adjacent categories of AQI. For instance, in some instances the poor observations were further categorized as moderate or very poor, and satisfactory observations were often rated either as good or as moderate. The confusion matrix in figure 4 demonstrates that the model performed best in predicting the categories of Good, Satisfactory and Moderate, and the highest number of errors were made between adjacent AQI categories.

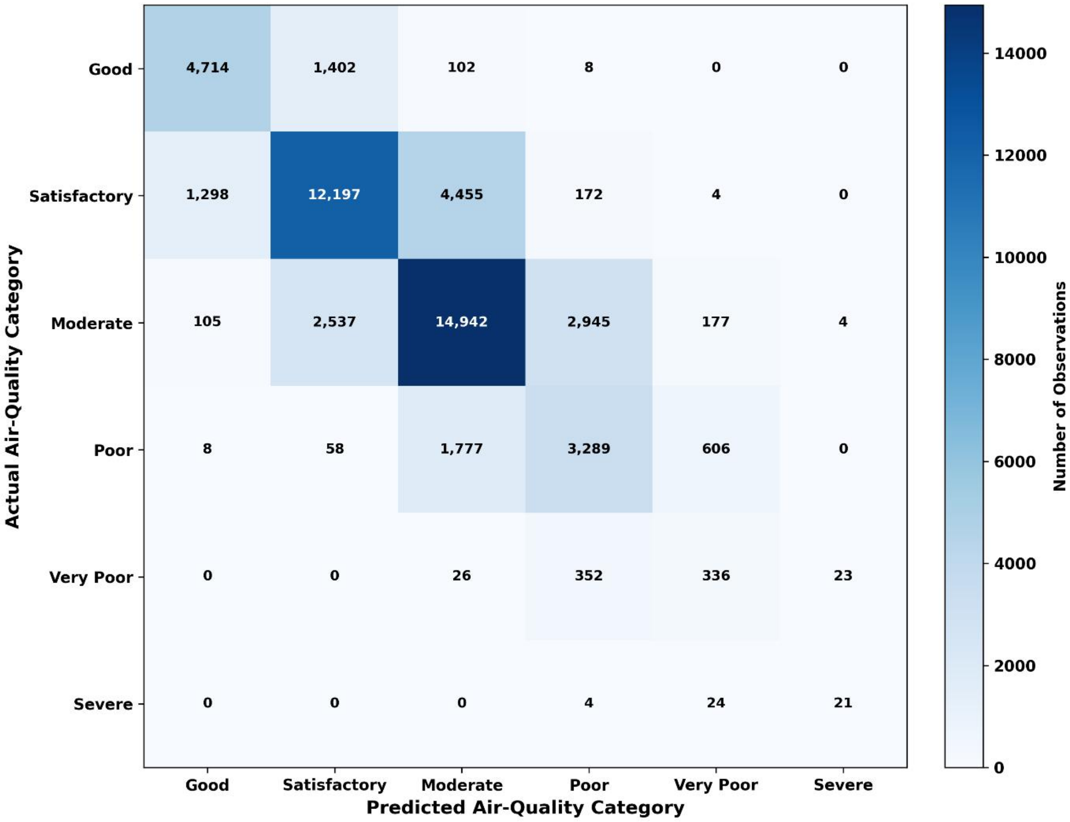

Figure 4. Confusion matrix for the random forest model

3. Discussion

It is natural there's confusion between the adjacent categories as relatively small changes in AQI may result in an observation shifting from one category to another. This limited number of Severe observations also took away the ability of the model to learn regular patterns in this class. The overall findings suggest that historical AQI, temporal features, spatial location, and main pollutants and monitoring information can be useful sources of pre-knowledge to estimate next-day air-quality levels in Indian cities. There are other meteorological and pollutant concentration variables that could be used to help better define Very Poor and Severe conditions.

The present study showed that machine learning based models can predict air quality categories of future day with useful accuracy for Indian cities, by considering historical air quality index, temporal characteristics, location, predominant pollutant and monitoring information. The results corroborate the general trend that machine learning is capable of transforming vast environmental datasets into useful forecasting data for urban pollution control (Gangwar et al., 2023). Chronological training and testing also reinforced the models' applicability since it emulated real-life forecasting scenarios instead of random data splitting. The best overall accuracy was obtained for the model of histogram gradient boosting, the highest for macro F1-score – for random forest, and comparatively balanced in six categories of AQI – for logistic regression. This outcome is in line with comparative studies that have revealed that none of the algorithms are consistently superior in all air quality assessment metrics (Gupta et al., 2023).

Modeling with ensembles is likely to be effective due to its ability to account for nonlinear relations between pollutants, as well as the interactions between temporal and geographical variables (Kothandaraman et al., 2022). But, with the use of class weighting and well-structured lag variables, it is possible that simpler models like multinomial logistic regression can be competitive because of their higher balanced accuracy. The random forest model was especially effective for Good, Satisfactory and Moderate classes. A similar study was conducted in multiple Indian cities conducted by Kumar and Pande (2023) that shows machine learning algorithms performed better on common pollution types. The successful prediction of the dominant AQI levels is also consistent with the optimized modelling results reported for major cities in India by Natarajan et al. (2024). Research using deep learning has revealed that there is significant predictive information in historical sequences, especially when air-quality conditions are similar on back-to-back days (Janarthanan et al., 2021).

The research findings of regional forecasting also validate that continuity in time is one of the determinants of the predictive performance (Abirami & Chitra, 2021). The air quality rating of Very Poor and Severe had lower performance. Many of the reasons for this were due to the few observations for these categories, particularly Severe conditions. One of the problems commonly faced by air-quality classification is the class imbalance, as rare pollution episodes do not offer enough examples to reliably learn a stable model (Liu et al., 2024). With the hybrid deep-learning approaches, more complicated temporal feature can be extracted to enhance the recognition rate of minority categories (Sarkar et al., 2022). Even with the small sample size, however, the Severe-category recall from this study showed that the random forest model had some accuracy for identifying high-risk conditions.

The confusion matrix demonstrates that the majority of errors were between neighbouring AQI categories. This is a reasonable pattern as small changes in AQI can move an observation either up or down a classification from Satisfactory to Moderate or Moderate to Poor. Similar classification issues have been reported in the city of Ahmedabad, where the AQI classes in nearby cities had similar environmental parameters (Maltare & Vahora, 2023). Moreover, studies conducted in Hyderabad have also pointed to the difficulty in separating the levels of closely related pollutions in space and time (Gokul et al., 2023). Prediction performance differences between metropolitan areas may also be due to environmental heterogeneity at the city level (Mathew et al., 2023). The Monthly results showed that the AQI value shows a definite seasonal variation with higher AQI in the post-monsoon and winter season and lower AQI in the monsoon season. Meteorological research has shown that humidity, wind, rainfall and temperature have a significant effect on the accumulation and dispersion of pollutants (Mampitiya et al., 2023). The urban thermal environment can also amplify regional air pollution scenarios, as demonstrated in Bengaluru, where there is a correlation between LST and air pollution (Suthar et al., 2024). Delhi PM2.5 prediction studies similarly show such seasonal factors (Masood & Ahmad, 2023). Due to the lack of direct meteorological measurements in the present data set, this may result in limited predictive accuracy. The results also demonstrate the need for city-specific modelling. Pollution behaviour might vary between cities in coastal regions and metropolitan areas in the interior due to the wind circulation and influence of the sea (Ravindiran et al., 2023b).

Particulate pollution is found to be influenced by very localised emission and atmosphere condition over the Delhi region in the modelling exercise (Kumar et al., 2020). Wider climate-related pollution analyses indicate that climate change over longer time horizons could also influence AQI trends (Ravindiran et al., 2023b). The results can be used to promote the use of machine-learning models in smart-city monitoring systems as applied. Making predictions using sensors can enhance Early Warning System, regulatory planning and public communication (Iskandaryan et al., 2020). Real-time update of AQI predictions using the new monitoring data may be possible with the help of IoT-based systems (Karnati, 2023).

Furthermore, the statistical learning approach can enhance the sustainable environmental governance by the detection of emerging pollution threats (Imam et al., 2024). Concentrations of pollutants, weather, traffic conditions, land-use indicators and satellite data should be included in future models. This integration could enhance the prediction of rare Severe events and enhance the usefulness of the forecasing systems for public-health protection.

5.Conclusion

In this study the authors have shown that machine learning techniques can effectively predict air-quality categories for the next day in Indian cities from environmental monitoring data. The models developed were able to capture meaningful spatial variation and temporal characteristics of urban air quality with the inclusion of historical AQI values, temporal variation, geographical information, prominent pollutants, and monitoring-station coverage. The results indicated that the best overall accuracy was obtained by the use of the histogram gradient boosting and the best macro F1-score was obtained by the use of random forest with a balanced performance across the six air quality categories. The balanced accuracy was also comparable to that of multinomial logistic regression, which suggests that simpler classification models can be helpful when additional lagged and contextual variables are added. The models were better at predicting air-quality conditions of Good, Satisfactory and Moderate than air-quality conditions of Poor, Very Poor and Severe. This difference was primarily related to the considerable imbalance in category frequencies, especially the small number of Severe observations. The confusion matrix also revealed that there are a high number of classification errors between adjacent AQI levels, which is because of the similarity in environmental conditions near the thresholds between the adjacent levels. Seasonal pattern was also observed and showed an increase in pollution during winter and post-monsoon and decrease in AQI during monsoon months. In summary, the results provide strong evidence of the potential for machine learning to assist with air-quality monitoring, forecasting, and evidence-based urban environmental management in India. Future work should incorporate meteorological factors, individual pollutant levels, the intensity of traffic, land use factors, satellite observations, and advanced class-balancing methods to better characterize the occurrence of rare pollution events.

References:

1.Abirami, S., & Chitra, P. (2021). Regional air quality forecasting using spatiotemporal deep learning. Journal of cleaner production, 283, 125341.

2.Essamlali, I., Nhaila, H., & El Khaili, M. (2024). Supervised machine learning approaches for predicting key pollutants and for the sustainable enhancement of urban air quality: A systematic review. Sustainability, 16(3), 976.

3.Gangwar, A., Singh, S., Mishra, R., & Prakash, S. (2023). The state-of-the-art in air pollution monitoring and forecasting systems using IoT, big data, and machine learning. Wireless Personal Communications, 130(3), 1699-1729.

4.Gokul, P. R., Mathew, A., Bhosale, A., & Nair, A. T. (2023). Spatio-temporal air quality analysis and PM2. 5 prediction over Hyderabad City, India using artificial intelligence techniques. Ecological Informatics, 76, 102067.

5.Gupta, N. S., Mohta, Y., Heda, K., Armaan, R., Valarmathi, B., & Arulkumaran, G. (2023). Prediction of air quality

index using machine learning techniques: a comparative analysis. Journal of Environmental and Public Health, 2023(1), 4916267.
6.Hussain, A., Draz, U., Ali, T., Tariq, S., Glowacz, A., Irfan, M., ... & Rahman, S. (2020). Waste management and prediction of air pollutants using IoT and machine learning approach. Energies, 13(15), 1-22.
7.Imam, M., Adam, S., Dev, S., & Nesa, N. (2024). Air quality monitoring using statistical learning models for sustainable environment. Intelligent Systems with Applications, 22, 200333.
8.Iskandaryan, D., Ramos, F., & Trilles, S. (2020). Air quality prediction in smart cities using machine learning technologies based on sensor data: a review. Applied Sciences, 10(7), 2401.
9.Janarthanan, R., Partheeban, P., Somasundaram, K., & Elamparithi, P. N. (2021). A deep learning approach for prediction of air quality index in a metropolitan city. Sustainable Cities and Society, 67, 102720.
10.Karnati, H. (2023). IoT-based air quality monitoring system with machine learning for accurate and real-time data analysis. arXiv preprint arXiv:2307.00580.
11.Kothandaraman, D., Praveena, N., Varadarajkumar, K., Madhav Rao, B., Dhabliya, D., Satla, S., & Abera, W. (2022). Intelligent forecasting of air quality and pollution prediction using machine learning. Adsorption Science & Technology, 2022, 5086622.
12.Kumar, K., & Pande, B. P. (2023). Air pollution prediction with machine learning: a case study of Indian cities. International Journal of Environmental Science and Technology, 20(5), 5333-5348.
13.Kumar, S., Mishra, S., & Singh, S. K. (2020). A machine learning-based model to estimate PM2. 5 concentration levels in Delhi's atmosphere. Heliyon, 6(11).
14.Kumari, S., & Singh, S. K. (2023). Machine learning-based time series models for effective CO2 emission prediction in India. Environmental Science and Pollution Research, 30(55), 116601-116616.
15.Liu, Q., Cui, B., & Liu, Z. (2024). Air quality class prediction using machine learning methods based on monitoring data and secondary modeling. Atmosphere, 15(5), 553.
16.Maltare, N. N., & Vahora, S. (2023). Air Quality Index prediction using machine learning for Ahmedabad city. Digital Chemical Engineering, 7, 100093.
17.Mampitiya, L., Rathnayake, N., Leon, L. P., Mandala, V., Azamathulla, H. M., Shelton, S., ... & Rathnayake, U. (2023). Machine learning techniques to predict the air quality using meteorological data in two urban areas in Sri Lanka. Environments, 10(8), 141.
18.Masood, A., & Ahmad, K. (2023). Data-driven predictive modeling of PM2. 5 concentrations using machine learning and deep learning techniques: a case study of Delhi, India. Environmental Monitoring and Assessment, 195(1), 60.
19.Mathew, A., Gokul, P. R., Raja Shekar, P., Arunab, K. S., Ghassan Abdo, H., Almohamad, H., & Abdullah Al Dughairi, A. (2023). Air quality analysis and PM2. 5 modelling using machine learning techniques: A study of Hyderabad city in India. Cogent Engineering, 10(1), 2243743.
20.Natarajan, S. K., Shanmurthy, P., Arockiam, D., Balusamy, B., & Selvarajan, S. (2024). Optimized machine learning model for air quality index prediction in major cities in India. Scientific reports, 14(1), 6795.
21.Pande, C. B., Radwan, N., Heddam, S., Ahmed, K. O., Alshehri, F., Pal, S. C., & Pramanik, M. (2025). Forecasting of monthly air quality index and understanding the air pollution in the urban city, India based on machine learning models and cross-validation. Journal of Atmospheric Chemistry, 82(1), 1.
22.Rautela, K. S., & Goyal, M. K. (2024). Transforming air pollution management in India with AI and machine learning technologies. Scientific Reports, 14(1), 20412.
23.Ravindiran, G., Hayder, G., Kanagarathinam, K., Alagumalai, A., & Sonne, C. (2023). Air quality prediction by machine learning models: A predictive study on the indian coastal city of Visakhapatnam. Chemosphere, 338, 139518.
24.Ravindiran, G., Rajamanickam, S., Kanagarathinam, K., Hayder, G., Janardhan, G., Arunkumar, P., ... & Muniasamy, S. K. (2023). Impact of air pollutants on climate change and prediction of air quality index using machine learning models. Environmental Research, 239, 117354.
25.Sarkar, N., Gupta, R., Keserwani, P. K., & Govil, M. C. (2022). Air quality index prediction using an effective hybrid deep learning model. Environmental Pollution, 315, 120404.
26.Suthar, G., Kaul, N., Khandelwal, S., & Singh, S. (2024). Predicting land surface temperature and examining its relationship with air pollution and urban parameters in Bengaluru: A machine learning approach. Urban Climate, 53, 101830.
27.Udayana, S. K. (2025). India Air Quality Index (AQI) Dataset [2023–2025] [Data set]. Kaggle. https://www.kaggle.com/datasets/saikiranudayana/india-air-quality-index-aqi-dataset