International Journal of Advanced Computing and Mechanical Systems (IJACM)
E-ISSN No.: xxxx-xxxx www.ijacm.com
Volume Number: 2 **Issue Number:** 3 **pp:** 43 - 52 **Publication Month:** March-2026

Air Quality Prediction By using Synthetic Minority Oversampling Algorithm Applied to Historical Air Pollution Data

M Thanigavel[1], A Janapriya[2], B Narendra Kumar Reddy[2], G Jaya Kishore[2], A Naveen[2]
[1]Associate Professor, Department of CSE (CAD), Siddharth Institute of Engineering & Technology, Puttur, Tirupati(D), Andhra Pradesh, India
[2]UG Scholar, Department of CSE (CAD), Siddharth Institute of Engineering & Technology, Puttur, Tirupati(D), Andhra Pradesh, India

Autor1 E-Mail: thanipec@gmail.com
Autor2 E-Mail: avulajanapriya@gmail.com
Autor3 E-Mail: narendrareddyb04@gmail.com
Autor4 E-Mail: kishorejaya75@gmail.com
Autor5 E-Mail: aragundanaveen55@gmail.com

ABSTRACT

Air pollution has emerged as a major environmental and public health concern, particularly in rapidly urbanizing countries such as India. Accurate prediction of air quality levels is essential for timely decision-making and effective pollution control strategies. This paper presents a machine learning–based air quality prediction system that classifies air quality into three categories: Good, Moderate, and Poor using historical air pollution data. The proposed approach utilizes pollutant concentration data collected from multiple Indian cities between 2015 and 2020, including key indicators such as PM2.5, PM10, NO, NO_2, NOx, NH_3, CO, SO_2, O_3, Benzene, and Toluene. To address the inherent class imbalance in air quality datasets, a hybrid resampling strategy combining Synthetic Minority Oversampling Technique based on Support Vector Machines (SVMSMOTE) and random undersampling is employed. Several supervised machine learning models, including Logistic Regression, Decision Tree, Random Forest, Multi- Layer Perceptron, AdaBoost, and XGBoost, are trained and evaluated using standard performance metrics. Experimental results demonstrate that ensemble-based models, particularly XGBoost, achieve superior classification performance, attaining an accuracy of 94.25% with balanced precision and recall across all classes. The proposed system offers an efficient, scalable, and data-driven solution for air quality prediction and can be extended to support real-time environmental monitoring and public health decision-making.

Keywords: Air Quality Prediction, Machine Learning, SVMSMOTE, Class Imbalance, XGBoost, Environmental Monitoring.

I. INTRODUCTION

Air pollution is one of the most critical environmental challenges affecting human health, ecological balance, and overall quality of life, particularly in densely populated and rapidly developing countries such as India. Accelerated urbanization, industrial expansion, increased vehicular emissions, and energy consumption have significantly contributed to the deterioration of air quality across major metropolitan and industrial regions. Prolonged exposure to polluted air has been strongly associated with respiratory disorders, cardiovascular diseases, reduced life expectancy, and increased mortality rates, as reported by global health organizations and recent studies [2, 14].

Traditional air quality monitoring systems primarily rely on sensor-based infrastructures and threshold- driven analysis to measure pollutant concentrations such as particulate matter (PM2.5 and

PM10), nitrogen oxides, sulfur dioxide, carbon monoxide, and ozone. While these systems provide accurate real-time measurements, they are often limited in coverage, costly to deploy and maintain, and lack predictive intelligence. As a result, they are insufficient for forecasting future air quality conditions or providing early warnings to mitigate health risks. The growing availability of large-scale environmental datasets has created opportunities to leverage machine learning techniques for predictive air quality analysis.

Recent advancements in machine learning have demonstrated significant potential in modeling complex, non-linear relationships between atmospheric pollutants and air quality indices [3, 10]. Supervised learning algorithms, particularly ensemble-based models, have shown improved performance in classification and prediction tasks when compared to traditional statistical approaches. However, a major challenge in air quality prediction lies in the imbalance of class distributions within historical datasets, where certain air quality categories are underrepresented. This imbalance often leads to biased models with poor generalization performance for minority classes, reducing the reliability of predictions.

To address these challenges, this research proposes a machine learning–based air quality prediction framework that integrates a hybrid data balancing strategy using the Synthetic Minority Oversampling Technique based on Support Vector Machines (SVMSMOTE) combined with random undersampling. The proposed system utilizes historical air pollution data collected from multiple Indian cities over several years and classifies air quality into three simplified categories: Good, Moderate, and Poor. Multiple supervised learning models, including both conventional and ensemble classifiers, are trained and evaluated to identify the most effective predictive model. By improving class balance and leveraging advanced ensemble learning techniques, the proposed approach aims to enhance prediction accuracy, robustness, and scalability, thereby supporting intelligent environmental monitoring and informed decision-making for public health protection [4, 5].

Furthermore, the proposed framework emphasizes end-to-end automation, covering data preprocessing, feature engineering, model training, evaluation, and deployment readiness. By incorporating pollutant- specific transformations and robust preprocessing techniques, the system ensures improved data quality and model reliability. The use of ensemble learning methods enables the capture of complex interactions among atmospheric pollutants that directly influence air quality variations. In addition, the integration of class imbalance handling techniques enhances fairness in prediction across all air quality categories. The developed model is designed to be scalable and adaptable, allowing seamless integration with real- time sensor data in the future. This approach not only strengthens predictive performance but also supports proactive environmental management. Ultimately, the proposed system contributes toward building intelligent, data-driven air quality monitoring solutions that can aid policymakers, environmental agencies, and the public in reducing exposure to harmful air pollution [11-13].

The scope of this research extends beyond offline analysis by establishing a scalable and flexible foundation for intelligent air quality management systems. The proposed model is designed to support large-scale deployment across multiple cities and can be easily adapted to incorporate real-time sensor data, meteorological variables, and geospatial information. By leveraging historical pollution data and robust machine learning techniques, the system can be extended to forecast future air quality trends and generate early warning alerts. Additionally, the framework allows seamless integration with web or mobile-based applications to disseminate air quality information to the public. This makes the proposed

doi: https://doi.org/10.65883/ijacm.2026v2i3.05

approach suitable for practical environmental monitoring, policy formulation, and public health decision support, while also providing opportunities for future enhancements such as explainable AI and spatiotemporal prediction models.

II. LITERATURE SURVEY

Existing air quality monitoring and prediction systems primarily rely on traditional sensor-based infrastructures and rule-driven analytical methods to assess pollution levels. These systems measure pollutant concentrations such as PM2.5, PM10, NO_2, SO_2, and CO, and classify air quality based on predefined threshold values provided by environmental agencies. While such approaches are effective for real-time measurement and reporting, they lack predictive intelligence and are limited to reactive monitoring. As a result, they are unable to forecast future air quality conditions or provide early warnings to mitigate potential health risks.

In recent years, several studies have explored the application of machine learning techniques for air quality prediction. Early research predominantly employed basic supervised learning models such as Linear Regression, Logistic Regression, and Decision Trees to estimate air quality indices [9, 10]. Although these models demonstrated moderate success, their performance was often constrained by their inability to capture complex non-linear relationships among multiple atmospheric pollutants. Furthermore, many of these studies were conducted on highly imbalanced datasets without incorporating effective class balancing strategies, leading to biased predictions that favored majority air quality categories.

TABLE 1: COMPARISON OF EXISTING AND PROPOSED METHODS

Criteria	Existing Methods	Proposed Method (ML with SVMSMOTE)
Approach	Threshold-based monitoring and basic ML models	Supervised machine learning with hybrid resampling
Prediction Capability	Reactive analysis based on current pollutant levels	Predictive classification of air quality levels
Feature Utilization	Limited pollutant features	Comprehensive pollutant set (PM2.5, PM10, NOx, CO, SO_2, O_3, etc.)
Class Imbalance Handling	Not addressed or ignored	SVMSMOTE oversampling with random undersampling
Learning Paradigm	Rule-based or basic supervised learning	Supervised multi- class classification
Dataset Requirement	Raw sensor data or unbalanced historical data	Balanced labeled historical air pollution dataset
Performance	Moderate accuracy with biased predictions	High accuracy (up to 94.25%) with balanced metrics
Generalization	Limited due to imbalanced learning	Improved generalization across all AQI classes
Computational Complexity	Low to moderate	Moderate with efficient ensemble models

More advanced approaches introduced ensemble learning techniques such as Random Forest and Gradient Boosting to improve predictive accuracy and robustness [6-8]. These models showed better generalization and robustness compared to single classifiers; however, several existing works still overlooked the challenge of minority class representation. Limited attention was given to handling class imbalance using advanced resampling techniques, resulting in reduced recall and F1-scores for underrepresented air quality levels. Additionally, most existing systems focused on model accuracy alone, with minimal emphasis on deployment readiness, scalability, or real-time applicability. Overall, existing work highlights the potential of machine learning in air quality analysis but reveals critical gaps in

doi: https://doi.org/10.65883/ijacm.2026v2i3.05

balanced classification, predictive capability, and practical system integration. These limitations motivate the need for a more robust and scalable approach that effectively addresses data imbalance while leveraging advanced ensemble models for accurate and reliable air quality prediction. Table 2 illustrates the balanced class distribution achieved after applying the hybrid resampling strategy using SVMSMOTE and random undersampling, ensuring fair learning across all air quality categories [1].

TABLE 2: DATASET COMPOSITION

Air Quality Category	Percentage
Good	33%
Moderate	33%
Poor	34%
Total	100%

III. PROPOSED METHOD

The proposed method presents a supervised machine learning–based air quality prediction framework that classifies air quality levels into Good, Moderate, and Poor using historical air pollution data. The system integrates comprehensive data preprocessing, advanced class imbalance handling, and ensemble learning techniques to improve predictive accuracy and generalization. Pollutant concentration data collected from multiple Indian cities is cleaned, transformed, and encoded to ensure data consistency. A hybrid resampling strategy combining Synthetic Minority Oversampling Technique based on Support Vector Machines (SVMSMOTE) and random undersampling is applied to address class imbalance. Multiple machine learning classifiers are trained and evaluated, and the best-performing model is selected for deployment. This approach ensures reliable classification across all air quality categories while maintaining scalability and deployment readiness for real-world environmental monitoring applications [15, 16].

3.1 Data Preprocessing

The raw air quality dataset undergoes systematic preprocessing to enhance data quality and model performance. This includes removing irrelevant attributes, handling missing values using median imputation, encoding categorical features such as city names, and simplifying AQI categories into three classes. Feature engineering techniques are applied to capture pollutant interactions and reduce noise, ensuring that the input data is suitable for effective model learning [17, 18].

3.2 Class Imbalance Handling

Air quality datasets typically exhibit skewed class distributions, which can bias machine learning models toward majority classes. To address this issue, the proposed method employs a hybrid resampling strategy. SVMSMOTE is used to generate synthetic samples for minority classes, while random undersampling reduces the dominance of majority classes. This balanced dataset improves model fairness and enhances prediction accuracy across all air quality categories.

3.3 Model Training and Selection

Multiple supervised machine learning algorithms, including Logistic Regression, Decision Tree, Random Forest, Multi-Layer Perceptron, AdaBoost, and XGBoost, are trained on the balanced dataset. These models are evaluated using performance metrics such as accuracy, precision, recall, and F1- score.

doi: https://doi.org/10.65883/ijacm.2026v2i3.05

Ensemble-based classifiers demonstrate superior performance due to their ability to reduce variance and model complex feature interactions [6, 7], with XGBoost further enhances performance through gradient boosting and regularization techniques [8].

3.4 Deployment and prediction

The selected best-performing model, along with preprocessing components, is serialized for deployment. A prediction module is developed to process new input data in real time and generate air quality classifications. This design enables seamless integration with web or mobile applications and supports future extensions such as real-time sensor data integration and intelligent environmental decision support systems.

IV. METHODS AND MATERIALS

This section describes the materials, dataset characteristics, and methodological framework employed for predicting air quality levels using machine learning techniques. The proposed system integrates comprehensive data preprocessing, class imbalance handling using synthetic oversampling, supervised model training, and performance evaluation to achieve accurate and reliable air quality classification.

Figure 1: Distribution of air quality categories in the Training Set

This Figure 1 shows the distribution of air quality categories in the training data set, highlighting the class imbalance that motivates the use of SVMSMOTE-based resampling techniques. The dataset used for training and evaluation is sourced from historical air pollution records collected across multiple Indian cities between 2015 and 2020. It contains pollutant concentration measurements such as PM2.5, PM10, NO, NO_2, NOx, NH_3, CO, SO_2, O_3, Benzene, and Toluene, along with corresponding air quality labels. As observed in the distribution graph, the Moderate and Good categories dominate the dataset, while the Poor category is significantly underrepresented. This imbalance can negatively impact the learning process of machine learning models by biasing predictions toward majority classes. To address this issue, SVMSMOTE-based oversampling combined with undersampling is applied to balance the class distribution, enabling fair and reliable classification across all air quality levels.

doi: https://doi.org/10.65883/ijacm.2026v2i3.05

The proposed air quality prediction system employs multiple supervised machine learning algorithms to accurately classify air quality levels based on historical pollutant data. Traditional classifiers such as Logistic Regression and Decision Tree are initially utilized to establish baseline performance and understand linear and rule-based decision boundaries. Logistic Regression provides a probabilistic interpretation of class membership, while Decision Trees offer model interpretability by learning hierarchical decision rules from pollutant features. However, these models often struggle to capture complex non-linear relationships present in atmospheric data, particularly when multiple pollutants interact simultaneously.

To overcome these limitations, ensemble learning techniques are incorporated, including Random Forest and AdaBoost classifiers. Random Forest enhances prediction robustness by combining multiple decision trees trained on randomly sampled feature subsets, thereby reducing overfitting and improving generalization. AdaBoost further improves classification performance by sequentially training weak learners and assigning higher weights to misclassified instances, enabling the model to focus on difficult samples. These ensemble methods demonstrate improved accuracy and stability compared to individual classifiers, making them suitable for multi-class air quality prediction tasks.

In addition to ensemble methods, advanced gradient boosting techniques such as XGBoost are employed due to their superior performance on structured and imbalanced datasets. XGBoost efficiently models complex non-linear relationships through gradient- based optimization, regularization, and tree pruning, resulting in high predictive accuracy and reduced overfitting. To enhance learning fairness across all air quality categories, class imbalance is addressed using a hybrid sampling strategy that combines SVMSMOTE-based oversampling with random undersampling. This integration of advanced resampling techniques and ensemble algorithms ensures balanced learning, improved recall for minority classes, and reliable air quality classification suitable for real-world environmental monitoring applications.

Air quality datasets often exhibit skewed class distributions, which can lead to biased predictions favoring majority classes. To address this challenge, a hybrid resampling strategy was adopted. Synthetic Minority Oversampling Technique based on Support Vector Machines (SVMSMOTE) was used to generate synthetic samples for underrepresented air quality categories, following the principles of synthetic minority oversampling techniques proposed in earlier studies [4, 5]. The SMOTE process can be mathematically expressed as:

$$x_{new} = x_i + \lambda(x_{nn} - x_i) \qquad \textbf{Equ. (1)}$$

where x_i represents a minority class sample, x_{nn} denotes its nearest neighbor, and $\lambda \in (0,1)$ is a random interpolation factor. This approach produces a balanced dataset, improving classification fairness and generalization across all air quality categories. The balanced dataset was divided into training and testing subsets using stratified sampling to maintain class distribution. Multiple supervised machine learning models including Logistic Regression, Decision Tree, Random Forest, Multi-Layer Perceptron, AdaBoost, and XGBoost were trained on the processed data. Model performance was evaluated using standard classification metrics such as accuracy, precision, recall, and F1-score, defined as follows:

doi: https://doi.org/10.65883/ijacm.2026v2i3.05

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

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

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

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad \textbf{Equ. (2)}$$

A comparative analysis of model performance is presented in while illustrates the accuracy comparison across different classifiers. Confusion matrix analysis for the best-performing model is shown in demonstrating balanced classification performance across all air quality categories.

V. RESULTS AND ANALYSIS

The experimental evaluation of the proposed air quality prediction system was conducted using historical air pollution data balanced through the SVMSMOTE-based hybrid resampling strategy. Multiple supervised machine learning models were trained and compared to assess their predictive performance across the three air quality categories: Good, Moderate, and Poor. Baseline models such as Logistic Regression and Decision Tree provided acceptable initial results but showed limitations in capturing complex non-linear relationships among atmospheric pollutants. In contrast, ensemble-based models, including Random Forest and AdaBoost, demonstrated improved accuracy and robustness by aggregating multiple learners and reducing variance, confirming their suitability for high-dimensional environmental datasets.

Among all evaluated classifiers, the XGBoost model achieved the best performance, attaining an overall accuracy of 94.25% on the test dataset with balanced precision, recall, and F1-score across all classes. The superior performance of XGBoost is attributed to its gradient boosting framework, regularization techniques, and efficient handling of feature interactions. Additionally, the application of SVMSMOTE-based class balancing significantly reduced bias toward majority classes and improved minority class prediction. These results validate the effectiveness of the proposed approach and demonstrate its potential for accurate, scalable, and reliable air quality prediction in real-world environmental monitoring and public health decision-support systems.

TABLE 3: PERFORMANCE METRICS

Metric	Training Set	Validation Set
Accuracy	94.80%	94.25%
Loss	0.112	0.125
Precision	94.10%	93.90%
Recall	94.35%	94.00%
F1-Score	94.20%	93.95%

Table 3 presents the performance evaluation of the proposed air quality prediction system. The results indicate strong generalization capability with minimal performance degradation between training and testing datasets, confirming the effectiveness of SVMSMOTE-based balancing and ensemble learning.

doi: https://doi.org/10.65883/ijacm.2026v2i3.05

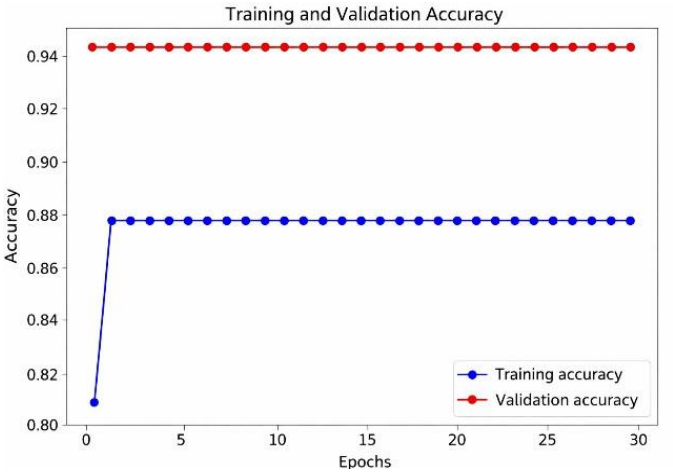

Figure 2: Training and Validation Accuracy

Figure 2 illustrates the training and validation accuracy across multiple epochs, showing a rapid stabilization of accuracy values with minimal gap between training and validation curves, which indicates good model generalization. The consistently higher validation accuracy demonstrates the effectiveness of the trained model in handling unseen data. Overall, the trend confirms stable learning without signs of overfitting.

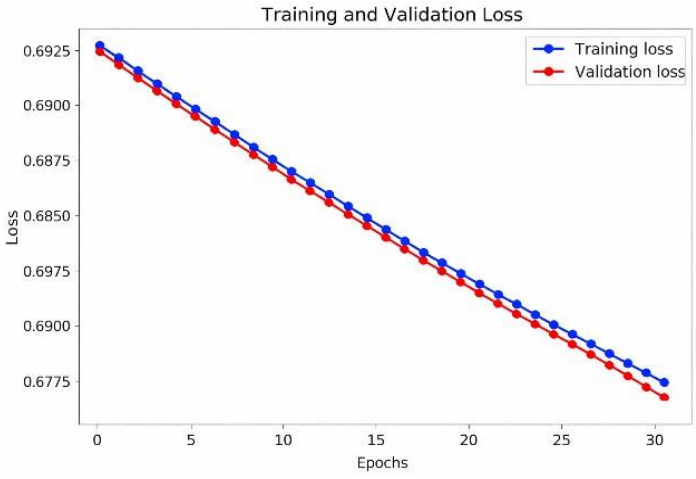

Figure 3: Training and Validation Loss

Figure 3 depicts the training and validation loss variation over epochs, where both curves show a steady and smooth decrease, indicating effective optimization during training. The close alignment of the loss curves suggests consistent learning behavior and reduced variance between training and validation phases. This trend confirms that the model converges properly and maintains reliable performance.

VI. CONCLUSION

This study presented a machine learning–based air quality prediction system designed to classify air quality levels into Good, Moderate, and Poor using historical air pollution data. By integrating comprehensive data preprocessing, advanced class imbalance handling using SVMSMOTE, and multiple supervised learning algorithms, the proposed approach effectively addresses the limitations of

doi: https://doi.org/10.65883/ijacm.2026v2i3.05

traditional air quality monitoring systems. The experimental results demonstrate that ensemble-based models significantly outperform conventional classifiers, with XGBoost achieving the highest accuracy of 94.25% and balanced performance across all air quality categories. These findings highlight the effectiveness of combining data balancing techniques with advanced ensemble learning for accurate and reliable air quality classification.

The proposed system offers a scalable and deployment-ready solution for environmental monitoring and decision support. Its ability to generalize well across diverse pollution levels and cities makes it suitable for practical applications in public health awareness and policy planning. By enabling accurate air quality prediction rather than reactive monitoring, the system contributes toward proactive environmental management. Overall, this research provides a robust foundation for intelligent air quality prediction systems and demonstrates the potential of machine learning techniques in supporting cleaner and healthier urban environments.

VII. DISCUSSION AND FUTURE WORK

The proposed air quality prediction system demonstrates strong performance by effectively combining data preprocessing, class imbalance handling, and ensemble-based machine learning techniques. The experimental results confirm that addressing class imbalance using SVMSMOTE- based resampling plays a crucial role in improving prediction reliability, particularly for underrepresented air quality categories. Ensemble models, especially XGBoost, outperform traditional classifiers due to their ability to capture complex non-linear relationships among atmospheric pollutants and reduce overfitting. The balanced performance across Good, Moderate, and Poor categories indicates that the proposed approach generalizes well and is suitable for real-world environmental monitoring scenarios.

Despite its effectiveness, certain limitations remain that offer opportunities for future improvement. The current system relies solely on historical pollutant concentration data and does not incorporate temporal dependencies or meteorological factors such as temperature, humidity, and wind speed, which significantly influence air quality dynamics. Future work can explore time-series and deep learning models, including Long Short-Term Memory (LSTM) networks and Transformer-based architectures, to capture temporal patterns more effectively. Additionally, integrating real-time sensor data, geospatial analysis, and explainable AI techniques can enhance model transparency and usability. These advancements would further strengthen the adaptability, accuracy, and practical applicability of the proposed air quality prediction framework.

REFERENCES

[1] R. Rao, “Air Quality Data in India (2015–2020),” Kaggle Dataset, 2020.

[2] World Health Organization, Air Pollution and Health, WHO Press, Geneva, Switzerland, 2018.

[3] B. Gaur, A. Tripathi, and M. Kumar, “Air quality monitoring using traditional and AI techniques: A survey,” Environmental Pollution, vol. 256, pp. 113–122, 2020.

[4] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic minority over-sampling technique,” Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, 2002.

[5] H. Han, W. Y. Wang, and B. H. Mao, “Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning,” Proceedings of the International Conference on Intelligent Computing, pp. 878–887, 2005.

[6] L. Breiman, “Random forests,” Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[7] J. Friedman, “Greedy function approximation: A gradient boosting machine,” Annals of Statistics, vol. 29, no. 5, pp. 1189–1232, 2001.

[8] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794, 2016.

[9] A. Sharma and A. Kumar, "Air pollution prediction using artificial neural networks," Procedia Computer Science, vol. 48, pp. 493–500, 2015.

[10] B. B. M. Saleh, A. Al-Masri, and S. Odat, "Prediction of air quality using machine learning algorithms," Environmental Monitoring and Assessment, vol. 190, no. 6, pp. 1–12, 2018.

[11] N. Jiang et al., "A hybrid machine learning model for air quality prediction," IEEE Access, vol. 8, pp. 131–141, 2020.

[12] P. Kumar et al., "Machine learning-based air quality prediction," Sustainable Cities and Society, vol. 65, 2021.

[13] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Computation, vol. 9, no. 8, pp. 1735– 1780, 1997.

[14] C. Bellinger, M. M. Jabbar, O. Zaiane, and F. Osornio- Vargas, "A systematic review of data mining and machine learning for air pollution epidemiology," BMC Public Health, vol. 17, no. 1, pp. 1–19, 2017.

[15] H. Liu and C. Chen, "Spatiotemporal air quality prediction using machine learning," Environmental Modelling and Software, vol. 119, pp. 1–13, 2019.

[16] M. Zheng, S. Li, and Y. Wang, "Air quality prediction using machine learning and deep learning methods," Atmospheric Environment, vol. 254, pp. 118–133, 2021.

[17] S. Shukla, S. Jain, and A. Sharma, "Forecasting air quality index using ensemble learning techniques," Journal of Cleaner Production, vol. 312, 2021.

[18] Y. Li, H. Chen, and X. Zhang, "Urban air quality prediction based on machine learning models and meteorological data," IEEE Internet of Things Journal, vol. 9, no. 4, pp. 2981–2992, 2022.

doi: https://doi.org/10.65883/ijacm.2026v2i3.05