GIET UNIVERSITY JOURNAL, 2026, VOL. 6, NO. 1, PP. 1-6
Manuscript received 10th October 2025; revised 21st January 2026;
Accepted 30th January 2026

ISSN: 2583-5955

PREDICTIVE MODELING OF AIR QUALITY INDEX USING MACHINE LEARNING TECHNIQUES

Manish Pradhan, Dept. of CSA, GIET University Gunupur, 24bca014.manishpradhan@giet.edu
Supratika Padhi, Dept. of CSA, GIET University Gunupur, 24bca054.supratikapadhi@giet.edu
Subham Nahak, Dept. of CSA, GIET University Gunupur, 24bca089.subhamnahak@giet.edu
Soumya Ranjan Mishra, Dept. of CSA, GIET University Gunupur, soumyaranjan@giet.edu
Prahallad Kumar Sahu, Dept. of CSA, GIET University Gunupur, prahalladsahu@giet.edu

Abstract: The air quality index is a critical measure for human health regarding the effect of air pollution. This paper presents the contributions to prediction research by describing various methodologies to predict AQI focusing on the most recent application such as synthesic minority oversampling technique (SMOTE). It was observed that semi-automatic feature engineering improves upon unsmoted datasets due to high prediction accuracy of RFR in conjunction with the enhanced datasets and low root mean square error (RMSE) measurements for most cities. Moreover, CatBoost regression establishes a top accuracy of 85.08% for New Delhi and 90.31% for Bangalore. From the support vector regression and CatBoost regression methods, it is evident that Kolkata (93.74%) and Hyderabad (97.60%) provide better performance. The novelty of this study incorporated model selection, data set balancing through SMOTE, and detailed graphical and metric-based comparisons. The results provide sufficient evidence of the growing potential for prediction accuracy improvement using SMOTE techniques to be brought into the view for climate control and air quality management.

***Keywords*— Pollution levels, PM2.5, PM10, Ozone (O_3), Carbon Monoxide (CO), Sulfur Dioxide (SO_2), Nitrogen Dioxide (NO_2), Particulate Matter.**

1. Introduction:

Man cannot live without air. It is incumbent upon us to keep air quality monitoring and understanding well designed to be beneficial to human health. Every year, tens of thousands of people fall sick from cardio-respiratory problems due to air pollution. For example, scientific evidence suggests that the single most serious environmental risk is air pollution [1]. Owing to rapid industrial development, there is an increase in the population levels, and the emissions of toxic gases deteriorate the environment. As a result, our health has been severely affected because of the contamination of air with harmful pollutants. These factors, when combined, have greatly worsened the air quality.

The air quality indicator, AQI, refers to an index which measures the degree of air pollution. Defined 12 parameters (air pollutants) as part of determining AQI of a region's air quality as NO_2, SO_2, CO, O_3, PM10 and PM2.5, NH_3, and Benzene. Other six for simply calculating AQI are: PM10, PM2.5, SO_2, NO_2, CO, and O_3 [2]. The choice for actual selection of pollutants changes largely around the specific objective and around a number of other factors like data, measurement

GIET UNIVERSITY JOURNAL, 2026, VOL. 6, NO. 1, PP. 1-6
Manuscript received 10th October 2025; revised 21st January 2026;
Accepted 30th January 2026

ISSN: 2583-5955

methods, and the rate of monitoring. Thus, high AQI indicia tors means that the person survives in heavily polluted air, which adversely produces serious effects on human health. The AQI can be used for real-time monitoring of air quality [3]. There are many weather stations capturing daily and hourly information of AQI right from where we breathe. Data will be mined and harvested for use in suggested work.

2. Literature Review

This study began with an examination of the relationship between various air indicators, including AQI, PM2.5 concentrations, total NOx (nitrogen oxides) concentrations, among other variables cited in this article [1Similarly, the predictions made by the researchers using random forest regression (RFR) and support vector regression (SVR) were evaluated using RMSE, coefficient of determination (R-SQUARE) and correlation coefficient r [4]. The surface ozone particulates PM2.5, along with California's hourly AQI can be predicted with a greater degree of certainty if SVR with RBF kernel is used. However, validation of unseen data is categorized into six categories of AQI as defined by the US Environmental Protection Agency (dataset) at an accuracy of 94.1 percent [5].

The AQI foretelling from machine learning methods, be it time-series analysis or linear Regression. The ARIMA and multiple linear regression techniques, as well as supervised machine learning, were used for predicting the AQI [6]. Diverse metrics quantitatively evaluated the performance. The model should predict AQI in future, ARIMA being over this time series model implicated. Both models were accurate and scored high in efficiency forecasting the AQI [7].

Another model uses a combination of an artificial neural network along with Kreiging method for estimating concentration of air pollutants at different locations Mumbai and Navi Mumbai [8]. There were quite high R values which implied good fit between anticipated and observed values. As far as R value on projection is concerned, ANN was found to outdo simple regression models [9]. It will model predictions of AQI concentrations according to parameters PM2.5, PM10, SO2, and NO2. As a conclusion, among these algorithms, the that achieved the greatest accuracy of prediction that reached 0.99985 on test data was random forest regression with a minimum mean square error of 0.00013 and mean absolute error of 0.00373 as compared to linear regression, decision tree regression, and SVR. [10].

3. Methodology:

Comparative analysis of AQI values for the cities of Bangalore, Kolkata, Hyderabad based on readings of parameters such as PM2.5, PM10, NO, NO_2, NO_x, NH_3, CO, SO_2, O_3, the next step being to analyze the three competing algorithms for accuracy and efficiency [11]. The authors would intend to investigate and present this in a more efficient manner-values that could lead to

GIET UNIVERSITY JOURNAL, 2026, VOL. 6, NO. 1, PP. 1-6
Manuscript received 10th October 2025; revised 21st January 2026;
Accepted 30th January 2026

ISSN: 2583-5955

further very interesting and insightful information. What pollution is in a South Asian was represented by these cities, noticeable for their high population densities. Other cities were discarded from this study because that would make the research paper rather lengthy. Thus, major Indian cities were chosen for pollution level analysis in urban settings as these contribute the most to pollution levels. The following are existing algorithms [12]. This gives an outline of a few algorithms, like Naive Bayes-a classifier grounded on Bayes' theorem. Artificial Neural Networks-an elaborative paradigm based on real brain neurons; Gradient Boosting- ensemble methods using weak predictive models; Decision Trees-a predictive modeling method grounded on object attributes; K-Nearest Neighbor-an instance-based nonparametric supervised learning method [13].

3.1 Data Collection: Dataset used for the study included historical AQI readings that were accessed from public sources as well:

- Environmental Protection Agencies
- Open AQ API and/or Kaggle datasets

The dataset contains daily or hourly concentrations of following key pollutants:

- SO_2, O_3
- Meteorological parameters: temperature, humidity, wind velocity

3.2 Data Preprocessing: The following steps in the preprocessing stage have to be done to get the inputs of Prophet model to highest quality:

- Handling Missing Data – Here the missing values were put in by means of linear interpolation or removed due to high messiness.
- Outlier Detection – Z-score or IQR (Interquartile Range) methods were used to eliminate extreme outlier values.
- Feature Engineering – Results included the creation of lagged AQI values and rolling averages as new variables.
- Data Formatting – It shall have two columns for the input to Prophet:
  1. ds (Date/Time Stamp) – Timestamps at which AQI readings were taken
  2. y (AQI Value) – AQI readings corresponding to the time stamps.

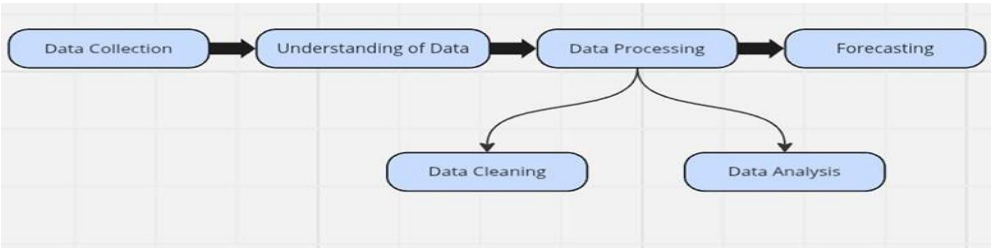

Fig. 1. Flow chart of the proposed model

4. Result & Discussion

GIET UNIVERSITY JOURNAL, 2026, VOL. 6, NO. 1, PP. 1-6
Manuscript received 10th October 2025; revised 21st January 2026;
Accepted 30th January 2026

ISSN: 2583-5955

In the proposed research, the aforementioned data set was cleaned to include only the cities of Bangalore, Hyderabad, Kolkata, New Delhi [14]. Dataset predefined was utilized first in its imbalanced form and then in its balanced state by using SMOTE. Graphs were done to be able to prove the improved accuracy of the balanced dataset models. Three prediction algorithms were evaluated: support vector regression, random forest regression, and Cat Boost regression [15]. This time, plots showing the graph inputs of the test data versus the actual predicted data were displayed. These metrics calculated were R-SQUARE, MSE, RMSE, and MAE for each of the algorithms. It generated in order to compare the results from both the balanced dataset for training and the imbalanced dataset for training, indicating the respective influence of applying the balanced dataset on various algorithms on the ultimate accuracies [17]. But then, in the light of this research output, the parameters for choosing a model, such as RMSE and R-square, among others, have been derived and discussed in papers [18-19], as well as the effective way by which such parameters can be used. Performance Metrics relate to a model while being trained and tested. Such metrics provide information about the accuracy of predictions and error with respect to the actual value due to all algorithms being applied from regression models. We provide a comparative report of the accuracy results from all four Cities Delhi, Bengaluru, Kolkata, and Hyderabad-entailing machine learning methods like support vector regression, random forest regression, and Cat Boost regression, on imbalanced data set where SMOTE was not applied [20].

5. Conclusion

It is a worldwide problem that air pollution has an impact on humanity, and researchers from several countries are trying to find a solution. Machine learning techniques were considered to predict AQI accurately. This research will critically assess the performances of three best data mining models involved in the prediction of accurate AQI data of the most crowded, polluted cities in India, namely, SVR, RFR, and CR. In order to fix class imbalance, model builders commonly use Synthetic Minority Over-Sampling Technique (SMOTE) to copy examples from the minority classes, enhancing their analysis. This method of balancing datasets and applying them was very promising for high accuracy in comparison with the imbalanced results. Using statistical methods, like RMSE, MAE, MSE, and R-SQUARE, to confirm results could contribute toward execution of this unique approach. The study done balanced versus imbalanced datasets used such applications extensively tabulated which may serve to be a good reference in future.

6. References

1. H. Liu, Q. Li, D. Yu, and Y. Gu, “Air quality index and air pollutant concentration prediction based on machine learning algorithms,” Applied Sciences, vol. 9, p. 4069, 2019.
2. M. Castelli, F. M. Clemente, A. Popovic, S. Silva, and L. Vanneschi, “A machine learning approach to predict air quality in California,” Complexity, vol. 2020, Article ID 8049504, 23 pages, 2020.

ISSN: 2583-5955

3. Mishra, S. R., Dash, S., Padhy, S., Kumar, N., & Dash, Y. (2024, September). Integrating Multi-Omics Data for Advanced Diabetes Prediction and Understanding. In 2024 7th International Conference on Contemporary Computing and Informatics (IC3I) (Vol. 7, pp. 1447-1453). IEEE.
4. G. Mani, J. K. Viswanadhapalli, and A. A. Stonie, "Prediction and forecasting of air quality index in Chennai using re gression and ARIMA time series models," Journal of Engi neering Research, vol. 9, 2021.
5. S. V. Kottur and S. S. Mantha, "An integrated model using Artifcial Neural Network (ANN) and Kriging for forecasting air pollutants using meteorological data," Int. J. Adv. Res. Comput. Commun. Eng, vol. 4, pp. 146–152, 2015.
6. Dash, S., Mishra, S. R., & Baboo, A. (2025, January). Enhancing Diabetes Prediction using Hybrid Ensemble Approach. In 2025 International Conference on Intelligent Systems and Computational Networks (ICISCN) (pp. 1-6). IEEE.
7. S. Halsana, "Air quality prediction model using supervised machine learning algorithms," International Journal of Sci entifc Research in Computer Science, Engineering and Infor mation Technology, vol. 8, pp. 190–201, 2020.
8. A. G. Soundari, J. Gnana, and A. C. Akshaya, "Indian air quality prediction and analysis using machine learning," International Journal of Applied Engineering Research, vol. 14, p. 11, 2019
9. Dash, S., Mishra, S. R., & Baboo, A. (2025, January). Efficient Prediction of Diabetes Mellitus Through Hybrid Ensemble Machine Learning Model Using IoT. In 2025 1st International Conference on AIML-Applications for Engineering & Technology (ICAET) (pp. 1-6). IEEE.
10. C. R.Aditya, C. R. Deshmukh, N.DK,P.Gandhi,andV.astu, "Detection and prediction of air pollution using machine learning models," International Journal of Engineering Trends and Technology, vol. 59, no. 4, pp. 204–207, 2018.
11. J. Kleine Deters, R. Zalakeviciute, M. Gonzalez, and Y. Rybarczyk, "Modeling PM2. 5 urban pollution using machine learning and selected meteorological parameters," Journal of Electrical and Computer Engineering, vol. 2017, Article ID 5106045, 14 pages, 2017.
12. Mishra, S. R., & Dash, S. (2024, December). Machine Learning Based Diabetes Prediction Using the PIMA Indian Dataset. In 2024 2nd International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES) (pp. 1-6). IEEE.
13. P. Bhalgat, S. Pitale, and S. Bhoite, "Air quality prediction using machine learning algorithms," International Journal of Computer Applications Technology and Research, vol. 8, pp. 367–370, 2019.
14. Dora, N., Dash, S., Baboo, A., & Mishra, S. R. (2025, August). Efficient Nail Disease Diagnosis Using Deep Neural Networks for Predicting Abnormalities. In 2025 International Conference on Next Generation of Green Information and Emerging Technologies (GIET) (pp. 1-5). IEEE.
15. Khuntuli, B., Dash, S., Pradhan, P. C., & Mishra, S. R. Combating food insecurity through remote sensing and machine learning for enhanced crop yield prediction. In Intelligent Computing Techniques and Applications (pp. 135-140). CRC Press.
16. Sahu, P. K., Biswal, B. B., Mishra, S. R., Padhy, J., & Kumar, D. (2025, March). Demand-Based Secured Data Transmission in WSN. In International Conference on Next Generation Computing and Communication Applications (pp. 37-44). Cham: Springer Nature Switzerland.
17. M. Bansal, "Air quality index prediction of Delhi using LSTM," Int. J. Emerg. Trends Technol. Comput. Sci, vol. 8, pp. 59–68, 2019.

GIET UNIVERSITY JOURNAL, 2026, VOL. 6, NO. 1, PP. 1-6
Manuscript received 10th October 2025; revised 21st January 2026;
Accepted 30th January 2026

ISSN: 2583-5955

18. A. Shishegaran, M. Saeedi, A. Kumar, and H. Ghiasinejad, "Prediction of air quality in Tehran by developing the nonlinear ensemble model," Journal of Cleaner Production, vol. 259, Article ID 120825, 2020.
19. Mishra, S. R., Dash, S., & Rath, L. (2024, November). Effective Diabetes Mellitus Prediction Using a Hybrid Ensemble Machine Learning Model with Iot. In 2024 International Conference on Integrated Intelligence and Communication Systems (ICIICS) (pp. 1-8). IEEE.
20. L. Tuan-Vinh, "Improving the awareness of sustainable smart cities by analyzing lifelog images and IoT air pollution data," in Proceedings of the 2021 IEEE International Conference on Big Data (Big Data), IEEE, Orlando, FL, USA, September 2021.
21. R. Kumar, P. Kumar, and Y. Kumar, "Time series data prediction using IoT and machine learning technique," Procedia Computer Science, vol. 167, no. 2020, pp. 373–381, 2020.
22. H. Maleki, A. Sorooshian, G. Goudarzi, Z. Baboli, Y. Tahmasebi Birgani, and M. Rahmati, "Air pollution pre diction by using an artifcial neural network model," Clean Technologies and Environmental Policy, vol. 21, no. 6, pp. 1341–1352, 2019.
23. K. P. Singh, S. Gupta, and P. Rai, "Identifying pollution sources and predicting urban air quality using ensemble learning methods," Atmospheric Environment, vol. 80, pp. 426–437, 2013.
24. S. Hansun and M. Bonar Kristanda, "AQI measurement and prediction using B-wema method," International Journal of Engineering Research and Technology, vol. 12, pp. 1621– 1625, 2019.
25. Mishra, S. R., & Dash, S. (2026). AI-Driven Remote Health Monitoring for Predicting Diabetes and Heart Diseases Using ULMCSO and PGND Models. Hyper-Intelligent Networks: Exploring the Future of Connectivity for Society 5.0, 219-247.
26. R. Janarthanan, P. Partheeban, K. Somasundaram, and P Navin Elamparithi, "A deep learning approach for pre diction of air quality index in a metropolitan city," Sus tainable Cities and Society, vol. 67, no. 2021, Article ID 102720, 2021.
27. M. Londhe, "Data mining and machine learning approach for air quality index prediction," International Journal of Engi neering and Applied Physics, vol. 1, no. 2, pp. 136–153, May 2021.
28. R. W. Gore and D. S. Deshpande, "An approach for classi fcation of health risks based on air quality levels," in Pro ceedings of the 20171st International Conference on Intelligent Systems and Information Management (ICISIM), pp. 58–61, Aurangabad, India, October 2017.
29. X. Zhao, M. Song, A. Liu, Y. Wang, T. Wang, and J Cao, "Data-Driven temporal-spatial model for the prediction of AQI in nanjing," Journal of Artifcial Intelligence and Soft Computing Research, vol. 10, no. 4, pp. 255–270, 2020.