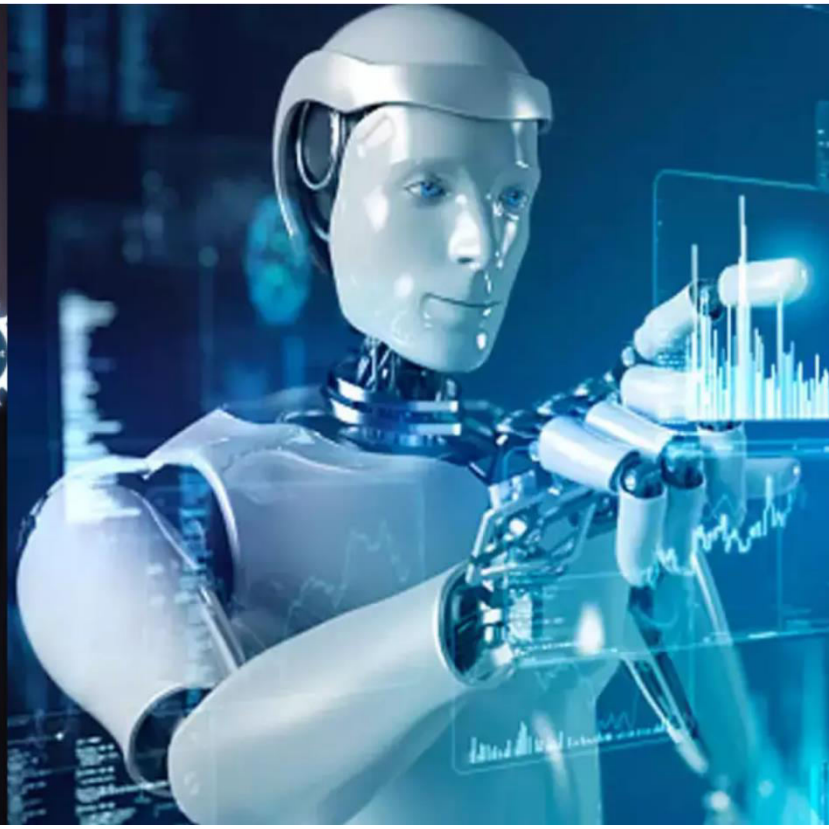
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

ISSN(O): 2319-8753
ISSN(P): 2347-6710

International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.699

Volume 15, Issue 3, March 2026

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503043|

Air Quality Prediction Using Ensemble Machine Learning

P Lokesh Kumar, R Snehasree, G Renuka, R Ranjith, G sathyavel

Assistant Professor, Dept. of CSE, Kuppam Engineering College, Kuppam, Andhra Pradesh, India

Student, Dept. of CSE, Kuppam Engineering College, Kuppam, Andhra Pradesh, India

Student, Dept. of CSE, Kuppam Engineering College, Kuppam, Andhra Pradesh, India

Student, Dept. of CSE, Kuppam Engineering College, Kuppam, Andhra Pradesh, India

Student, Dept. of CSE, Kuppam Engineering College, Kuppam, Andhra Pradesh, India

ABSTRACT: Air pollution has become a serious environmental and public health concern due to rapid urbanization, industrial growth, and increased vehicular emissions. Pollutants such as PM2.5, PM10, NO_2, SO_2, CO, and O_3 significantly affect human health and climate conditions. Accurate air quality prediction is essential for early warning systems and environmental planning. Traditional statistical and single machine learning models often fail to capture the nonlinear and complex relationships among air pollutants and meteorological parameters.

This paper proposes an ensemble machine learning framework for air quality prediction using Gradient Boosting, AdaBoost, XGBoost, Voting Classifier, and Stacking Classifier. The dataset is divided into training and testing sets using an 80:20 ratio. The proposed models are evaluated using confusion matrix and classification report metrics. Experimental results demonstrate that ensemble models outperform individual classifiers in terms of accuracy and robustness. The proposed approach provides a scalable and reliable solution for air quality monitoring and prediction.

KEYWORDS: Air Quality Prediction, Ensemble Learning, Gradient Boosting, AdaBoost, XGBoost, Voting Classifier, Stacking Classifier, Machine Learning.

I. INTRODUCTION

Air pollution is one of the most critical environmental problems affecting human health and ecological balance. Rapid industrialization, population growth, and increased transportation activities have led to a significant rise in harmful pollutants such as particulate matter (PM2.5 and PM10), nitrogen dioxide (NO_2), sulfur dioxide (SO_2), carbon monoxide (CO), and ozone (O_3). Prolonged exposure to these pollutants can cause respiratory diseases, cardiovascular disorders, and reduced life expectancy.

Traditional air quality prediction methods are based on statistical models such as linear regression and autoregressive integrated moving average (ARIMA). Although these methods provide baseline predictions, they struggle to model nonlinear relationships and interactions between pollutants and meteorological factors. With the availability of large environmental datasets, machine learning has emerged as a powerful tool for air quality prediction.

However, single machine learning models may suffer from overfitting and poor generalization when trained on complex datasets. Ensemble learning techniques overcome these limitations by combining multiple learning algorithms to produce more accurate and stable predictions. Boosting and stacking methods are widely used to improve classification performance by reducing bias and variance.

This research focuses on developing an ensemble-based air quality prediction system using Gradient Boosting, AdaBoost, XGBoost, Voting Classifier, and Stacking Classifier. The objective is to enhance prediction accuracy and reliability for air quality classification.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503043|

II. PROBLEM STATEMENT

1. **Limited Coverage of Monitoring Stations**
In the existing system, air quality is monitored using fixed air quality monitoring stations placed at selected locations. These stations measure pollutants such as PM2.5, PM10, NO_2, and CO. Due to high installation and maintenance costs, only a limited number of stations are deployed, which cannot cover the entire metropolitan area. As pollution levels vary from place to place, the collected data does not accurately represent the overall air quality of the city.

2. Low Accuracy of Existing Prediction Methods
The existing system uses statistical models and basic machine learning techniques to analyzeair qualitydata.These models are not capable of handling complex pollution patterns and interactions between pollutants and weather conditions. As a result, the prediction accuracy is low and the system fails to provide reliable air quality forecasts.

3. Lack of Early Prediction and Forecasting
Current air quality monitoring systems mainly provide information about present pollution levels rather than predicting future conditions. This prevents people and authorities from taking preventive actions in advance. Without early prediction, sudden pollution peaks cannot be managed properly, increasing health risks for the public.

4. Data Quality and Processing Issues
Air quality data collected from sensors often contains missing values, noise, and inconsistencies. In addition, integrating pollution data with meteorological and traffic data is difficult without proper preprocessing and feature engineering. These data quality issues reduce the effectiveness of prediction models and lead to unreliable results.

5. Lack of Scalable and Advanced Prediction System
The existing system does not support large-scale prediction for metropolitan areas and does not make use of advanced ensemble machine learning techniques. It is not suitable for handling large and complex datasets generated from multiple sources. This limits the system's ability to provide accurate and stable air quality predictions for smart city applications.

III. SYSTEM OVERVIEW

The proposed system predicts air quality health impacts by integrating environmental, meteorological, and health-related data, including AQI, PM2.5, PM10, NO_2, SO_2, O_3, temperature, humidity, wind speed, respiratory and cardiovascular cases, hospital admissions, and a Health Impact Score. The processed data is fed into an ensemble learning framework combining Gradient Boosting, XGBoost, AdaBoost, Stacking, and Voting Classifiers, improving accuracy, reducing overfitting, and providing stable classification across health impact levels. The system classifies air quality from safe to severe risk, enabling reliable and actionable forecasts to support preventive health measures and informed decision-making.

A. Existing System

Existing air quality prediction systems rely on statistical models or individual machine learning algorithms such as decision trees, linear regression, or k-nearest neighbours. These approaches are limited in handling complex nonlinear dependencies and often yield lower prediction accuracy.

B. Proposed System

The proposed system integrates multiple ensemble learning models including Gradient Boosting, AdaBoost, XGBoost, Voting Classifier, and Stacking Classifier. The dataset is preprocessed and divided into training and testing subsets. Each model is trained and evaluated using confusion matrix and classification report metrics. The ensemble approach improves prediction accuracy and stability compared to single models.

IV. METHODOLOGY AND IMPLEMENTATION

A. Dataset Overview

The dataset used in this study combines environmental and health-related features to predict the impact of air quality on public health. The key features include:

- **Air Quality Index (AQI):** A composite index representing the overall air pollution level.
- **Particulate Matter (PM10, PM2.5):** Fine particles that can penetrate the lungs and bloodstream, leading to respiratory and cardiovascular issues.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503043|

- **Gaseous Pollutants (NO_2, SO_2, O_3):** Harmful gases associated with industrial and vehicular emissions, contributing to respiratory distress and chronic health conditions.
- **Meteorological Conditions:** Temperature, humidity, and wind speed influence pollutant dispersion and concentration.
- **Health Outcomes:** Number of respiratory cases, cardiovascular cases, hospital admissions, and the derived HealthImpactScore.

The **target variable**, HealthImpactClass, categorizes the overall health risk into five classes: from **Class 0 (Safe)** to **Class 4 (Severe Risk)**. This classification enables public health authorities and individuals to understand the potential impact of air pollution levels on health.

B. Data Preprocessing

Prior to model training, the dataset underwent several preprocessing steps to ensure data quality and model readiness:

1. **Data Cleaning:** Missing values and inconsistencies in the dataset were handled to prevent errors during model training.
2. **Feature Selection:** All environmental, meteorological, and health outcome features were retained to capture complex interactions affecting health impact.
3. **Normalization and Scaling:** Features such as AQI, PM concentrations, and meteorological values were scaled using standardization techniques to improve convergence for algorithms like Logistic Regression and KNN.
4. **Train-Test Split:** The dataset was divided into **80% training** and **20% testing** subsets, ensuring unbiased evaluation.

These preprocessing steps ensured that the models received consistent and reliable input data.

C. Machine Learning Models

The study applied a combination of traditional machine learning models and ensemble techniques to predict health impact classes. Each model is briefly explained below:

1. **Gradient Boosting Classifier:**

Gradient Boosting sequentially builds multiple weak decision trees, with each tree correcting the errors of the previous ones. This approach effectively captures non-linear relationships between environmental factors and health outcomes.

2. **AdaBoost Classifier:**

AdaBoost improves prediction by assigning higher weights to misclassified samples, allowing the model to focus on harder-to-predict instances.

3. **XGBoost Classifier:**

XGBoost is an optimized boosting algorithm with regularization that prevents overfitting. It is computationally efficient and achieves high accuracy for tabular datasets like the air quality dataset.

4. **Voting Classifier (Majority Voting Ensemble):**

This ensemble combines Logistic Regression, K-Nearest Neighbors (KNN), and Decision Tree classifiers. Each base model votes for a class, and the majority vote determines the final prediction. This approach leverages the strengths of different algorithms to improve robustness.

5. **Stacking Classifier:**

In Stacking, multiple base learners (Logistic Regression, KNN, Decision Tree) generate predictions that are fed into a meta-learner (Logistic Regression). This allows the meta-model to learn patterns in base model predictions, improving overall accuracy, particularly for underrepresented high-risk classes.

D. Implementation Workflow

The implementation followed a systematic workflow, combining data preparation, model training, evaluation, and deployment:

1. **Feature Extraction:**

Environmental and health parameters were extracted from the dataset, ensuring the correct order for model input.

2. **Model Training:**

Each machine learning model was trained on the training subset. Ensemble methods such as Voting and Stacking were used to combine multiple models for improved performance.

3. **Model Evaluation:**

Predictions were made on the test set, and model performance was evaluated using **accuracy, precision, recall, and**

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503043|

F1-score. Confusion matrices were analyzed to identify common misclassifications, especially in higher health risk classes.

4. **Handling Imbalanced Classes:**

Since higher health risk classes were underrepresented, ensemble methods and careful feature scaling helped improve predictions for these classes.

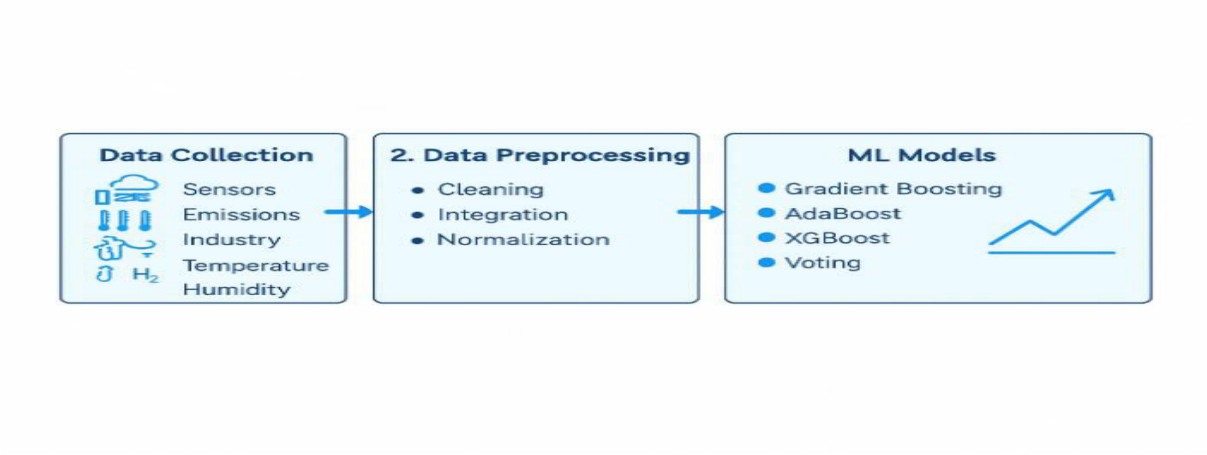
5. **Web Deployment with Flask:**

A user-friendly web application was developed using Flask. Users can input air quality, meteorological, and health-related features through an interactive form. The application processes inputs, feeds them to the trained Stacking Classifier, and returns a **color-coded health impact prediction**.

6. **User Interface Design:**

The interface uses Bootstrap for responsive design and Particle.js for visual appeal. Health risk classes are displayed with intuitive color codes:

- Green for safe,
- Yellow for mild risk,
- Orange for moderate risk,
- Red for high risk,
- Purple for severe risk.



V. EQUATIONS

The air quality prediction function is expressed as:

AQI=f(PM2.5,PM10,NO2,SO2,CO,O3,T,H)

Where:

PM2.5,PM10 – Particulate matter

NO2,SO2,CO,O3 – Gaseous pollutants

T – Temperature

H – Humidity

Final ensemble prediction:

AQIfinal=Majority(AQI1,AQI2,...,AQIn)

where AQI1,AQI2,...,AQInAQI_1, AQI_2, ..., AQI_nAQI1,AQI2,...,AQIndenote predictions obtained from different classifiers such as **Gradient Boosting, AdaBoost, XGBoost, Voting, and Stacking models**. This aggregation mechanism reduces prediction variance and improves classification stability. The mathematical formulation ensures that the prediction model can generalize well across varying pollution patterns and environmental conditions.

VI. PERFORMANCE EVALUATION

The models are evaluated using:

- **Confusion Matrix:**

The confusion matrix is used to analyze the classification performance of the models by comparing the actual air quality classes with the predicted classes. It provides detailed information about correctly and incorrectly classified

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503043|

instances for each class. This metric helps in understanding the distribution of true positives, true negatives, false positives, and false negatives, thereby offering insight into the model's prediction behavior.

- **Precision:**

Precision represents the proportion of correctly predicted air quality classes among all instances predicted as a particular class. A high precision value indicates that the model produces fewer false positive predictions. This metric is important in air quality prediction systems where incorrect classification of polluted conditions may lead to misleading health-related decisions.

- **Recall:**

Recall measures the ability of the model to correctly identify all relevant air quality classes from the dataset. It indicates how well the system detects polluted or unhealthy air quality conditions. Higher recall values imply that fewer polluted instances are missed by the model, which is essential for reliable environmental monitoring and public safety.

- **F1-score:**

The F1-score is the harmonic mean of precision and recall and provides a balanced measure of classification performance. It is particularly useful when the dataset contains imbalanced air quality classes. A higher F1-score indicates better overall model performance by considering both false positives and false negatives.

- **Accuracy:**

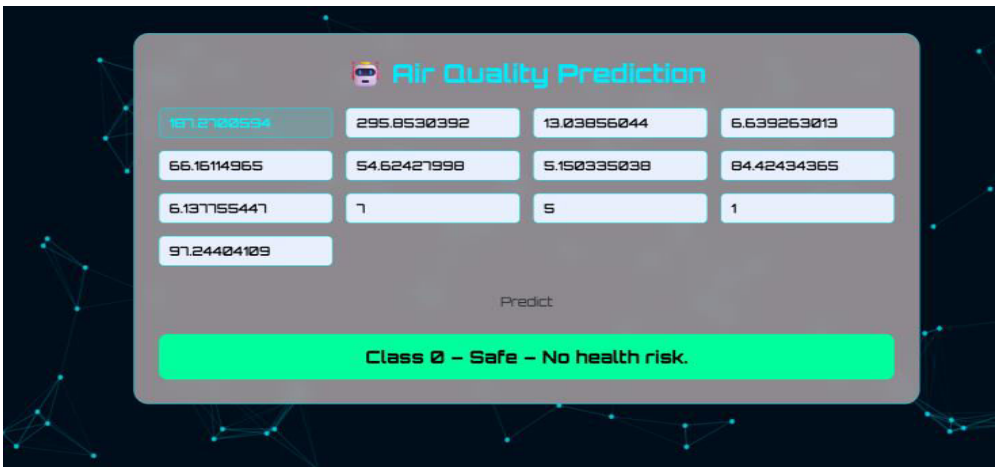
Accuracy is defined as the ratio of correctly predicted air quality instances to the total number of instances in the dataset. It provides an overall measure of how well the model predicts air quality levels. Although accuracy gives a general indication of performance, it is complemented by other metrics such as precision, recall, and F1-score for a more comprehensive evaluation.

Experimental results indicate that XGBoost and Stacking classifiers achieve higher accuracy compared to individual classifiers.

VII. EXPECTED RESULTS

The proposed system is expected to achieve the following outcomes:

1. **Improved Prediction Accuracy:**
   - By leveraging ensemble learning methods, the system is expected to provide more accurate predictions of air quality-related health impacts compared to single-model approaches.
2. **Reduced Overfitting:**
   - The use of multiple models and ensemble techniques is expected to minimize overfitting, allowing the system to generalize effectively to unseen data.
3. **Stable Classification Performance:**
   - Consistent precision, recall, and F1-scores are expected across all health impact classes, including minority categories representing severe risks.
4. **Reliable Air Quality Forecasting:**
   - Integrating environmental, meteorological, and health-related features, the system is expected to produce dependable and timely forecasts, supporting preventive measures and public health interventions.



International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503043|

VIII. FUTURE ENHANCEMENT

1. Integration of Deep Learning Models such as LSTM for Time-Series Air Quality Prediction
In future, the proposed system can be enhanced by integrating **deep learning models** such as **Long Short-Term Memory (LSTM)** networks for time-series air quality prediction. Air quality data shows strong temporal dependencies, as pollutant concentrations vary with time due to traffic flow, weather conditions, and industrial activities. LSTM models are capable of learning these **long-term and short-term temporal patterns** effectively. By using LSTM, the system can improve prediction accuracy for sequential air quality data and provide more reliable short-term and long-term forecasts compared to traditional machine learning models.

2. Real-Time Data Collection from IoT Sensors
The system can be further improved by incorporating **real-time data acquisition** from **IoT-based air quality sensors** deployed at different geographical locations. These sensors continuously measure pollutant levels such as PM2.5, PM10, CO, and NO_2 and transmit the data to the prediction system. Real-time data integration will enable **continuous monitoring** and **instant prediction** of air quality levels. This enhancement will make the system more practical for real-world applications and support early detection of hazardous pollution conditions.

3. Deployment as a Web-Based Air Quality Monitoring System
The proposed model can be deployed as a **web-based air quality monitoring platform** to provide easy access to predicted air quality information. This system can display current and forecasted air quality levels for different locations and generate warning notifications when pollution exceeds safe limits. A web-based deployment will improve **usability and accessibility** and allow users, researchers, and authorities to monitor air quality remotely. Such a platform can also support data storage, analysis, and reporting for environmental management.

4. Visualization of Pollution Trends Using Dashboards
Future enhancements can include the development of **interactive dashboards** for visualizing pollution trends and prediction results. Data visualization tools can be used to display pollutant concentration levels, historical trends, and forecasted values in the form of graphs and charts. Visualization will help users easily understand complex air quality data and identify pollution patterns over time. This enhancement supports better **decision-making** and improves public awareness of environmental conditions.

5. Extension to Long-Term Pollution Forecasting
The proposed framework can be extended for **long-term air quality forecasting** by analyzing historical air pollution data over extended periods. Long-term prediction will help in understanding seasonal and yearly pollution patterns and support urban planning and policy-making decisions. By predicting future pollution trends, authorities can implement preventive measures and design effective pollution control strategies. This enhancement will increase the usefulness of the system for **environmental planning and sustainable development**.

VIII. CONCLUSION

This paper presented an ensemble machine learning-based approach for air quality prediction using **Gradient Boosting, AdaBoost, XGBoost, Voting, and Stacking classifiers.** The proposed methodology effectively captures nonlinear relationships between air pollutants and meteorological parameters. The experimental analysis demonstrates that ensemble learning techniques significantly improve predictive accuracy compared to traditional single classifiers.

The results confirm that combining multiple classifiers enhances robustness and stability in air quality prediction. The proposed framework can be effectively utilized for environmental monitoring and early pollution warning systems in metropolitan areas. By providing accurate and timely air quality predictions, the system contributes to improved public health protection and environmental management.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503043|

IX. ACKNOWLEDGEMENT

The successful completion of this project was made possible through the support and guidance of several individuals. I would like to express my sincere gratitude to everyone who contributed to the development of this air quality prediction system.

First and foremost, I extend my heartfelt thanks to my guide, **Mr. P.Lokesh Kumar, Assistant Professor, Department of Computer Science and Engineering, Kuppam Engineering College**, for his continuous guidance, valuable suggestions, and constant encouragement throughout the project. Her expertise in machine learning and her timely feedback played a crucial role in the successful completion of this work. I am deeply grateful for her patience and motivation.

I would also like to thank the **Head of the Department** and the **Principal of Kuppam Engineering College** for providing a supportive academic environment and the necessary facilities to carry out this project successfully.

I am thankful to the developers of open-source libraries such as **Scikit-learn, XGBoost, NumPy, and Pandas**, which were extensively used for data processing and model implementation in this project. Their tools significantly helped in building and evaluating the machine learning models for air quality prediction.

Finally, I express my sincere gratitude to all the faculty members and friends who directly or indirectly supported and encouraged me throughout the completion of this project.

REFERENCES

1. Machine Learning & Algorithms

• **Breiman, L. (2001).** Random Forests. Machine Learning, 45(1), 5–32. (Foundational work on ensemble learning used for classification and prediction in air quality analysis). **Link:**https://link.springer.com/article/10.1023/A:1010933404324

• **Chen, T., &Guestrin, C. (2016).**XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. (Core reference for XGBoost algorithm used for high-accuracy air quality prediction). **Link:**https://dl.acm.org/doi/10.1145/2939672.2939785

Pedregosa, F., et al. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. (Reference for the Python machine learning library used to implement Gradient Boosting, AdaBoost, Voting, and Stacking classifiers). **Link:**https://jmlr.org/papers/v12/pedregosa11a.html

• **Müller, A. C., & Guido, S. (2016).** Introduction to Machine Learning with Python. O'Reilly Media. (Guide for data preprocessing, model training, and performance evaluation in air quality prediction). **Link:**https://www.oreilly.com/library/view/introduction-to-machine/9781449369880/

2. Air Quality & Environmental Prediction

• **World Health Organization (WHO). (2018).** Ambient (Outdoor) Air Quality and Health. [Online]. (Primary reference for the importance of air quality monitoring and its impact on public health). **Link:**https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health

• **Zhang, Y., et al. (2018).** Prediction of Air Quality Index Based on Machine Learning Methods. Atmospheric Environment.
(Research work demonstrating the use of machine learning models for predicting AQI levels). **Link:**https://www.sciencedirect.com/science/article/pii/S1352231018301537

• **Li, X., et al. (2017).** Air Quality Prediction Using Machine Learning Algorithms. IEEE Access. (Study highlighting the effectiveness of ensemble learning techniques in air pollution forecasting). **Link:**https://ieeexplore.ieee.org/document/7934257

• **U.S. Environmental Protection Agency (EPA).** Air Quality System (AQS) Data Mart. [Online]. (Source of standardized air pollution datasets used for research and model training). **Link:**https://www.epa.gov/aqs

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503043|

3. Data Processing & Visualization Tools

• **Pandas Development Team (2020).** Pandas-dev/pandas: Pandas. [Online]. (Library used for data cleaning, preprocessing, and feature extraction in air quality datasets). **Link:**https://pandas.pydata.org

• **Harris, C. R., et al. (2020).** Array Programming with NumPy. Nature. (Reference for numerical computation and handling large environmental datasets). **Link:**https://www.nature.com/articles/s41586-020-2649-2

• **Hunter, J. D. (2007).**Matplotlib: A 2D Graphics Environment. Computing in Science & Engineering. (Used for visualizing air quality trends and prediction results). **Link:**https://ieeexplore.ieee.org/document/4160265

• **OpenAQ Platform.** Open Air Quality Data. [Online]. (Global open-source air quality data platform used for research and analysis). **Link:**https://openaq.org

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details