

A Novel Air Quality Labelling and Classification Approach Using Supervised Learning

*Towqir Ahmed Shaem¹, Jeba Shahana Azad², Aonmoy Das³, Jannatul Arifa Sweetty⁴

^{1, 3, 4} *Electrical and Electronic Engineering, University of Chittagong*

² *Materials Science and Engineering, Rajshahi University of Engineering and Technology (RUET)*

*Corresponding Author

DOI : <https://doi.org/10.51583/IJLTEMAS.2025.140300061>

Received: 16 March 2025; Accepted: 27 March 2025; Published: 21 April 2025

Abstract: Air quality assessment is crucial for environmental monitoring, public health and decision-making. Air quality does not depend solely on the concentration of certain gasses; rather, it is also influenced by other pollutants that are challenging to measure individually for further investigations. In this study, we used the UCI Air Quality dataset, which included key pollutants such as CO, NO₂, NO_x, benzene, temperature, humidity and sensor data for other forms of pollutants. We proposed a novel labeling scheme based on weighted pollutant concentrations, enabling more precise air quality classification into Good, Moderate, and Unhealthy categories. After that, we evaluated five supervised learning models—Random Forest, Decision Tree, Support Vector Machine, K-Nearest Neighbors, and Gradient Boosting—for classification, considering all types of measured pollutants, and assessed their performance using accuracy, confusion matrices, classification reports, and ROC-AUC curves. Our research also highlights the potential of AI-driven techniques in comprehensive air quality assessment as well as real-time air pollution prediction and classification for environmental protection.

Keywords: Air quality index, Supervised learning, Labelling, Weights, Key Pollutants, ML models.

I. Introduction

Air pollution is becoming a major concern in the modern world, significantly affecting human health as well as overall quality of life. Presence of harmful substances, gases and biological molecules such as carbon monoxide, nitrogen oxides (NO_x), sulfur dioxide (SO₂), benzene (C₆H₆) and particulate matter are responsible for polluting the air and environment. Industrial wastes or emissions, vehicular emissions and natural disasters like wildfires and volcanic eruptions are the potential sources of air pollution. Moreover, due to rapid industrialization and unplanned urbanization, air pollution has been increasing at an alarming rate.

As a result, it not only harms the surroundings but also increases health risks for humans. According to many research studies, respiratory diseases, cardiovascular conditions, and premature deaths are becoming common due to excessive air pollution [1]. For example, people are most likely to suffer from chronic bronchitis, lung cancer, and heart attack due to prolonged exposure to some particular pollutants like NO_x and CO [2]. Not only that, air pollution can also exacerbate pre-existing health diseases such as asthma and other respiratory disorders. Alongside these, air pollution is becoming a potential threat to the environment and wildlife by contributing to acid rain, damaging ecosystems and reducing biodiversity as well.

Therefore, it is crucial to predict the air quality levels accurately so that governments and policymakers can take preventive measures. However, traditional methods of data collection mostly rely on sensor based techniques and satellites data, which requires high costs while lacking real time adaptability. To address these challenges, researchers are exploring AI-driven or machine learning (ML) based techniques for air quality monitoring. Mostly, they collect the data from IoT based sensors, which lack proper data labelling and target values for supervised learning. Also many people use their own threshold sensors values to create air quality indexes/labels. Besides, different manufacturers maintain varying measurement units, making it difficult to establish universal threshold values [3]. As a result, many researchers need to work with unsupervised learning to identify patterns.

Not only that, IoT based sensor data often exhibit inconsistencies such as sensor noise, calibration issues, and nonlinear pollutant behavior [4]. Moreover, air quality data from IoT and electronic sensors may exhibit variability due to environmental factors like temperature, humidity and pressure fluctuations [5]. To address these challenges, we calculated auxiliary data to create target features for supervised learning. These auxiliary data are derived from historical trends and numerous meteorological conditions. Besides, it will help us to reduce sensor variability, while improving classification accuracy [6].

In our work, we used standard parameters for data labelling [7]. After that, ML models are crucial to learning complex data patterns beyond classical threshold-based rules. Also real world data is often noisy and fixed thresholds may become outdated over time. ML models can handle these ambiguities better and can be fine-tuned for better performance as well. To provide comprehensive insights into air quality trends we also employed five distinct classification models, including random forest, support vector machine (SVM), decision tree, K-nearest neighbors and gradient boosting. Regarding performance, we evaluated their accuracy, precision, recall, confusion metrics and receiver operating characteristic (ROC) curves.

II. Related Works

Researchers have been using various statistical, machine learning and deep learning models to enhance air quality forecasting accuracy. In this section, we are going to review significant studies that have contributed to this domain, highlighting their methodologies, findings and limitations.

Traditional Statistical Approaches

In 1976, Box and Jenkins introduced the autoregressive integrated moving average (ARIMA) model. It had been widely used for time-series forecasting of air pollution levels [8]. Since then, early research on air pollution prediction relied heavily on statistical models such as ARIMA. As ARIMA assumes linear relations between air quality factors, it has lower efficiency while calculating complex interactions. Later, Seinfeld and Pandis utilized MLR to establish relationships between meteorological variables and pollutant concentrations in 1998 [9]. Though MLR can provide interpretable results, it lacks adaptability to non-linear patterns for high-dimensional air quality data.

Machine Learning-Based Approaches

Due to the advancement of semiconductor technology, researchers are getting access to high computational devices, allowing them to train machine learning models for air pollution forecasting. At the same time, edgeAI and TinyML are becoming popular for low power IoT devices. For instance, Jiang et al. (2017) used meteorological and pollution data and applied Random Forest and SVM to predict PM_{2.5} concentrations in Beijing [10]. They proposed that unlike traditional statistical methods, machine learning based algorithms can perform better due to the ability to interpret complex correlations between different features. Similarly, Li et al. implemented Gradient Boosting Machines (GBM) for air pollution classification. He claimed higher prediction accuracy for GBM as compared to ARIMA [11].

However, researchers are now able to unlock convoluted air quality patterns by using deep learning strategies. In 2020, Zhang et al. developed a Long Short-Term Memory (LSTM)-based air pollution prediction model. It leveraged historical pollutant data and meteorological factors to forecast PM_{2.5} levels [12]. According to their determinations, LSTM networks outperformed traditional ML models due to their ability to retain long-term dependencies. Therefore, Recurrent Neural Networks (RNNs) and LSTM networks are being widely used for time-series forecasting for high computational devices. Due to their higher computational cost, deep learning models cannot yet replace machine learning models for daily use.

Researchers are now planning to work with hybrid models to enhance efficiency, combining statistical and machine learning techniques, specially designed for low power devices. Additionally, it can increase accuracy and efficiency while consuming low power and can be implemented in embedded systems and wearable devices. In 2021, Feng et al. integrated ARIMA with deep learning models to enhance predictive accuracy [13]. Results highlighted that hybrid models can outperform previous methods. In 2022, an IoT-enabled smart monitoring air quality system was introduced by Kumar et al. He utilized ML algorithms to predict pollution levels based on real-time sensor data [14].

Despite significant progress, challenges remain in air pollution forecasting. Though many studies have focused on specific pollutants (eg, PM_{2.5}, NO₂) but have failed to provide a comprehensive air quality classification framework. Moreover, the scalability of ML model across different geographical areas remains a concern. Our research addresses these research gaps by introducing robust classification models incorporating multiple pollutants parameters and creating labels of three different classes while ensuing broader adaptability.

III. Methodology

Dataset

Authentic data is crucial to maintaining accuracy while training ML models. In our work, we used the dataset of the UC Irvine Machine Learning Repository, a reputable provider of authentic research data [15]. It contains real time sensor values and auxiliary data of key pollutants collected from an air pollution monitoring station in a town in Italy. Our dataset has 9471 records and 15 features. However, it has some missing values in some features, which were handled during the preprocessing stage. The key variables in the dataset include:

CO (GT): Concentration of Carbon Monoxide (mg/m³)

NO₂ (GT): Concentration of Nitrogen Dioxide (µg/m³)

NO_x (GT): Concentration of Nitric Oxide and Nitrogen Dioxide (µg/m³)

C₆H₆ (GT): Benzene concentration (µg/m³)

NMHC (GT): Non Metanic Hydrocarbons concentration (µg/m³)

T: Temperature (°C)

RH: Relative Humidity (%)

AH: Absolute Humidity (g/m³)

PT08.S1 (CO) : Sensors reading of CO

PT08.S2 (NMHC): Sensors reading of NMHC

PT08.S3 (NO_x) : Sensors reading of NO_x

PT08.S4 (NO₂) : Sensors reading of NO₂

PT08.S5 (O₃) : Sensors reading of O₃

Date & Time: Timestamps of data collection (Removed during preprocessing)

The dataset provides hourly measurements over a period, allowing for the study of air pollution variations over time.

Data Preprocessing

A) Handling null values

Having null values in a dataset might impact the accuracy of ML models. Initially, we analyzed our entire dataset to identify the missing value of the features. It has 114 missing values across various columns. In that case, the traditional way of handling null values is by replacing the null values with the mean or median of the entire column. In most cases, it is acceptable if the values are equally correlated. However, in our dataset, all the features are independent variables. Since null values in our dataset can lead to inconsistencies in training an ML model, we decided to do a complete case analysis and removed the rows with missing values. This decision was based on ensuring data integrity and preventing synthetic information. After removing the missing values, our dataset contained 9357 total records for training and evaluation.

B) Air Quality Labeling

Unlike conventional air quality assessment ML models, we introduced a novel scheme of custom air quality labelling. Air quality dataset might contain some key pollutants data along with sensors reading for particular gas. We cannot directly score weighted values based solely on sensor data because sensor measurements are prone to inaccuracies due to environmental factors, calibration errors, and sensor drift [16]. Additionally, air quality indicators often rely on integrating multiple pollutant concentrations and meteorological variables. Relying solely on raw sensor data can lead to biased or inconsistent scoring. Instead, incorporating auxiliary data, such as meteorological variables and pollutant ratios, enhances robustness and accuracy [17]. That's why we assigned weighted score based on pollution severity and thresholds value derived from environmental safety standards.

Table 1: Pollutant Concentration Ranges and Their Assigned Score Values

Individual Score Value	CO (ppm)	NO ₂ (ppb)	Benzene (µg/m ³)	NO _x (ppb)	Temperature (°C)	Humidity (%)
0	≤5	≤100	≤8	≤75	0 – 35	20 - 80
1	5 – 10	100 – 200	8 – 16	75 – 150	<0 or >35	<20 or >80
2	>10	>200	>16	>150		

Table 2: Air Quality Index Classification Based on Total Score

Total Score for Each Instance	Air Quality Index/Level
≤2	Good
3	Moderate
≥4	Unhealthy

Table 1, presents the standard threshold values of key pollutants. We have three classes (Good, Moderate and Unhealthy) and their corresponding score ranges in Table 2. We calculated score value for each key pollutants and calculated total score to identify air quality label for each record. Categorizing air quality levels based on the score range is essential for any machine learning model. It enables a strategic classification approach. Our labelling scheme will aid in real-time application by allowing authorities to assess pollution effectively.

C) Feature Selection

Feature correlation is known as the relationship between different features in a dataset. Understanding this correlation between different variables is crucial for feature selection.

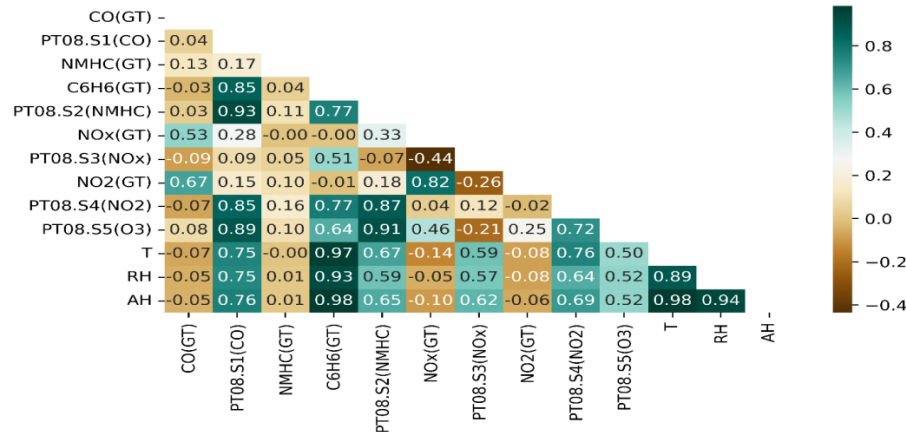


Figure 1: Feature Correlation Matrix

Figure 1, is the visual representation of the correlation of our dataset, where each cell in the matrix contains the Pearson correlation coefficient. The color gradient further helps in identifying patterns. The darker shades there is, the stronger correlations there will be. For example, CO (GT) has a strong positive correlation (0.67) with NO₂ (GT). On the other hand, temperature (t) and relative humidity (RH) both has a strong negative correlation (-0.89).

D) Encoding

Machine learning models require numerical values to process the data effectively. Therefore, encoding categorical variables is crucial to training any model. Models may misinterpret categorical variables due to the inappropriate encoding, resulting in performance reduction [18]. As we have three ordinal classes in the air quality level or index, we used label encoding to encode them. Therefore, good, moderate and unhealthy classes were converted to class 0, class 1 and class 2 respectively.

E) Scaling

We used Min-Max scaling techniques to preprocess our data. Basically, It's a normalization technique used to transform numerical data into a fixed range, typically [0, 1]. The formula is

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

X is the original value, Xmin and Xmax is the minimum and maximum value in the feature, respectively. Using min max scaling helps maintain feature relationships and prevent dominance by large values while improving model performance.

F) Train Test Split

Train-test splitting is important to evaluate a machine learning model. The training set is used to train the model and the testing set is for evaluating the model. Studies suggest that an 80-20% or 70-30% split is a common practice [19]. This technique prevents overfitting and helps to perform effectively. As a result, we used a 70-30% split before training the model.

G) Supervised Learning Models

Our study implements five different supervised learning models to classify air quality index which was created through a novel scheme. We used Random Forest, Decision Tree, Support Vector Machine (SVM), K-Nearest Neighbors (KNN) and lastly Gradient Boosting. We will evaluate the performance of our models while classifying air quality levels. Fig. 2 depicts the workflow of our air quality prediction approaches.

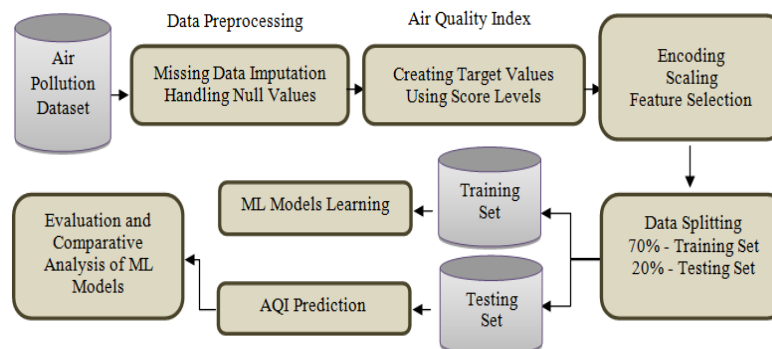


Figure 2: Flowchart of Our Methodology

IV. Results and Discussion

We evaluated our models in terms of four different metrics: Accuracy, Precision, Recall and F1 Score. Below are the formulas for those metrics:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

$$Precision = \frac{TP}{TP+FP}$$

$$Recall = \frac{TP}{TP+FN}$$

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision+Recall}$$

TP (True Positive), TN (True Negative), FP (False Positive), and FN (False Negative) represent classification outcomes. Table 3 indicates the performance evaluation of five different supervised learning models—Random Forest, Decision Tree, Support Vector Machine (SVM), K Nearest Neighbors (KNN) and Gradient Boosting. Overall, Random Forest, Decision Tree and Gradient Boosting achieved the highest scores on all metrics, indicating their ability to classify air pollution data correctly. However, KNN and SVM showed decent performance having accuracy around 0.92, making them moderately effective model.

Table 3: Performance Metrics of Machine Learning Models

Models	Accuracy	Precision	Recall	F1-Score
Random Forest	0.99	0.99	0.99	0.99
Decision Tree	0.99	0.99	0.99	0.99
Support Vector Machine	0.90	0.91	0.90	0.91
K-Nearest Neighbors	0.92	0.93	0.92	0.92
Gradient Boosting	0.99	0.99	0.99	0.98

Fig.3 depicts the confusion matrices of five different models—Random Forest, Decision Tree, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Gradient Boosting—evaluated on the testing dataset. Gradient Boosting, Random Forest and Decision Tree had the lowest misclassifications while predicting multiple classes or labels, demonstrating near-perfect classification, indicating that tree-based models performed significantly better. However, SVM and KNN struggled, particularly misclassifying moderate and unhealthy categories, leading to more misclassifications. The novel labelling scheme contributed to improving classification performance, providing clearer class separations and making supervised learning a powerful tool for air quality prediction.

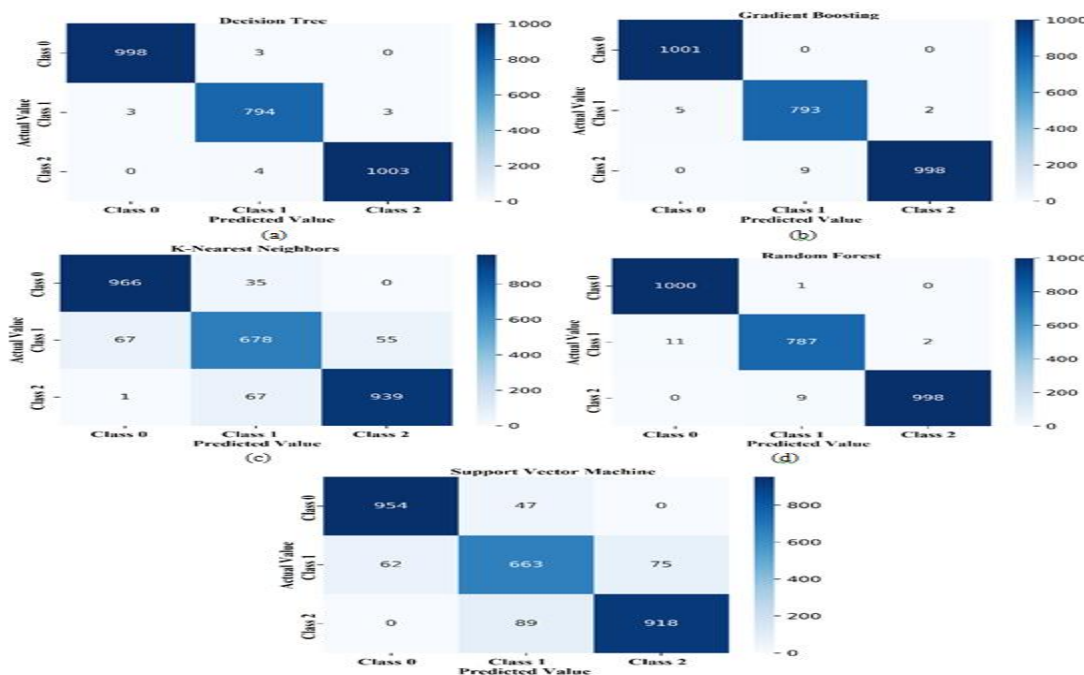


Figure 3: Confusion Matrix of (a) Decision Tree (b) Gradient Boosting (c) KNN (d) Random Forest (e) SVM

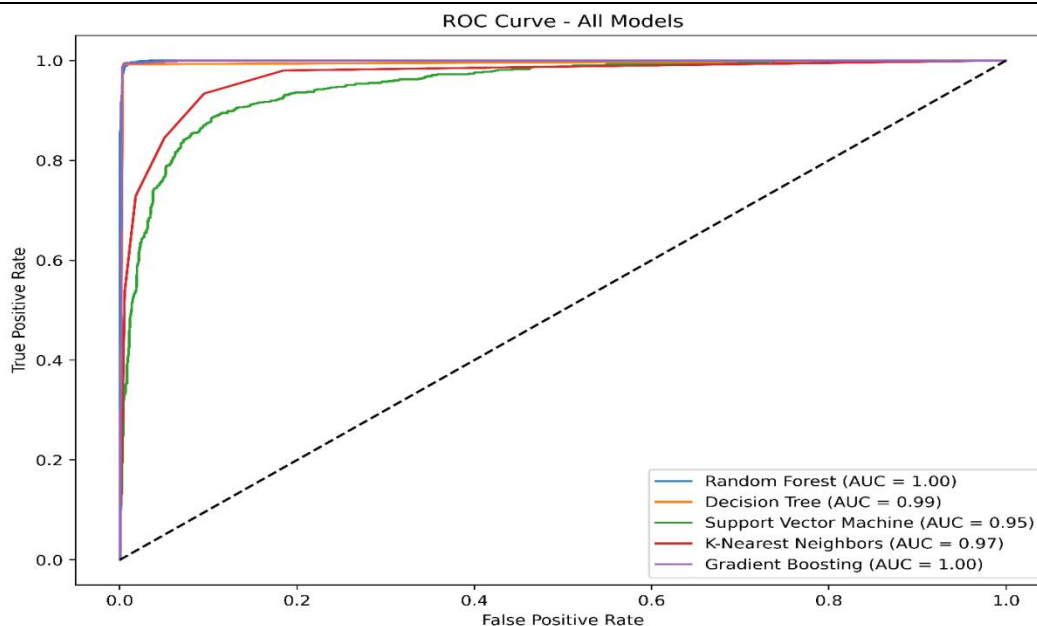


Figure 4: ROC curve of our models

Fig.4, presents the Receiver Operating Characteristic (ROC) curve of our classification models, indicating their performance in distinguishing air quality categories. The Area under the Curve (AUC) values highlight the effectiveness of our labelling scheme, with Random Forest and Gradient Boosting achieving an AUC of 1.00, indicating perfect classification ability. Decision Tree follows closely with an AUC of 0.99, while KNN and SVM achieve 0.97 and 0.95, respectively. Additionally, the sharp rise in True Positive Rate (TPR) with minimal False Positive Rate (FPR) signifies high model reliability and classification accuracy. The findings suggest that our thresholds-based labelling method provides a reliable classification framework, making it valuable for real-world air quality assessment.

V. Conclusion

This study proposed a novel labelling method and evaluated five supervised learning models—Random Forest, Decision Tree, SVM, KNN and Gradient Boosting—for air quality assessment. According to our results, tree based models, particularly Random Forest and Gradient Boosting, achieved the highest classification accuracy, making them highly suitable for real world air quality prediction. Additionally, our novel labelling scheme, based on pollutant concentrations and standard weighted scores, provided a scalable framework for categorizing air quality into multiple classes which improved adaptability comparing with traditional methods. In future, we will integrate SMD sensor data and employ our models using edgeAI and tinyML for low power IoT devices.

References

1. Dockery, D. W., Pope, C. A., Xu, X., Spengler, J. D., Ware, J. H., Fay, M. E., ... & Speizer, F. E. (1993). An association between air pollution and mortality in six U.S. cities. *New England Journal of Medicine*, 329(24), 1753-1759.
2. Pope, C. A., Burnett, R. T., Thun, M. J., Calle, E. E., Krewski, D., Ito, K., & Thurston, G. D. (2002). Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. *JAMA*, 287(9), 1132-1141.
3. Kim, S. et al. (2018). "Standardization Challenges in Air Quality Monitoring Sensors." *Journal of Sensor Technology*, 12(1), pp. 45-60.
4. Zhou, X., et al. (2020). "Challenges in Low-Cost Sensor-Based Air Quality Monitoring." *Environmental Monitoring Journal*, 34(4), pp. 512-530.
5. Smith, J., & Brown, K. (2019). "Environmental Influences on Air Pollution Measurements." *Atmospheric Research*, 28(3), pp. 210-225.
6. Qiu, M., Zigler, C. and Selin, N.E., 2022. Statistical and machine learning methods for evaluating trends in air quality under changing meteorological conditions. *Atmospheric chemistry and physics*, 22(16), pp.10551-10566.
7. World Health Organization (2022). "Global Air Quality Guidelines." WHO Press.
8. Box, G. E., & Jenkins, G. M. (1976). *Time Series Analysis: Forecasting and Control*. Holden-Day.
9. Seinfeld, J. H., & Pandis, S. N. (1998). *Atmospheric Chemistry and Physics: From Air Pollution to Climate Change*. Wiley.
10. Jiang, Y., Li, C., & Zhao, J. (2017). Air pollution prediction in Beijing using machine learning models. *Environmental Science & Technology*, 51(18), 10281-10289.
11. Li, W., Wang, Y., & Xie, L. (2019). Gradient boosting machines for air pollution forecasting: A case study in China. *Journal of Environmental Management*, 252, 109682.

12. Zhang, R., Wu, Y., & Chen, X. (2020). LSTM-based deep learning model for air quality prediction. *Neural Networks*, 132, 312-321.
13. Feng, J., Liu, X., & Song, Q. (2021). Hybrid ARIMA-deep learning model for improved air quality prediction. *Atmospheric Environment*, 244, 117941.
14. Kumar, R., Singh, A., & Verma, P. (2022). IoT-enabled air quality monitoring using machine learning algorithms. *IEEE Internet of Things Journal*, 9(5), 3290-3301.
15. Vito, S. (2008). Air Quality [Dataset]. UCI Machine Learning Repository. <https://doi.org/10.24432/C59K5F>
16. Jayaratne, R., et al. (2018). "Evaluation of the performance of low-cost PM sensors." *Environmental Science & Technology*, 52(9), 5135–5141. DOI: 10.1021/acs.est.8b01826
17. Li, T., et al. (2020). "Effects of meteorological factors on PM_{2.5} concentrations in Beijing, China: A spatiotemporal analysis." *Scientific Reports*, 10, 10235. DOI: 10.1038/s41598-020-71338-7
18. Zheng, A., & Casari, A. (2018). *Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists*. O'Reilly Media.
19. Sharma, A., & Kaur, H. (2022). A Comparative Study on Data Splitting Strategies for Machine Learning Models. *International Journal of Computer Science and Artificial Intelligence*, 10(3), 45-52.