Grenze International Journal of Engineering and Technology, June Issue

A Data-Driven Approach to Air Quality Prediction in Jharkhand's Coalfields using Machine Learning

Kajal Kumari[1], Dr. Sudip Kumar Sahana[2] and Dr. Debjani Mustafi[3]
[1-3]Computer Science & Engineering, Birla Institute of Technology, Ranchi, India
Email: Ka123jal000@gmail.com, {sudipsahana@bitmesra.ac.in, debjani.mustafi@bitmesra.ac.in}

***Abstract*—Air pollution is a considerable issue to human health and requires prompt reduction in anthropogenic zones like industrial plants. Every day, harmful pollutants including CO, CO2, particulate matter, NO2, SO2, and NH3 are discharged into our environment. The mining industry stands out as a key source of air pollution, particularly due to the substantial generation of Particulate Matter. This leads to an unhealthy atmosphere for both mine workers and neighboring communities. This study focuses on monitoring air quality in Jharkhand, known as India's largest coal-producing state in India. The state hosts 38 opencast coal mines and 5 underground coal mines. Accurate prediction of air quality is crucial. Traditional measurement methods often yield imprecise results and require complex mathematical computations. A variety of supervised algorithms used for prediction logistic regression with 0.7482, decision tree classifier with 0.9989, random forest with 0.9989, and KNN with 0.9898 accuracy.**

***Index Terms*— Environmental Protection Agency (EPA), National Environment Protection Measure (NEPM), Support Vector Regression (SVR), Radial Basis Function (RBF), Random Forest Classification (RFC).**

I. INTRODUCTION

Air pollution poses a serious impact to human health and must be reduced in anthropogenic areas like industrial plants. Harmful pollutants such as CO, CO2, particulate matter, NO2, SO2, and NH3 are continuously being discharged into our environment on a daily basis. The primary industrial plant which involves in air pollution is mining industry, where a substantial amount of Particulate Matter is generated. This eventually leads to an unhealthy environment for both the mine workers and the surrounding communities. Our dataset pertains to the monitoring of air quality in Jharkhand, a state renowned as the largest coal-producing region in India. Here, there are 38 opencast coal mines and 5 underground coal mines. The mining operations involve on-site drilling, blasting, coal cleaning, and transportation, all of which result in the production of hazardous air pollutants. Accurately predicting air quality is essential. Traditional methods for measuring it often yield less precise results and require extensive mathematical calculations.

II. LITERATURE REVIEW

In [1] aims on forecasting the hourly air quality in the state of California. The findings reveal that using SVR with the RBF kernel enables accurate predictions of pollutant concentrations. Additionally, the classification into six Air Quality Index (AQI) categories achieved an impressive accuracy of 94%.

Grenze ID: 01.GIJET.10.2.622_2

In [2], utilizing an 11-year dataset gathered by Taiwan's Environmental Protection Administration, the study achieved promising results. Stacking ensemble and AdaBoost emerged as the top performers, offering a remarkable 94% accuracy in predicting the target variable across three datasets. The Root Mean Squared Error (RMSE) with imputation stood at 22.618, Mean Absolute Error (MAE) at 16.105, and R-squared (R2) at 0.735.
In [3], this study, a machine learning approach was employed to forecast PM2.5 levels utilizing a dataset retrieved from the UCI repository. The dataset comprises five key features: temperature, wind speed, dewpoint, pressure, and the target variable, PM2.5. The chosen model for this predictive task was logistic regression. The model demonstrated exceptional performance, with a mean accuracy of 0.9988 and a very low standard deviation of 0.000612. This outcome suggests that the logistic regression model accurately predicts PM2.5 levels based on the atmospheric conditions provided in the dataset. Further analysis could involve exploring feature importance, evaluating additional performance metrics, and considering methods such as cross-validation and hyperparameter tuning to enhance the model's capabilities.
In [4], this study aimed to assess the air quality index using a dataset sourced from Kaggle, which includes three primary features: temperature, wind speed, and humidity. Various machine learning algorithms were applied, namely linear regression, decision tree, random forest, artificial neural network, and support vector machine. Among these models, the highest accuracy was achieved with the neural network and boosting models. These results suggest that the neural network and boosting algorithms exhibit superior performance in predicting the air quality index based on the provided dataset's atmospheric conditions. Further analysis may include exploring feature importance, assessing additional performance metrics, and optimizing the chosen models for even better results.
In [5], this study presents a novel comparative analysis of air quality index monitoring utilizing various machine learning techniques. The dataset, obtained from Kaggle, comprises five primary features: SO2, NO2, RSPM, SPM, and PM2.5. Before applying the Synthetic Minority Over-sampling Technique (SMOTE), the CatBoost regression model achieved an accuracy of 79.86% for New Delhi and 68.68% for Bangalore. Upon implementing SMOTE to address class imbalance, the accuracy improved significantly. Specifically, the accuracy for New Delhi increased to 85.084%, while Bangalore saw a notable increase to 90.30%. These results highlight the effectiveness of SMOTE in enhancing the predictive performance of the CatBoost regression model.

III. Machine Learning Algorithms

A. Logistic Regression

Logistic Regression is an algorithm designed for binary classification tasks, where the outcome variable y has two possible values 0 and 1. The logistic regression model estimates the probability that a given input x belongs to the positive class. The mathematical formula is:

$$P(y=1|\ x) = \sigma(\beta 0+\beta 1x) \quad (1)$$

In"1", The value of x is a series of x_n. $P(y = 1|x1, x2,\ldots,xn)=\sigma(\beta 0+\beta 1x1+\beta 2x2+\ldots+\beta nxn)$ β_0, β_1, ..., β_n are the coefficients of logistic regression.
x_1, x_2, ..., x_n are the input features.

B. Decision Tree Classifier

A Decision Tree classifier is a type of supervised learning algorithm that is non-parametric and commonly used for solving classification problems. It operates by partitioning the data into subsets based on the values of its features. This process is repeated recursively on each subset, creating a tree-like structure where the nodes represent the features, the branches represent the decisions based on those features, and the leaves represent the class labels. The mathematical essence of Random Forest lies in how these Decision Trees are aggregated to make predictions.

$$\text{input sample } x = [x_1, x_2, \ldots, x_n] \quad (2)$$

In "2", Starting at the root node j = 0: If x $X_0 \leq t_0$, go to the left branch: j_{left} - left_branch(0), If x $X_0 > t_0$,go to the right branch: j_{right} – right_branch(0).
And, at each subsequent node j: If x $X_j \leq t_j$, go to the left branch: j_{left} – left_branch(j). If x $X_j > t_j$, go to the right branch: : j_{right} – right_ branch(j) and repeat the process until reaching a leaf node L. The predicted class $y^{\wedge}$ for the input X is the majority class of the samples in leaf node L

C. Random Forest Classifier

A Random Forest is an ensemble learning technique that works by creating numerous decision trees during training and then providing the mode of the classes (for classification) or the average prediction (for regression) of the individual trees as the output.

$$RF: \hat{y} - \text{mode}(\{T_1(X), T_2(X)\ldots, T_N(X)\}) \quad (3)$$

In "3", N: The total number of Decision Trees in The Random Forest. T_i (X):Prediction of the ith Decision Tree for the input X. Mode ($\{T_1(X), T_2(X)\ldots, T_N(X)\}$): Mode function that selects the most frequently occurring Prediction among all Trees for input X.

D. K-Nearest Neighbors

K-nearest neighbors algorithm is based on classifies data point along with how it is similar to their neighbor. KNN is best model for the pattern recognition tasks on different features. It uses the features and labels of training data to predict the classification of unlabeled data. It is useful in prediction of disease.

Consider, test point as ***x***.

set of the *k* nearest neighbors of *x* as *Sx*.

Formally, *Sx* is defined as , $Sx \subseteq D$ where *D* is a subset such that: $|Sx|=k$

For all (*x'*, *y'*) in *D* except those in S_x, the distance between *x* and *x'* is greater than or equal to the maximum distance between *x* and any point *x''* in *Sx*.

The predicted label for *x*, denoted as *h (x)*, is determined as the mode of the set of labels $\{y'' : (x'',y'') \in S_x\}$.

IV. Proposed Methodology

This research utilizes a dataset sourced from Kaggle titled 'Air Quality Index', a trusted repository often utilized for research purposes. Focused on studying the air quality of Jharkhand, specific data from the Jharkhand State Pollution Control Board years (2004-2015) was extracted. The sourcing and extraction of high-quality data from air quality datasets is a critical step in ensuring the accuracy and reliability of the final research findings.

A. Proposed model

Air quality testing devices will be installed in various locations near coal mines to gather information on the composition of different gases in the air. For real-time monitoring of coal miners' health, wearable sensors such as LM35, BP-BMP180, and ECG – AD8232 will be utilized. In addition, environmental sensors like DHT22, MQ2, MQ5, MQ7, MQ135, DS18B20, as well as sound and audio detection sensors, will be placed near the mines to capture data related to temperature, sound levels,and other environmental factors. In "Figure 1" the flowchart of proposed approach is given,

B. Data Collection

This research utilizes a dataset sourced from Kaggle titled 'Air Quality Index', a trusted repository often utilized for research purposes. Focused on studying the air quality of Jharkhand, specific data from the Jharkhand State Pollution Control Board years(2004-2015) was extracted. The sourcing and extraction of high-quality data from air quality datasets is a critical step in ensuring the accuracy and reliability of the final research findings.

C. Data Visualization

The count of each type present in a dataset is essential for understanding the data distribution, selecting appropriate evaluation metrics, improving model performance, and making informed decisions throughout the machine learning pipeline. It ensures that the model learns from and represents the underlying data accurately.

In "Figure 2", The bar graph illustrates the count of categories such as Residential, Rural and Other Areas, and Industrial zones, which contribute to the emission of air pollutants.

Pair plot is useful in exploratory data analysis to understand the relationship between pairs of variables in dataset. So, In "Figure 3" dataset parameters are plotted.

D. Data Pre-processing

The dataset comprises nine significant parameters: location, agency, type (Residential/Industrial), SO2, NO2, RSPM, SPM, PM 2.5, and date. The processing stage plays a crucial role in data analysis as it enhances data quality. During this phase, the Air Quality Index (AQI) has been computed using the most influential parameters from the dataset. Subsequently, air samples have been categorized based on their AQI values.

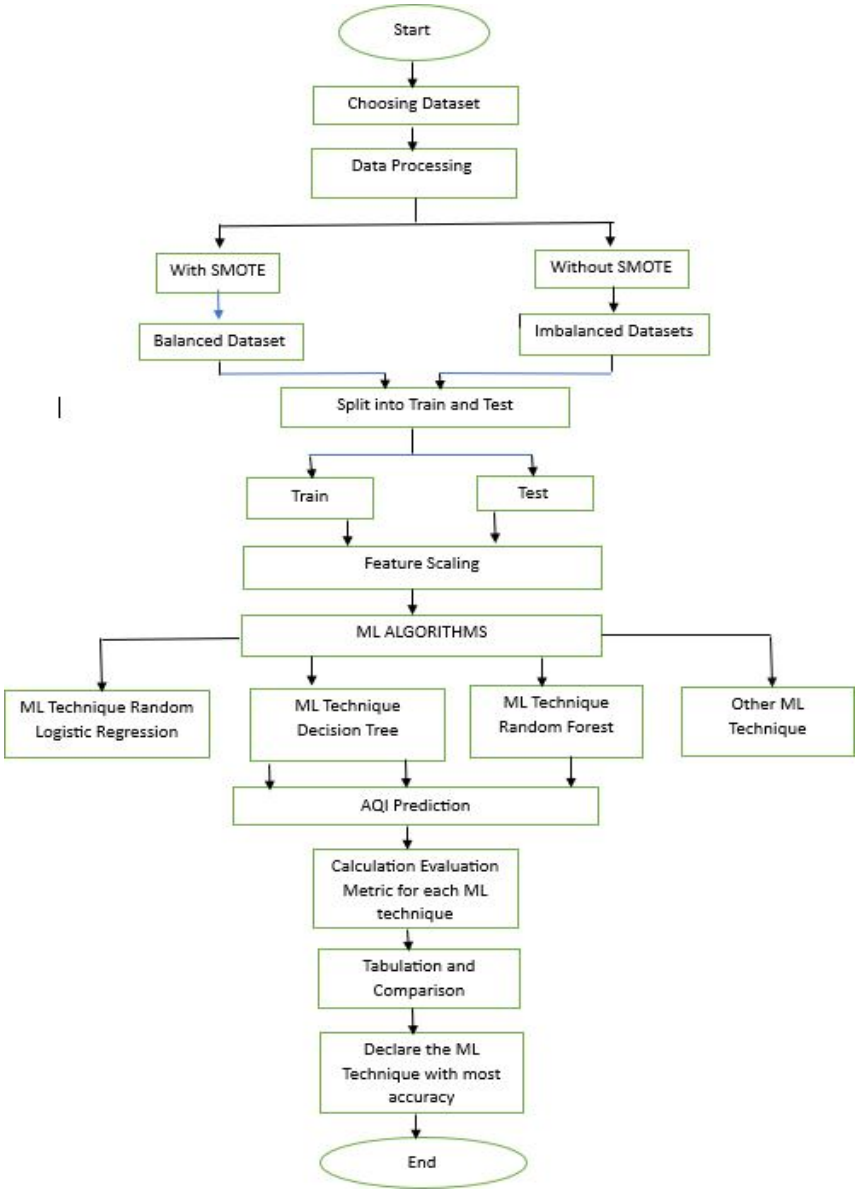

Figure 1: Workflow of proposed model

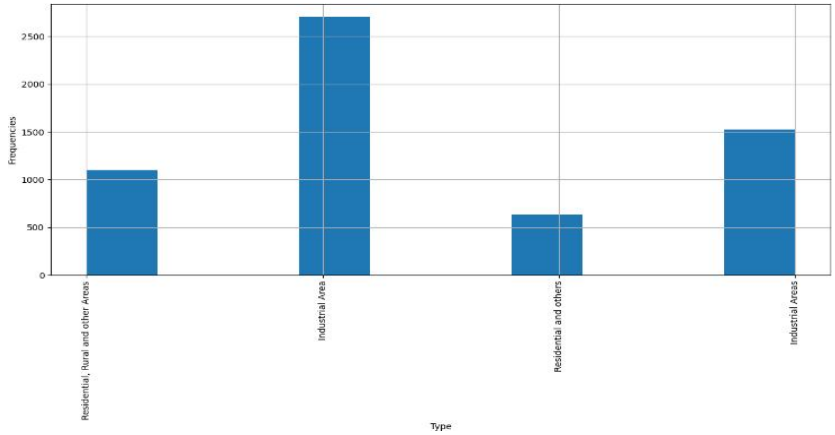

Figure 2: Bar graph for count of type present in dataset

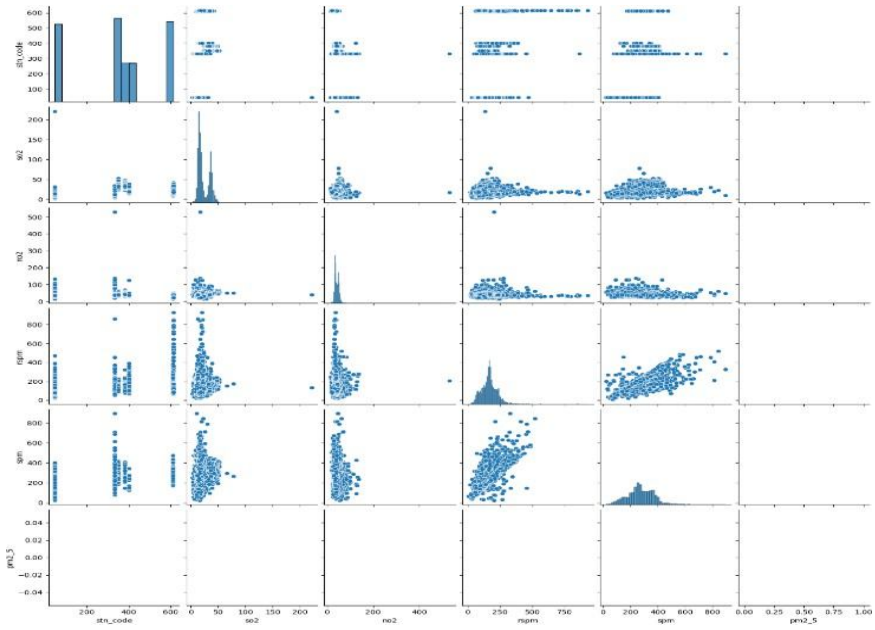

Figure 3: Pair plot of six features

E. WQI Calculation

The Air Quality Index (AQI) is determined by the concentration of pollutants present in the air. Each pollutant's AQI value is represented as a percentage of the level set by the National Environment Protection Measure for Ambient Air (NEPM) standard. Six specific pollutants are monitored to calculate the AQI. These pollutants include PM10, PM2.5, nitrogen dioxide, sulphur dioxide, carbon monoxide. The mathematical formula is,

$$Ip = [\ IHi\text{-}ILo \mid BPHi - BPLo] (Cp – BPLo) + ILo \qquad (4)$$

In “4”, Ip =index of pollutant p
Cp =truncated concentration of pollutant p
BPHi = concentration breakpoint i.e greater than or equal to Cp The AQI Values from 0 -300 is listed in “Table II”.

TABLE II: SIX AIR QUALITY INDEX (AQI) CATEGORIES DEFINED BY THE EPA

AQI Values	**Level**	**Daily AQI Colour**	**Description**
0-50	Good	Green	Satisfactory and air pollution poses little or no risk.
51-100	Moderate	Yellow	Acceptable
101-150	Unhealthy for sensitive group	Orange	Members of sensitive group may experience health effects.
151-200	Unhealthy	Red	More serious health effects.
201-300	Very unhealthy	Purple	Health alert: The risk of health effects is increased for everyone.
301 and Higher	***Hazardous***	**Maroon**	**Health warning of emergency conditions: Everyone is more likely to be affected.**

V. RESULT AND DISCUSSION

The graphical representation displays in figure 4, first four measured values of sulphur dioxide (SOi), nitrogen dioxide (NOi), Respirable Particulate Matter (RPi), and Suspended Particulate Matter (SPM), along with their corresponding Air Quality Index (AQI) values.

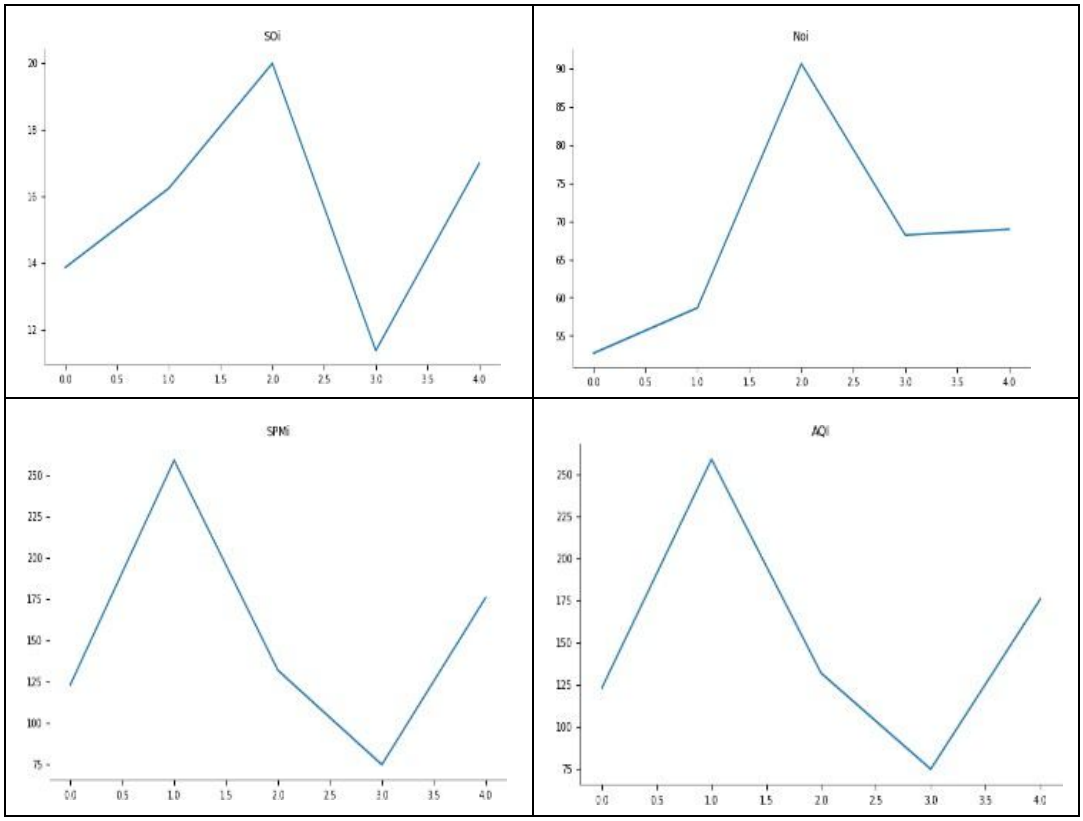

Figure 4: The graphical representation of features

In this study, Cohen's Kappa coefficient (kappa score) is used for imbalanced class distributions. 1 signifies perfect agreement between predicted and actual classes, while 0 shows agreement equivalent to chance. In Figure 5, The result is displaying with Random Forest algorithm accuracy 0.9989 with kappa score of 0.9986.

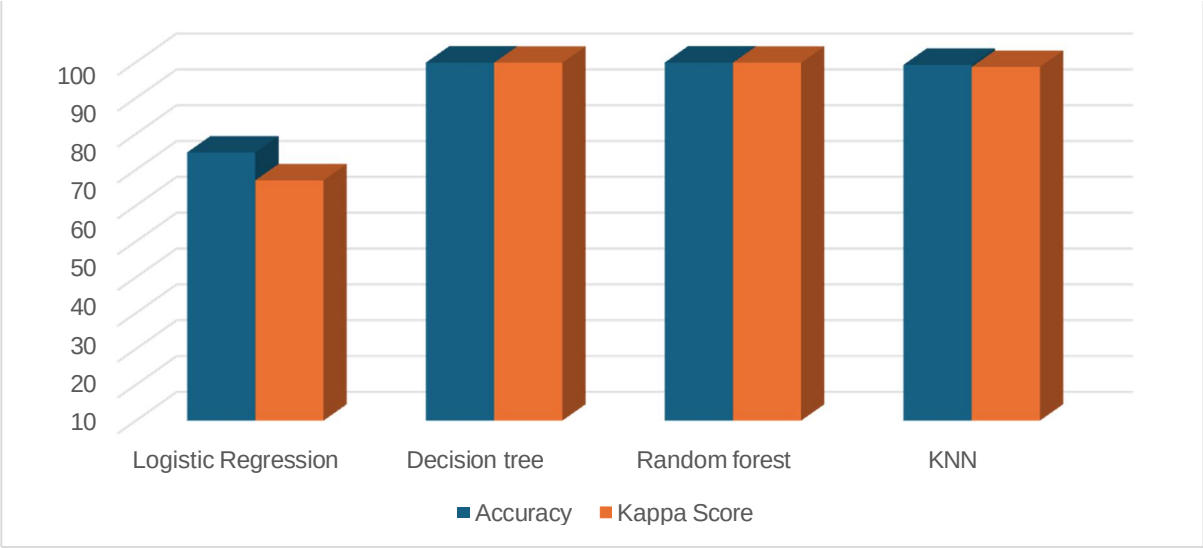

Figure 5: Comparison of different machine learning algorithms

VI. Conclusion

In this study, Logistic regression, decision tree, random forest, and KNN were employed, ensuring a high level of accuracy with a focus on the kappa score. The analysis was conducted on a Kaggle dataset, and graphical

representations were provided for a comparative analysis of all applied machine learning algorithms, along with their respective kappa scores. The dataset utilized in this study included significant features such as SO2, NO2, RSPM, and SPM. However, it is noteworthy that the PM2.5 values in the Jharkhand state were predominantly missing, indicated by NaN (Not a Number) values. This observation underscores the challenge of missing data in certain features within the dataset, particularly in the context of PM2.5 levels for Jharkhand state. It highlights the importance of handling missing data effectively through techniques such as imputation or model adjustments to ensure robust and reliable analysis results. This study also analyses the air quality surrounding open-cast coal mines by deploying sensors to measure key air quality parameters. These sensors will transmit real-time data to our monitoring system. Once we've gathered sufficient data, we will apply machine learning techniques to derive insights and investigate the correlation between Air Quality Index (AQI) values and the prevalence of chronic health conditions among coal miners and their local communities.

ACKNOWLEDGMENT

This research work is part of the project supported by Central Coalfield Limited under the CSR scheme.

REFERENCES

[1] M. Castelli, F. M. Clemente, A. Popovič, S. Silva, and L. Vanneschi, "A machine learning approach to predict air quality in California," Complexity, 2020.
[2] Y. C. Liang, Y. Maimury, A. H. L. Chen, and J. R. C. Juarez, "Machine learning-based prediction of air quality," Applied Science, vol. 10, no. 24, pp. 9151, 2020.
[3] C. R. Aditya, C. R. Deshmukh, D. K. Nayana, and P. G. Vidyavastu, "Detection and prediction of air pollution using machine learning models," International Journal of Engineering Trends and Technology(IJETT), vol. 59, no. 4, pp. 204-207, 2018
[4] T. Madan, S. Sagar, and D. Virmani, "Air quality prediction using machine learning algorithms–a review," in 2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), pp. 140-145, IEEE, Dec. 2020
[5] N. S. Gupta, Y. Mohta, K. Heda, R. Armaan, B. Valarmathi, and G. Arulkumaran, "Prediction of air quality index using machine learning techniques: a comparative analysis," Journal of Environment and Public Health, pp. 1-26, 2023
[6] P. Bhalgat, S. Pitale, and S. Bhoite, "Air quality prediction using machine learning algorithms,"International Journal of Computer Applications Technology and Research, vol. 8, no. 9, pp. 367-370, 2019.
[7] K. Kumar and B. P. Pande, "Air pollution prediction with machine learning: a case study of Indian cities,"International Journal of Environment Science and Technology, vol. 20, no. 5, pp. 5333-5348, 2023.