Grenze International Journal of Engineering and Technology, January Issue

Air Quality Forecasting using Near-Miss Approaches: A Predictive Model

Diksha Mehra[1] and Davinder Kumar[2]
[1-2]Chandigarh University/Department of Mathematics, Mohali, India
Email: diksha.mehra2001@gmail.com, davinder.e15548@cumail.in

***Abstract*—One reporting metric for air quality is the air quality index (AQI). The purpose of the AQI is to alert the public, for a limited time, to the harmful impact that local air pollution has on health. Delhi's air pollution levels have significantly increased. The objective of this project is to determine how best to use prediction of AQI to assist temperature regulation. In order to guarantee that the most practical solution to the problem of air quality is found, this job combines a detailed investigation with the incorporation of Near Miss approaches. Three different techniques have been used in the proposed work: XGBoost regression (XGR), K-nearest neighbour regression (KNNR), and decision tree regression (DTR) to determine the AQI of New Delhi. The investigation's findings demonstrate that, in terms of predicting New Delhi's air quality index, the XGBoost-NM model is clearly superior. It reaches the highest classification performance of 93.04% accuracy for AQI prediction by using the NM balanced approach. It is unequivocally shown that datasets treated with the Near Miss algorithm yielded results with greater accuracy.**

***Index Terms*— AQI, New Delhi, ML techniques, Near Miss, air quality, accuracy, prediction.**

I. INTRODUCTION

Numerous environmental changes over the last few decades have contributed to a decline in air quality, which is influenced by various contaminants and the current climate. Due to car and industrial emissions, high concentrations of pollutants contaminate the atmosphere, particularly in metropolitan areas, endangering human health as well as the ecosystem as a whole [1, 2]. The WHO estimates that nine out of ten people on the planet breathe contaminated air. According to estimates, air pollution causes 7 million deaths annually [3]. Air pollution is responsible for one-third of deaths from heart disease, lung cancer, and stroke [4]. A place's quality of life is determined by a number of elements, one of which is its air quality. Its measurement, known as the AQI, is dependent on the amount of contaminants in the atmosphere.

India is a nation in development. India's air pollution levels are rising in tandem with industrialization and urbanization. Delhi, the capital city of India, is ranked 11th overall and is the most polluted capital in the world [5]. On October 17, 2014, the Ministry of Forest, Climate Change and Environment developed and put into use the AQI system in order to monitor and assess the ambient air quality in India [6]. When determining the AQI, the Indian AQI System (IND-AQI) takes into account the following eight contaminants: NO2, CO, SO2, PM2.5, PM10, O3, Pb, and NH3.

As an outcome, the dataset that was employed included information of the AQIs in New Delhi. The most precise results will be ascertained by evaluating three different regression evaluation techniques. The effectiveness of a

Grenze ID: 01.GIJET.11.1.306_1

dataset is investigated in the suggested investigation both before and after the Near Miss approach is applied. The main purpose is how Near Miss is being used. Unlike other papers, this one makes utilises Near Miss to account for the consequences of an imbalanced dataset. Additionally, diagrams and metrics that detail each step have been used to display all methods and performance metrics under all datasets, in both its imbalanced and balanced forms. The ability of the suggested methodologies to accurately forecast future AQI values will serve as a warning and emphasise the urgency of reducing air pollution conditions.

II. LITERATURE REVIEW

A. Index of Air Quality (AQI)

Using a straightforward instrument known as a diffuse tube, the air quality index is used to measure air quality and forecast whether or not the air is good for health issues [7]. It also shows temperature changes by detecting air pollutants such as nitrogen dioxide and traffic-related pollutants. The diffuse tube reduces health risks associated with the air quality index by monitoring air pollutants and absorbing particular pollutants from the air without the need for a power source

B. Keeping an eye on air pollution

It's a good idea to monitor air pollution through the use of air quality measurement tools. The air quality index provides information on pollution levels, which have a greater impact on health [8]. As a result, it monitors improved air quality to help people live healthier lives and cope with life's challenges. By using this monitoring system, our health issues will be more effectively identified than without monitoring periods, allowing us to identify those controls to monitor the air pollutants in the right way to simulate the pollutant in which the controlled system is utilised to monitor the air, which is affecting our health harshly, and by using the correct monitor, we can survive the life in a good manner.

C. Existing System for AQI

Machine learning (ML) techniques like time series analysis and linear regression are used to produce the AQI prediction. Employing supervised machine learning and MLR techniques, the AQI was forecast. To evaluate the results, a number of quantitative measures were used. Second, the ARIMA time series method was used to predict the AQI into the future. It was found that the AQI could be predicted rather well and accurately by both models [9]. The AQI was estimated based on data from the prior year and predicted over a particular future year utilising gradient descent in order to solve the multivariable regression problem. They increased the model's effectiveness and surpassed conventional regression models by applying estimated costs for the prediction issue. In order to determine the order of selection depending on how near the ideal solution the choices were, they also employed the AHP MCDM approach [10]. A deep RNN based on LSTM was investigated as an effective method to forecast hourly pollution concentrations and, consequently, Delhi's AQI. The results were accurate even in hourly projections. The findings [11] showed that deep learning-based tactics outperformed conventional statistical techniques [12].

A combined approach that computed air pollution at various Mumbai locales employed Kriging and artificial neural networks methodology. A high R-value indicated that the required the difference between the observed and expected values had been reached. ANN performed better than basic regression models in terms of forecast and R value [13]. Among the four algorithms—linear regression, decision tree regression, SVR, and RFR— the random forest regression methodology produced the best accuracy of 0.98985 on the test data with the least mean square error of 0.00123 and the mean absolute error of 0.014773 when it came to predicting AQI author concentrations with parameters like PM2.5, PM10, SO2, and NO2 [14].

III. METHODOLOGY

In this study, variables like NO, $PM_{2.5}$, PM_{10}, NO_2, O, CO, NH_3 and SO_2 levels are used to assess the AQI values of New Delhi employing three different methods. The three algorithms are then compared to determine which approach is the most accurate and efficient. Analysing and presenting it effectively is the goal. Figure 1 outlines the several actions that will be taken in order to carry out this work and get the intended outcome. The diagram offers a comprehensive overview of all the steps in the process and is process-based.

A. Near Miss algorithm (NM)

An uneven dataset can be balanced with the help of a technique known as near-miss. It belongs to the class of under sampling algorithms and is an effective way to balance the data. The method looks at the class distribution and chooses samples at random from the bigger class to achieve this. In an attempt to balance the distribution, this strategy removes the data point of the bigger class when two points in the range that correspond to different classes are extremely close to one another.

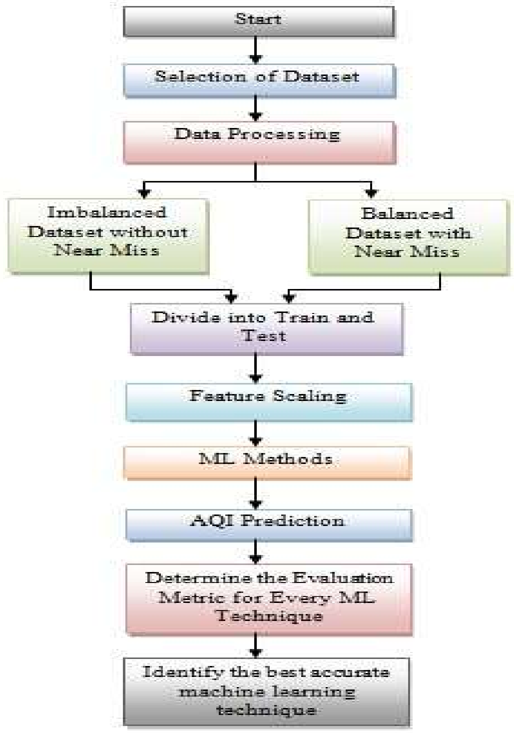

Figure 1: Proposed Methodology

B. Decision Tree Regression

Decision tree regression predicts future data by using an object's qualities to train a model inside the structure of a tree. This produces meaningful continuous output. Continuous output denotes a result or output that is not discrete, that is, not solely represented by a known, discrete collection of numbers or values. Based on a model of decisions and all possible outcomes, such as outputs, input costs, and utility, the Decision Tree is a tool for making decisions. It is designed to resemble a flowchart.

C. XG boost Regression

XGBoost is a useful tool for building supervised models for regression. The assertion will be supported by an awareness of its (XGBoost) goal function and base learners. The objective function consists of a loss function and a regularisation term. It makes clear the difference among actual and expected values, or how much the output of the model differs from the real data.

D. KNN Regression

A non-parametric ML method called K-Nearest Neighbours (KNN) can be used to address regression and classification problems. KNN is frequently referred to as "K-Nearest Neighbours Regression" or "KNN Regression" when talking about regression. This simple and easy to comprehend method produces predictions by identifying the majority class (for classification) or averaging the goal values of the K data points that are closest to an input (for numerical regression).

E. Dataset

The open Kaggle data collection of Indian air quality data was used for this investigation. The dataset contains hourly and daily data from many sites in different Indian cities regarding the quality of the air and the air quality index (AQI). The data spans from 2018 to 2024.The original dataset comprised all of India's cities in 38643 rows and 18 columns. The initial dataset, which only included data from one city New Delhi, for example has been selected and cleansed. The New Delhi dataset had 16 columns and 183 rows after cleaning. The sample dataset for New Delhi is shown in Tables 1.

TABLE I: DATASET FOR NEW DELHI

Date	CO	NO	NO_2	SO_2	$PM_{2.5}$	PM_{10}	NH_3	O_3	AQI	AQI_bucket
9/1/24	10.23	51.18	69.29	4.65	197.81	274.67	46.42	30.14	467	Severe
10/1/24	14.87	25.98	42.47	6.07	91.27	254.48	76.56	39.98	154	Moderate
11/1/24	9.65	13.67	43.37	5.25	173.47	226.31	143.63	27.56	298	Poor
12/1/24	11.73	20.43	44.55	3.13	167.98	265.56	137.23	32.07	320	Very poor
13/1/24	12.68	13.30	63.75	5.21	150.86	340.90	156.82	54.8	338	Very Poor
14/1/24	13.27	29.47	67.23	4.25	225.67	427.57	126.35	47.21	367	Very Poor
15/1/24	9.54	18.76	52.67	2.17	252.58	364.78	145.51	20.51	374	Very Poor

The original dataset's composition is unbalanced. A balanced dataset is created by using the synthetic minority oversampling technique (Near Miss) approach on the imbalanced dataset. This algorithm makes use of oversampling. For each class label in the dataset, an equal amount of rows, more rows, or less rows, added to any classes that have insufficient rows. A dataset that is unbalanced has asymmetry. A skewed class distribution is the result of an imbalanced dataset, and this has multiple effects on the accuracy of the model.
Thus, it is imperative to balance the facts. By oversampling the positive class label, the accuracy of the results can be increased. In this study, oversampling is done through the usage of Near Miss. By basing its estimation on the closest neighbours, the Near Miss approach increases the occurrence of the minority class or minority class group in the provided dataset. The imbalanced and balanced datasets for New Delhi are contrasted in Figure 2.

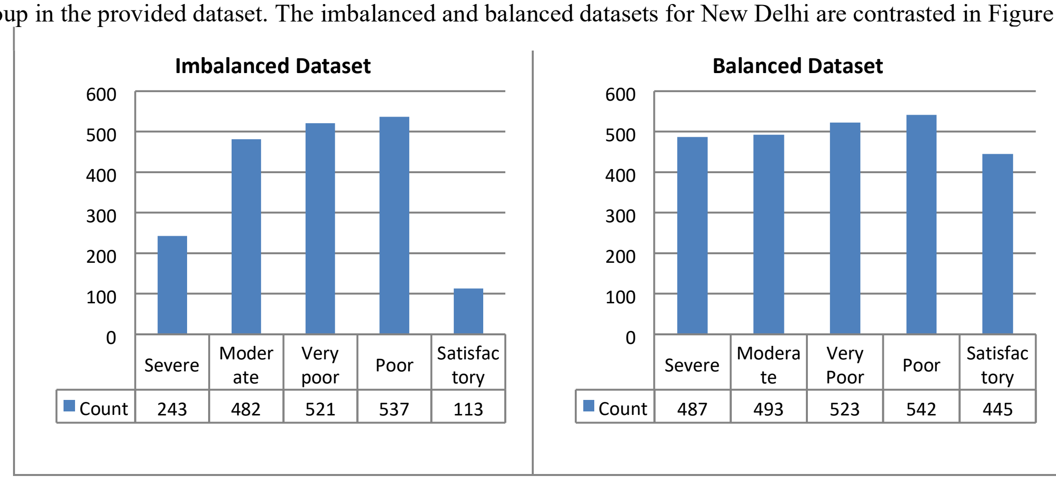

Figure 2: Imbalanced and Balanced Dataset for New Delhi

IV. RESULTS AND DISCUSSION

In the suggested strategy, the previously mentioned dataset was employed twice: once in an imbalanced type and once in a balanced type that made use of Near Miss. After being cleansed, the dataset only included values for the city of New Delhi. Plotting of the graphs revealed that the models using the balanced dataset had higher accuracy. Three algorithms Decision Tree regression, KNN regression, and XGBoost regression were applied to it in order to make predictions.
The study presented in this work indicates that the decision to employ statistical metrics—like R-SQUARE, RMSE and so forth—has been recognised and discussed in publications [15–18], along with the best ways to put them into practice. A model's performance is monitored and evaluated using metrics (during training and

testing). Since every algorithm used is based on a regression model; these metrics provide details on the degree of divergence from the actual values as well as the forecast' accuracy.
Table 2 presents a comparison of the accuracy of the dataset without and with the Near Miss method of New Delhi. The dataset with the Near Miss method obviously exhibits higher accuracy levels. It is also evident how radically New Delhi's precision has changed.
Figure 3 shows the accuracy of the AQI estimate in New Delhi using a bar graph with and without the Near Miss method, using different machine learning approaches like KNNR, DTR, and XGR.

TABLE II: COMPARISON OF THE DATASET'S ACCURACY VALUES USING AND WITHOUT NEW DELHI'S NEAR MISS METHOD

Methods	Accuracy
KNN	81.67
KNN-NM	89.96
DT	83.54
DT-NM	90.34
XG Boost	84.21
XG Boost-NM	93.04

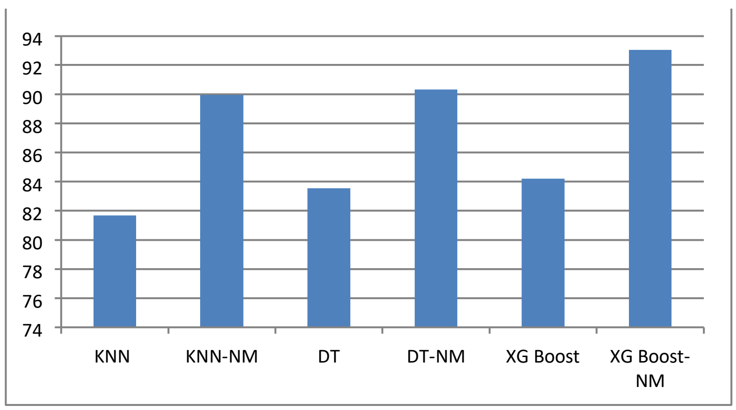

Figure 3: Accuracy comparison of different ML methods with and without Near Miss algorithm

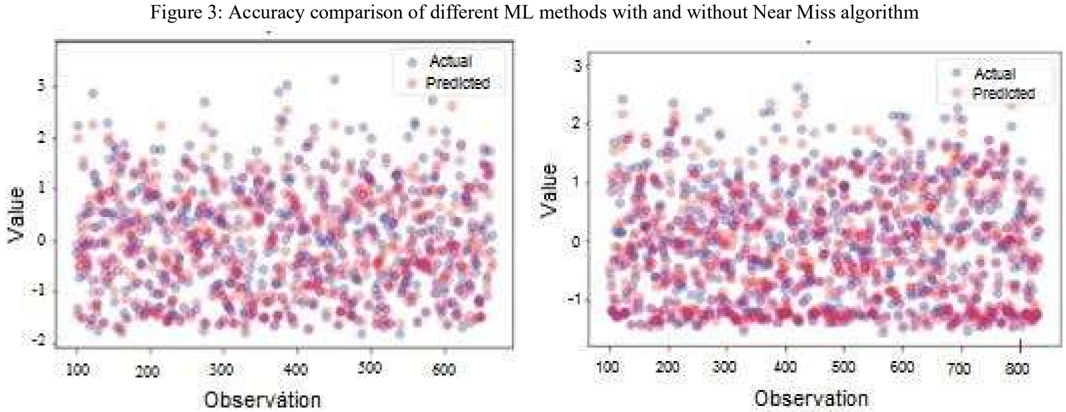

Figure 4: Scatter plot for New Delhi-KNNR

Figures 4 show the scatter plots for New Delhi and show the projected and actual values for two datasets: one that is balanced (by applying NM) and the other that is imbalanced (not applying NM) using KNN regressions.
The scatter plots for New Delhi are shown in Figures 5, which also show the predicted and actual values for a balanced dataset (with N)and an unbalanced dataset (without NM) employing DT regression.

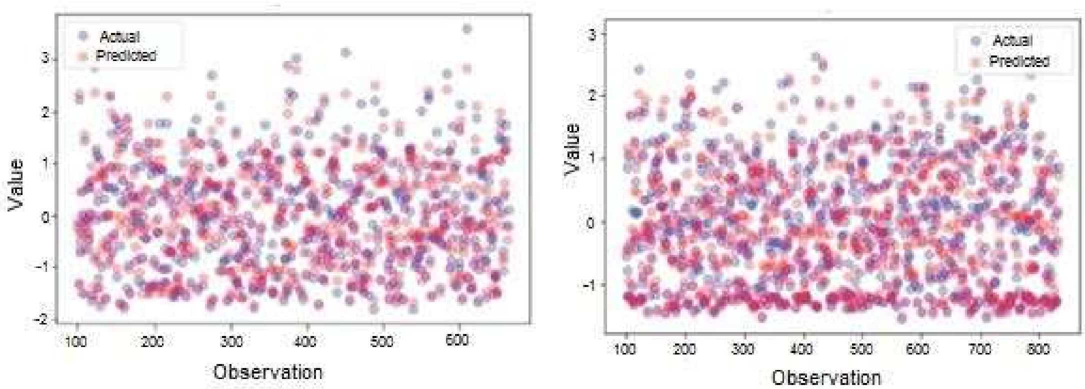

Figure 5: Scatter plot for New Delhi-DTR

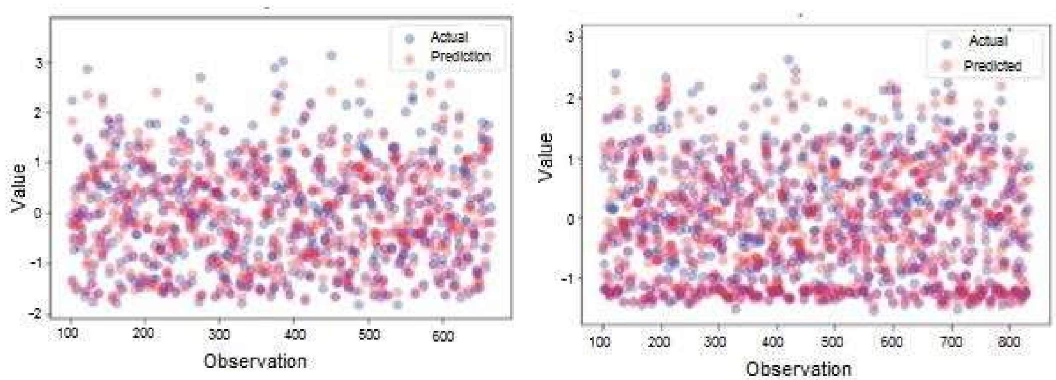

Figure 6: Scatter Plot for New Delhi-XGR

Scatter plots show the predicted values and actual for a balanced dataset with NM and an imbalanced dataset (without the usage of NM) employing XGBoost regression. Figures 6 show the scatter plots for New Delhi.

TABLE III: THE OUTCOME OF NEW DELHI CITY'S PERFORMANCE INDICATORS BOTH WITH AND WITHOUT THE NEAR MISS ALGORITHM

Algorithms	R-Square	MSE	RMSE	MAE
KNN	0.9256	0.0915	0.4124	0.3262
KNN-NM	0.9364	0.0893	0.3912	0.3158
DT	0.9376	0.0832	0.3957	0.3163
DT-NM	0.9452	0.0794	0.3738	0.3052
XGBoost	0.9413	0.0786	0.3873	0.3124
XGBoost-NM	0.9502	0.0753	0.3651	0.2968

Table 3 shows the results of performance measures for the imbalanced dataset for New Delhi City, both with and without the usage of the Near Miss method. These metrics include MSE, MAE, R-SQUARE and RMSE values for each of the three techniques—KNN, Decision Tree, and XGBoost regression.)When compared to DT and KNN regression, the XGBoost regression technique yields the best results.

The MSE, MAE, R-SQUARE and RMS of the KNN, Decision Tree, and XGBoost regression of the imbalanced dataset (i.e., without applying the Near Miss algorithm) and balanced dataset with the Near Miss technique are graphically compared in Figure 4. It demonstrates that the regression using XGBoost has the lowest values of RMSE, MSE, and MAE and the highest R-SQUARE.

According to the unbalanced dataset's results, XGBoost regression generates the highest accuracy (84.21%) and the minimum RMSE value (0.3873) in New Delhi. For balanced dataset with Near Miss algorithm XGBoost

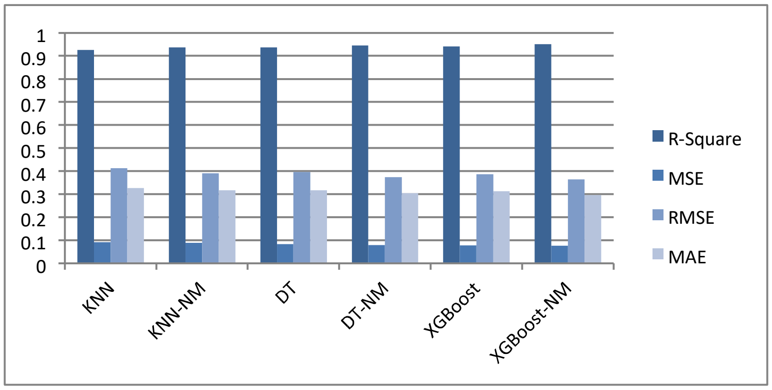

Figure 4: A comparison of the New Delhi unbalanced and balanced dataset's R-square, MSE, RMSE, and MAE machine learning approaches

generates the highest accuracy (93.04%) and the minimum RMSE value (0.3651) in New Delhi. The results of the investigation in Table 2 show that the XGBoost-NM model has a clear edge when it comes to air quality index prediction in New Delhi. By employing the NM unbalanced technique, it achieves the maximum classification performance of 93.04% accuracy for AQI prediction.

V. Conclusion

The global issue of air pollution is being studied by researchers from all across the world. To make an accurate prediction of the AQI, ML approach was examined. This research evaluate the efficacy of the three data mining algorithms (KNNR, DTR, and XGR) in accurately forecasting AQI data India's most populous and polluted city new Delhi. To achieve more precise and reliable results, the class data was equalized employing the synthetic minority oversampling technique (Near Miss). This distinctive strategy uses the balanced datasets and verifies that the balanced dataset is extremely accurate by comparing the findings of the balanced and imbalanced datasets sharply. Subsequently, statistical methods such as MSE, MAE, R-SQUARE, and RMSE are employed to verify that the improved outcomes were really effective in achieving increased precision.

According to the imbalanced dataset's results, XGBoost regression yields the highest accuracy (84.21%) and the minimum RMSE value (0.3873) in New Delhi. In New Delhi, XGBoost produces the lowest RMSE value (0.3651) and the highest accuracy (93.04%) for balanced datasets using the Near Miss algorithm. The investigation's findings demonstrate that, in terms of predicting New Delhi's air quality index, the XGBoostNM model is clearly superior. It reaches the highest classification performance of 93.04% accuracy for AQI prediction by using the NM imbalanced approach.

Hence, it seems that the XGBoost and Decision Tree algorithms can produce good results to predict air quality in the applied case of AQI in New Delhi when combined with Near Miss used datasets. This may force national and municipal governments, together with other civic organisations, to take action and regulate the quality of the air.

References

[1] K. Gu, J. Qiao, and X. Li, ”Highly efficient picture-based prediction of PM2.5 concentration,” IEEE Transactions on Industrial Electronics, vol. 66, no. 4, pp. 3176-3184, 2018.

[2] B. Zhai and J. Chen, ”Development of a stacked ensemble model for forecasting and analyzing daily average PM2.5 concentrations in Beijing, China,” Science of the Total Environment, vol. 635, pp. 644-658, 2018.

[3] S. C. Mangones, P. Jaramillo, N. Y. Rojas, and P. Fischbeck, ”Air pollution emission effects of changes in transport supply: the case of Bogota, Colombia,” Environmental Science and Pollution´ Research, vol. 27, pp. 35971-35978, 2020.

[4] World Health Organization, ”How air pollution is destroying our health,” 2019. [Online]. Available: https://www.who.int/airpollution/news-and-events/how-air-pollution-is-destroying-ourhealth.

[5] N. Van Mead, ”22 of world’s 30 most polluted cities are in India, Greenpeace says,” The Guardian, 2019. [Online]. Available: https://www.theguardian.com/cities/2019/mar/05/india-home-to22-of-worlds-30-most-polluted-citiesgreenpeace-says.

[6] K. Ramasamy, ”Enchanted improvements in air quality across India-A study from COVID-19 lockdown perspective,” Adalya Journal, vol. 9, no. 5, 2020.
[7] G. Oliveri Conti, B. Heibati, I. Kloog, M. Fiore, and M. Ferrante, ”A review of AirQ Models and their applications for forecasting the air pollution health outcomes,” Environmental Science and Pollution Research, vol. 24, pp. 64266445, 2017.
[8] X. Tan, L. Han, X. Zhang, W. Zhou, W. Li, and Y. Qian, ”A review of current air quality indexes and improvements under the multi-contaminant air pollution exposure,” Journal of Environmental Management, vol. 279, p. 111681, 2021.
[9] G. Mani and J. K. Viswanadhapalli, ”Prediction and forecasting of air quality index in Chennai using regression and ARIMA time series models,” Journal of Engineering Research, vol. 10, no. 2A, pp. 179-194, 2022.
[10] G. Soundari, J. G. Jeslin, and A. C. Akshaya, ”Indian air quality prediction and analysis using machine learning,” Int J Appl Eng Res, vol. 14, no. 11, pp. 181-186, 2019.
[11] M. Bansal, A. Aggarwal, T. Verma, and A. Sood, ”Air quality index prediction of Delhi using LSTM,” Int. J. Emerg. Trends Technol. Comput. Sci, vol. 8, no. 5, pp. 59-68, 2019.
[12] 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, p. 120825, 2020.
[13] S. V. Kottur and S. S. Mantha, ”An integrated model using Artificial Neural Network (ANN) and Kriging for forecasting air pollutants using meteorological data,” Int. J. Adv. Res. Comput. Commun. Eng, vol. 4, no. 1, pp. 146152, 2015.
[14] S. Halsana, ”Air quality prediction model using supervised machine learning algorithms,” International Journal of Scientific Research in Computer Science, Engineering and Information Technology, vol. 8, pp. 190-201, 2020.
[15] R. B. Abad, S. Mousavi, N. Mohamadian, D. A. Wood, H. Ghorbani, S. Davoodi, and K. Shahbazi, ”Hybrid machine learning algorithms to predict condensate viscosity in the near wellbore regions of gas condensate reservoirs,” Journal of Natural Gas Science and Engineering, vol. 95, p. 104210, 2021.
[16] Rajabi, M., Beheshtian, S., Davoodi, S., Ghorbani, H., Mohamadian, N., Radwan, A. E., & Alvar, M. A. (2021). Novel hybrid machine learning optimizer algorithms to prediction of fracture density by petrophysical data. Journal of Petroleum Exploration and Production Technology, 11(12), 4375-4397.