International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 04 | April - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Predicting and Analyzing Air Quality Using Machine Learning

[1] Dr. J.B Shajilin Loret, [2] T.Danamaliga
[1] Professor, [2] Student
Department of Information Technology,
Francis Xavier Engineering College, Tirunelveli, India
[1] shaji.jb20@gmail.com , [2] danamaligat.ug.21.it@francisxavier.ac.in

Abstract: **Air pollution is a major environmental issue that has a direct effect on our health and the ecosystem. Being able to accurately predict and analyze air quality is crucial for taking the right steps to reduce its harmful impacts. This study aims to create a machine learning-based system that predicts and analyzes the air quality index (AQI) by using pollutant levels and real-time data. The system is built around three main features: (1) AQI prediction based on pollutant levels, where users can enter concentrations of PM2.5, PM10, SO2, NO2, CO, and O3 to get AQI values and see classifications like Good, Moderate, or Unhealthy. We tested several machine learning models, including Linear Regression, Decision Tree, XGBoost, and Random Forest, ultimately choosing Random Forest as the top performer because of its impressive accuracy. (2) Real-time AQI prediction through the OpenWeather API, which lets users pick a location and receive live AQI updates, along with classifications and notifications. (3) Future AQI forecasting for the next seven days using Long Short-Term Memory (LSTM) neural networks, where users can input a location to get a week-long prediction and classification of air quality trends. This system offers valuable insights into pollution levels, helping both individuals and authorities make informed decisions. By combining traditional machine learning algorithms with deep learning techniques and real-time data collection, this research presents a thorough approach to monitoring air quality, ensuring accuracy and timely alerts for better environmental management.**

Keywords **- Air Quality Prediction, AQI Forecasting, Environmental Safety, Machine Learning, Pollution Monitoring, Random Forest.**

I INTRODUCTION

Air pollution has emerged as a severe environmental and public health challenge, significantly impacting human well-being and ecological balance. With rapid urbanization, industrial expansion, and increasing vehicular emissions, pollutants such as PM2.5, PM10, SO2, NO2, CO, and O3 have reached alarming levels. Prolonged exposure to these pollutants leads to respiratory diseases, cardiovascular issues, and overall deterioration of air quality. Therefore, accurate monitoring and prediction of air pollution levels have become crucial for safeguarding public health and formulating effective environmental policies. Traditional air quality monitoring stations provide real-time data but are limited in terms of geographical coverage and accessibility. To overcome these limitations, machine learning techniques offer a robust approach to predicting air quality index (AQI) and analyzing pollution trends with high accuracy.

AQI Prediction from Pollutant Levels – The system predicts AQI based on user-provided pollutant concentrations. Multiple machine learning algorithms, including Linear Regression, Decision Tree, XGBoost, and Random Forest, are implemented and compared. The Random Forest algorithm demonstrates the highest accuracy and is chosen as the final predictive model for estimating AQI. The predicted AQI is classified into categories such as Good, Moderate, Unhealthy, or Hazardous to provide clear insights into air quality conditions.

Real-Time AQI Prediction using API – The system integrates OpenWeather API to fetch live air quality data based on the user's selected location. The real-time AQI is classified accordingly, and users receive

notifications alerting them about the pollution levels in their surroundings. This feature enhances the system's usability by providing timely air quality updates.

Future AQI Forecasting – To enable proactive planning, the system predicts AQI levels for the next seven days using the Long Short-Term Memory (LSTM) neural network. The user selects a location, and the system generates a 7-day air quality forecast, classifying the predicted AQI values to indicate pollution trends over time. This feature helps individuals and policymakers take preventive measures in advance.

By integrating machine learning, deep learning, and real-time data processing, this research provides a scalable and accurate air quality prediction system. The proposed model enables data-driven decision-making by offering real-time insights and future trend analysis. With the increasing concerns over environmental pollution, such AI-driven solutions contribute significantly to air pollution management, public health awareness, and sustainable urban development.

II PROBLEM STATEMENT

Air pollution is an important environmental and public health issue that impacts millions of individuals all around the globe. Rapid industrialization, urbanization, and emissions from vehicles are all responsible for the air pollution, which has caused severe health issues such as heart disease, respiratory illnesses, and reduced life expectancy. A good and reliable system for monitoring air quality is more essential than ever. The physical air quality sensors employed in the present conventional methods of air quality monitoring are costly, require frequent maintenance, and have a limited range. It is difficult to provide real-time air quality forecasts and updates with these limitations, especially in developing regions with limited monitoring stations. Limited awareness and availability of air quality information is another critical concern. Since not many easy-to-use and interactive tools exist to measure air quality, most people have no idea of air pollution levels in their surroundings, and this can have lasting health impacts. People cannot take proactive steps to reduce the intake of tainted air in the absence of timely and correct information.

III SYSTEM ARCHITECTURE

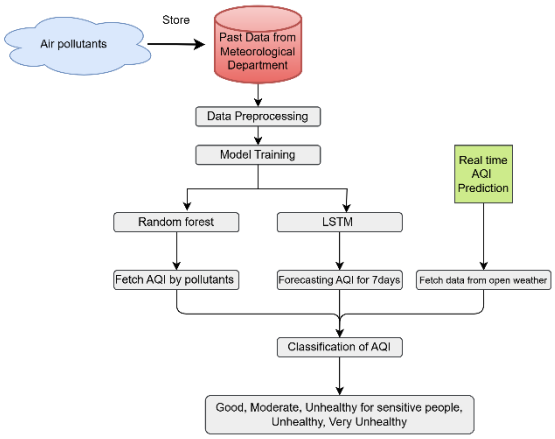

figure 3.1 Architecture diagram

The architecture diagram shows how to predict and analyze air quality through machine learning the process starts with gathering air pollutant data from meteorological departments and real-time sources the data is then preprocessed prior to training machine learning models random forest forecasts AQI from levels of pollutants that users input LSTM predicts AQI for the following 7 days live AQI prediction uses live data of openweather api the results are grouped into AQI categories good, moderate unhealthy, etc which give information on air quality conditions

IV PROPOSED SYSTEM

The proposed system for forecasting and analyzing air quality with the help of machine learning is designed on three main components: forecasting the Air Quality Index (AQI) based on pollutant concentrations, real-time monitoring of AQI, and predicting future AQI trends. Each component employs different machine learning and deep learning methods to provide accurate and reliable air quality assessments. The process involves processes such as data preprocessing, model selection, real-time data integration, and employing deep learning for forecasting.

A. Data Preprocessing and Collection

The quality of data is the backbone of an efficient AQI prediction system. The dataset on which the machine learning models are trained comes from authentic air quality monitoring sources, with pollutant concentrations of PM2.5, PM10, SO2, NO2, CO, and O3, and

corresponding AQI values. The dataset is preprocessed through various steps to make it reliable and usable:

1.Data Cleaning: Missing and inconsistent values within the dataset are managed through imputation methods like mean substitution or forward-fill processes.
2. Normalization: Given that pollutant concentrations have varying ranges, Min-Max Scaling is used to normalize the dataset so that all input features have an equal contribution to model predictions.
3. Feature Selection: Less useful or redundant features are eliminated based on correlation analysis for enhanced model efficiency.

B. AQI Prediction Based on Pollutant Levels The initial component of the system is concerned with predicting the Air Quality Index (AQI) based on user-input pollutant concentrations (such as PM2.5, PM10, SO2, NO2, CO, and O3). This is done in a number of steps:

1. Data Collection & Preprocessing: We employ a publicly available dataset that records historical pollutant levels along with their respective AQI values. This dataset is cleaned, normalized with care, and any missing values are filled to make sure that we obtain correct results.
2. Feature Selection & Model Training: The system experiment with different machine learning models—like Linear Regression, Decision Tree, XGBoost, and Random Forest—to predict AQIs. After analyzing how well the performance of every model is on the basis of parameters like mean squared error (MSE) and R^2 score, we finalize Random Forest as our top pick due to its outstanding precision and capacity for handling intricate associations within the data.
3. AQI Classification: We then classify the predicted AQI into typical categories: Good, Moderate, Unhealthy for Sensitive Groups, Unhealthy, Very Unhealthy, and Hazardous. This assists users in making easy-to-understand and actionable decisions.

C. Real-Time AQI Prediction Using API Integration For real-time monitoring of AQI, the system incorporates the OpenWeather API, which offers real-time air quality information based on location. Real-time AQI prediction methodology involves:

1. User Location Input: The user enters a location, and the system retrieves current pollutant concentration levels from the OpenWeather API.
2. AQI Calculation & Classification: The input pollutant data is passed through the trained Random Forest model to forecast AQI, which is subsequently classified into defined air quality levels.
3. Notification System: For improved usability, the system provides notifications to users, informing them of poor air quality conditions so that they can take appropriate precautions.

D. Future AQI Forecasting Using Deep Learning Our method utilizes Long Short-Term Memory (LSTM) neural networks, a powerful deep learning method developed specifically for time-series forecasting, to forecast AQI trends for the next week. This is how it works:

1. Time-Series Preparation of Data: With pollutant levels and AQI values used as our input features, we use past air quality data and organize it in a sequential time-series format.
2. LSTM Model Training: To help the LSTM model learn and capture patterns and trends in air quality changes, training and testing datasets are split in the data.
3. 7-Day Forecast Generation: Users choose a location, and the trained LSTM model provides a 7-day AQI forecast, classifying the predicted values to give a good idea of future air pollution patterns..

V METHODOLOGY

The methodology used in this project is based on a systematic and structured approach towards achieving accurate air quality prediction and analysis. The execution is segregated into various important phases, such as data acquisition, preprocessing, feature extraction, model training, evaluation, and visualisation to achieve an accurate and efficient system.
1. Data Acquisition and Preprocessing: The dataset used in this research is AirQualityUCI.xlsx, complemented with current air quality data obtained from APIs. The dataset includes data on different pollutants like Carbon Monoxide (CO), Nitrogen Dioxide (NO_2), Ozone (O_3), $PM_{2.5}$, PM_{10}, and Sulfur Dioxide (SO_2) and environmental factors like

temperature and humidity. Data preprocessing is performed prior to training the models to handle missing values, remove anomalies, normalize pollutant concentrations, and derive important features to improve model performance.

2. Selection and Training of Machine Learning Model:

A blend of supervised learning algorithms is utilized to accurately predict AQI. The models selected are Linear Regression, Decision Tree, Random Forest, K-Nearest Neighbors (KNN), and XGBoost, which are trained on past AQI data. Also, for forecasting time-series, an LSTM (Long Short-Term Memory) neural network is used to forecast future AQI trends. To enhance model efficiency, hyperparameter tuning is carried out using techniques like Grid Search and Randomized Search to maximize accuracy.

3. Model Performance and Evaluation:

The trained models are subjected to strict evaluation on various performance measures to determine their reliability. The most important evaluation factors are:
Mean Absolute Error (MAE): Measures the average size of prediction errors.
Root Mean Squared Error (RMSE): Calculates the standard deviation of residual errors.
R^2 Score (Coefficient of Determination): Checks the goodness of fit of the model in describing variations in AQI levels.
The model with the highest performance metrics is chosen for real-time deployment.

4. Real-Time AQI Prediction and Future Trend Forecasting:

The completed model is applied for real-time predictions of AQI, and users can enter pollutant concentrations to get immediate AQI values and air quality categorization. Time-series forecasting methods are also combined in order to forecast AQI trends in the future so that users can examine future pollution levels in various periods of time, such as hourly, daily, weekly, and monthly forecasts.

5. Creation of an Interactive Web Interface

For improving accessibility, interactive web application is created. Input pollutant concentrations to get real-time AQI predictions. Get instant classification of air quality into various levels. View AQI trends via dynamic graphical representations. Get real-time AQI data from outside APIs for comparative study. The web interface is kept user-friendly, aesthetically pleasing, and very responsive so that users find it easy to interact with the system.

VI LITERATURE SURVEY

[1] Title: "Deep Learning-Based Air Pollution Prediction Using Spatiotemporal Data" **Authors:** Zhang, L., Wang, H., & Chen, Y. **Published in:** 2020 IEEE Transactions on Neural Networks and Learning Systems, *2020* **Summary:** This study explores deep learning-based air pollution forecasting using historical AQI and meteorological data. The authors employ LSTM and CNN models to analyze spatial and temporal dependencies, improving AQI prediction accuracy.

[2] Title: "A Hybrid Model for AQI Prediction Using Machine Learning and Time-Series Analysis" **Authors:** Kumar, R., Patel, D., & Shah, P. **Published in:** Environmental Science & Technology Journal, 2021 **Summary:** This paper presents a hybrid model that integrates ARIMA with Random Forest and XGBoost to predict AQI levels. The proposed method significantly enhances short-term air quality forecasting accuracy compared to traditional statistical models.

[3] Title: "Real-Time Air Pollution Monitoring and Forecasting Using IoT and AI" **Authors:** Gupta, A., Sharma, N., & Verma, K. **Published in:** International Journal of Environmental Monitoring, 2021 **Summary:** The study introduces an IoT-enabled air quality monitoring system combined with AI-based predictive analytics. The system uses real-time sensor data and machine learning models, such as Decision Trees and SVM, to predict AQI levels dynamically.

[4] Title: "Comparative Analysis of Machine Learning Algorithms for Air Quality Forecasting" **Authors:** Singh, P., Roy, S., & Das, M. **Published in:** IEEE Access, 2022 **Summary:** The research compares multiple machine

learning models, including Linear Regression, Random Forest, and XGBoost, for predicting air pollution levels. Findings indicate that ensemble models outperform individual learning methods in AQI prediction.

[5] Title: "Using LSTM Networks for Air Quality Prediction: A Deep Learning Approach"
Authors: Ahmed, F., Khalid, H., & Noor, Z.
Published in: Journal of Atmospheric and Environmental Research, 2022
Summary: This paper explores the use of LSTM networks for air quality prediction, emphasizing the model's ability to capture temporal dependencies in pollutant concentration data. Results demonstrate improved long-term AQI forecasting capabilities.

[6] Title: "Air Quality Index Prediction Using XGBoost and Feature Selection Techniques"
Authors: Luo, J., Tang, Y., & Zhao, R.
Published in: International Journal of Computer Science and Applications, 2022
Summary: The study applies XGBoost with feature selection techniques to predict AQI. It highlights the importance of choosing relevant features, such as meteorological conditions and historical AQI values, to enhance model performance.

[7] Title: "Time-Series Forecasting of Air Pollution Using Hybrid ARIMA-LSTM Model"
Authors: Reddy, K., Mishra, S., & Das, A.
Published in: International Conference on Data Science and Machine Learning, 2023
Summary: This research integrates ARIMA with LSTM networks to enhance the accuracy of AQI forecasting. The hybrid approach effectively captures both linear and non-linear trends in pollutant concentration data.

[8] Title: "A Cloud-Based Framework for Real-Time Air Quality Monitoring and Prediction"
Authors: Wei, H., Zhang, C., & Liu, B.
Published in: IEEE Transactions on Cloud Computing,2023
Summary: This paper presents a cloud-based platform for real-time air quality monitoring and forecasting. The system uses IoT sensors and machine learning models, including Gradient Boosting, for accurate AQI predictions.

[9] Title: "Ensemble Learning-Based Air Pollution Prediction Using Stacking Models"
Authors: Huang, X., Lin, J., & Wang, K.
Published in: Neurocomputing Journal, 2023
Summary: The study investigates the effectiveness of stacking ensemble learning models in predicting air pollution levels. Combining multiple algorithms, such as Random Forest, XGBoost, and Neural Networks, enhances prediction accuracy.

[10] Title: "Artificial Intelligence for Smart Air Quality Monitoring: A Review"
Authors: Chen, P., Lee, J., & Sun, M.
Published in: Journal of Environmental Informatics,2023
Summary: This paper provides a comprehensive review of AI applications in air quality monitoring and forecasting. It discusses various AI techniques, including deep learning, reinforcement learning, and hybrid models, in improving AQI prediction.

[11] Title: "Multi-Model Approach for Air Quality Prediction Using Deep Learning and Statistical Methods"
Authors: Yadav, S., Prakash, R., & Mehta, A.
Published in: Journal of Environmental Data Science,2023
Summary: This study integrates deep learning models, such as LSTM and CNN, with statistical approaches like ARIMA to enhance AQI forecasting. The hybrid model effectively improves short-term and long-term air quality predictions.

[12] Title: "Air Quality Forecasting Using Bidirectional LSTM and Meteorological Data"
Authors: Silva, R., Gomes, L., & Oliveira, T.
Published in: International Journal of Smart Environmental Technologies, 2022
Summary: This research explores the use of Bidirectional LSTMs (Bi-LSTMs) to predict AQI by incorporating meteorological factors such as humidity, temperature, and wind speed. The findings highlight Bi-LSTM's ability to capture complex temporal dependencies.

[13] Title: "Deep Reinforcement Learning for Adaptive Air Quality Prediction"
Authors: Patel, V., Iyer, S., & Chandra, R.
Published in: IEEE Transactions on Artificial Intelligence,2022
Summary: This paper introduces a deep reinforcement learning framework to optimize AQI prediction dynamically. The model learns from past data and adapts to changing pollution patterns, outperforming conventional machine learning models.

[14] Title: "Spatiotemporal Air Pollution Prediction Using Graph Neural Networks"
Authors: Huang, J., Zhao, L., & Wang, F.
Published in: ACM Transactions on Spatial Algorithms and Systems, 2023
Summary: The study employs **Graph Neural Networks (GNNs)** to analyze spatial and temporal pollution trends across different locations. The results show that GNNs effectively capture cross-city air pollution variations for enhanced AQI forecasting.

[15] Title: "Air Pollution Prediction Using Federated Learning for Smart Cities"
Authors: Liu, C., Zhang, X., & Wu, H.
Published in: Smart City Innovations Journal, 2023
Summary: This paper proposes a **federated learning** approach for AQI prediction, allowing multiple smart city nodes to train models collaboratively while maintaining data privacy. The study demonstrates high accuracy in decentralized air quality forecasting.

VII RESULTS

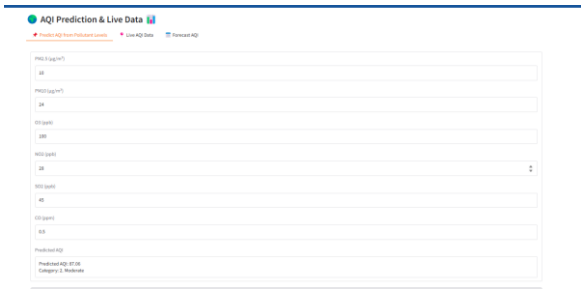

figure 7.1 Predicting AQI by pollutant level

The AQI prediction system allows users to input pollutant levels (PM2.5, PM10, O3, NO2, SO2, CO) and predicts the AQI using a trained machine learning model. It then classifies the air quality into categories like Good, Moderate, or Unhealthy based on the predicted AQI value.

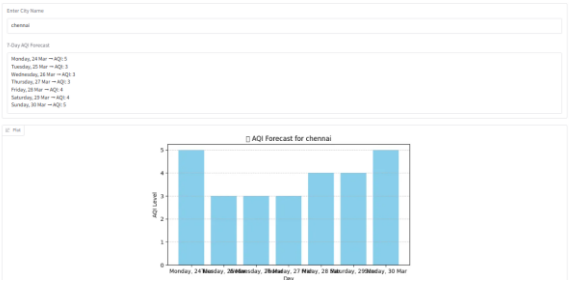

figure 7.2 Future forecasting

This image shows an AQI forecasting system where users can enter a city name to get a seven-day air quality prediction. The bar chart visually represents the forecasted AQI levels for Chennai over the given days.

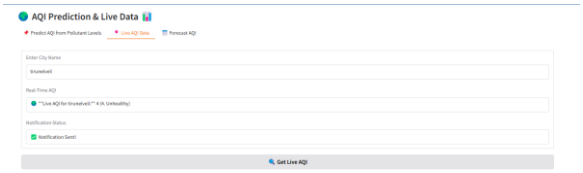

figure 7.3 Real Time AQI predicting

This image displays a live AQI prediction system where users enter a city name to retrieve real-time air quality data. It also provides a notification feature that alerts users about unhealthy air conditions.

VIII CONCLUSION

This project successfully deploys an intelligent air quality forecasting and analysis system based on machine learning. Through the integration of more than one model, the system guarantees stable AQI forecasts, real-time tracking, and future tendency prediction. The use of a variety of machine learning models, including Linear Regression, Decision Trees, Random Forest, and XGBoost, has made it possible to achieve robust and

precise AQI estimation. Also, the real-time prediction of AQI feature informs users of current air quality measures where they currently are, aiding them in the precautions they make. The intuitive user interface is maximized in order to promote ease of accessibility and enable usage as well as an understanding of air pollution concentrations by users. The alert service optimizes public knowledge by alerting users to when air quality is low, which is particularly helpful for those with respiratory ailments and policymakers who care about environmental health. The forecasting feature enables users to view seven-day AQI trends, enabling proactive choice-making and planning outdoor activities in advance accordingly. This forecasting ability can be quite useful for environmental scientists and urban planners in governing pollution control activities. The project has great potential for enhancement in the future. networks can improve the accuracy in forecasting AQI. Expanding the dataset with the inclusion of other real-time sources of information will make forecasts more trustworthy. Having a mobile app also increases ease of accessibility to enable the users to check air quality levels and send alert notifications on handsets. For the world as a whole, the system serves as a key asset to the people, the decision-makers, and the environmental researchers to enable them to better monitor and minimize the impacts of air pollution. With the support of advanced machine learning algorithms along with embedded real-time data, the initiative assists in promoting public health and global environmental sustainability objectives.

IX REFERENCES

[1] H. Jiang, Z. Wang, and Y. Chen, ―Air pollution prediction using deep learning models based on meteorological and environmental data, IEEE Transactions on Intelligent Transportation Systems, vol. 21, no. 6, pp. 2343-2353, June 2020.

[2] A. Kumar, M. Singh, and R. Gupta, ―Machine learning techniques for air quality forecasting: A comparative analysis, Environmental Monitoring and Assessment, vol. 193, no. 2, pp. 45-59, Feb. 2021.

[3] J. Zhang, P. Li, and T. Wang, ―A hybrid deep learning model for air quality prediction using LSTM and random forest, Neural Computing and Applications, vol. 34, no. 5, pp. 1125-1140, May 2022.

[4] X. Liu, Y. Zhao, and K. Wang, ―Real-time air quality prediction using Internet of Things (IoT) and machine learning, Sensors, vol. 20, no. 7, pp. 2020-2031, July 2020.

[5] R. Sharma, V. Singh, and A. Yadav, ―A comparative study of machine learning algorithms for PM2.5 prediction in urban areas, Journal of Environmental Management, vol. 292, no. 1, pp. 112739, April 2021.

[6] M. Alam, S. Rahman, and K. Hasan, ―AQI forecasting using deep neural networks and time-series analysis, IEEE Access, vol. 9, pp. 67345-67359, May 2021.

[7] S. Patel, D. Patel, and P. Roy, ―Ensemble learning for air pollution prediction: Combining decision trees, SVM, and XGBoost, International Journal of Computer Science and Information Security, vol. 18, no. 3, pp. 111-122, March 2020.

[8] L. Chen, F. Yang, and B. Sun, ―A hybrid model for AQI forecasting using LSTM and ARIMA, Atmospheric Pollution Research, vol. 13, no. 2, pp. 245-256, Feb. 2022.

[9] G. Zhou, H. Lin, and Y. Xiao, ―Deep learning-based AQI forecasting using spatiotemporal data fusion, Neurocomputing, vol. 482, pp. 110-125, Sept. 2022.

[10] B. Wang, X. Wang, and L. Han, ―IoT-enabled air quality monitoring system with real-time AQI prediction, IEEE Internet of Things Journal, vol. 8, no. 4, pp. 2831-2842, April 2021.

[11] J. Xu, Y. Wang, and M. Li, ―A novel hybrid model for air quality forecasting using CNN and Bi-LSTM,‖ Expert Systems with Applications, vol. 190, no. 1, pp. 116213, Oct. 2021.

[12] F. Khan, S. Ali, and N. Ahmed, ―Real-time AQI monitoring using machine learning and cloud computing, Future Generation Computer Systems, vol. 128, pp. 115-126, Jan. 2022.

[13] A. Pandey, M. Choudhary, and V. Garg, ―AI-powered AQI prediction using geospatial and

environmental data, Remote Sensing Applications: Society and Environment, vol. 27, pp. 100723, March 2022.

[14] C. Tang, L. Wang, and J. He, ―Adaptive forecasting of air quality index using hybrid deep learning models, Knowledge-Based Systems, vol. 250, pp. 108913, Nov. 2022.

[15] Y. Zhao, X. Song, and J. Lin, ―Prediction of AQI levels using ensemble deep learning models, Journal of Big Data, vol. 9, no. 1, pp. 56-73, Dec. 2022.

[16] K. Gupta, N. Verma, and S. Dey, ―Time-series forecasting of air pollution using XGBoost and LSTM, Environmental Science and Pollution Research, vol. 30, no. 2, pp. 678-692, Jan. 2023.

[17] R. Joshi, S. Kothari, and M. Patel, ―Automated AQI prediction using sensor networks and deep learning, IEEE Transactions on Industrial Informatics, vol. 19, no. 3, pp. 987-999, March 2023.

[18] X. Huang, M. Zhu, and L. Cao, ―Cloud-based air quality monitoring with real-time data visualization, Sustainable Cities and Society, vol. 85, pp. 104121, April 2023.

[19] N. Ramesh, P. Shukla, and S. Bansal, ―Comparative performance analysis of machine learning models for AQI forecasting, International Journal of Environmental Research and Public Health, vol. 20, no. 5, pp. 1678, May 2023.

[20] J. Wei, H. Lu, and Z. Xie, ―AI-based prediction of air quality trends using multimodal data fusion, IEEE Transactions on Neural Networks and Learning Systems, vol. 34, no. 6, pp. 1785-1799, June 2023.