Air Quality Prediction using Machine Learning

[1] S.Anu Priya, [2] B.Naga Saikiran, [3] B.Adi Mahesh, [4] B.Sita Ramudu, [5] CH.Deekshith

[1]*Assistant Professor, Department of Computer Science and Engineering, Bharath Institute of Science & Technology (BIST), 173, Agaram Road, Selaiyur, Chennai-600 073*

[2,3,4,5]*UG Student, Department of Computer Science and Engineering, Bharath Institute of Science & Technology (BIST), 173, Agaram Road, Selaiyur, Chennai-600 073*

Abstract— **Air pollution has become one of the most serious environmental challenges affecting human health and ecological balance across the world. Rapid urbanization, industrial activities, and increasing vehicular emissions have significantly contributed to the deterioration of air quality in many cities. Continuous exposure to polluted air can lead to severe health problems such as respiratory diseases, cardiovascular disorders, and reduced life expectancy. Therefore, predicting air pollution levels in advance is essential for enabling authorities and individuals to take preventive actions and reduce the harmful effects of poor air quality. This study presents a machine learning–based approach for predicting the Air Quality Index (AQI) using historical environmental data. The system analyzes various pollution parameters including PM2.5, PM10, nitrogen dioxide (NO_2), carbon monoxide (CO), sulfur dioxide (SO_2), temperature, and humidity. The dataset is preprocessed through data cleaning, normalization, and feature selection to improve model performance. Several machine learning algorithms such as Linear Regression, Random Forest, and XGBoost are applied to learn patterns from historical data and generate accurate AQI predictions. The performance of the proposed models is evaluated using standard evaluation metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and R^2 score. Experimental results demonstrate that machine learning models can effectively capture relationships between environmental factors and air pollution levels, providing reliable forecasts of future air quality conditions. The developed system can assist environmental authorities and the general public by providing early warnings of pollution levels, thereby supporting better environmental monitoring and decision- making.**

Keywords - Machine Learning, Air Quality Index (AQI), Random Forest, XGBoost, Linear Regression, Pollution Forecasting, PM2.5 and PM10 Analysis, Environmental Data Analysis

I.INTRODUCTION

Air pollution has become one of the most critical environmental problems affecting human health and ecological balance worldwide. Rapid industrial growth, urbanization, and increasing vehicular emissions have significantly contributed to the deterioration of air quality in many cities. Polluted air contains harmful substances such as particulate matter, nitrogen oxides, carbon monoxide, and sulfur dioxide, which can cause severe respiratory and cardiovascular diseases.

Air quality monitoring has become essential for understanding pollution levels and protecting public health. Governments and environmental agencies install monitoring stations to measure pollutants and calculate the Air Quality Index (AQI). The AQI is an important indicator that reflects the level of air pollution and its potential impact on human health. Traditional air monitoring systems mainly focus on measuring the current level of pollutants in the atmosphere. Although these systems provide valuable real-time information, they are not capable of accurately predicting future air quality conditions. Without forecasting capabilities, it becomes difficult for authorities and citizens to take preventive actions in advance.

With the rapid growth of data science and artificial intelligence, Machine Learning has emerged as a powerful tool for analyzing large environmental datasets. Machine learning algorithms can identify hidden patterns in historical data and use them to predict future outcomes. This capability makes machine learning highly suitable for environmental prediction problems such as air quality forecasting. Air pollution prediction involves analyzing various environmental factors that influence pollutant concentration. These factors include meteorological parameters such as temperature, humidity, wind speed, and atmospheric pressure. In addition, pollutants like PM2.5, PM10, nitrogen dioxide (NO_2), carbon monoxide (CO), and sulfur dioxide (SO_2) play a major role in determining overall air quality.

Particulate matter such as PM2.5 and PM10 are among the most harmful pollutants affecting human health. These microscopic particles can penetrate deep into the lungs and bloodstream, causing respiratory illnesses and heart problems. Monitoring and predicting particulate matter concentration is therefore essential for maintaining public health safety.

Machine learning techniques enable researchers to analyze large volumes of historical air quality data collected from monitoring stations. By training models on past pollution data, it is possible to identify relationships between different environmental variables and pollutant concentrations. These relationships help build predictive models capable of estimating future air pollution levels. Data pre-processing plays a critical role in building accurate machine learning models. Environmental datasets often contain missing values, noise, and inconsistent measurements. Proper pre-processing techniques such as data cleaning, normalization, and feature selection are necessary to improve the reliability and accuracy of the predictive models.

Feature selection is another important step in air quality prediction. Not all environmental parameters contribute equally to pollution levels. Identifying the most relevant features helps reduce model complexity and improves prediction performance. This process ensures that the machine learning models focus on the most influential variables. Several machine learning algorithms have been successfully applied to air quality prediction problems. Regression algorithms are particularly suitable because air quality prediction involves estimating continuous values such as AQI. These algorithms learn patterns from historical data and generate predictions for future conditions.

Linear Regression is one of the most commonly used algorithms for predicting numerical values. It establishes a linear relationship between input features and the target variable. Although simple, it provides a good baseline model for understanding the relationship between environmental parameters and AQI levels. Random Forest is another widely used machine learning algorithm for environmental prediction tasks. It is an ensemble learning technique that builds multiple decision trees and combines their outputs to improve prediction accuracy. Random Forest models are known for handling complex data relationships and reducing overfitting.

XGBoost (Extreme Gradient Boosting) is a powerful boosting algorithm that has gained popularity in many data science applications. It improves prediction performance by sequentially correcting the errors of previous models. XGBoost is highly efficient and often provides better accuracy compared to traditional algorithms. Comparing multiple machine learning models helps identify the most suitable algorithm for air quality prediction. Different models may perform differently depending on the dataset characteristics. Evaluating their performance ensures that the most accurate and reliable model is selected for forecasting air pollution levels.

Model evaluation is an essential part of machine learning system development. Various performance metrics are used to assess the accuracy of predictive models. Common evaluation metrics for regression problems include Mean Absolute Error (MAE), Mean Squared Error (MSE), and the coefficient of determination (R^2 score).

The Mean Absolute Error measures the average difference between predicted and actual AQI values. A lower MAE value indicates better prediction accuracy. Similarly, Mean Squared Error measures the average squared difference between predicted and actual values, emphasizing larger errors. The R^2 score is another important metric that indicates how well the model explains the variation in the target variable. An R^2 value closer to 1 indicates a better model fit. Using these evaluation metrics helps ensure that the developed machine learning system produces reliable predictions.

Accurate air quality prediction systems can provide significant benefits for society. Early warnings about high pollution levels allow governments to implement control measures such as traffic restrictions or industrial emission reductions. These preventive actions can reduce health risks for the population. In addition, air quality forecasting systems can help individuals plan their daily activities more safely. People with respiratory problems or other health conditions can take precautions when pollution levels are expected to rise. This improves public awareness and promotes healthier living conditions.

II. REVIEW OF LITERATURE

Environmental datasets often suffer from missing or incomplete values due to sensor failures or data transmission issues. Techniques for imputing missing data, such as statistical interpolation and machine learning-based imputation, have been proposed to enhance data quality. Improving dataset reliability is essential for building accurate prediction models and ensuring consistent environmental monitoring [1].

Early studies in air quality modeling focused on analyzing environmental data using structured and rule-based approaches. These models helped in understanding the relationships between atmospheric variables such as temperature, humidity, and pollutant concentrations. Such foundational work laid the groundwork for modern data- driven air quality prediction systems [2].

Traditional time-series forecasting methods such as Autoregressive Integrated Moving Average (ARIMA) have been widely used to model temporal patterns in air pollution data. These models are effective in identifying trends and seasonality but are limited in capturing nonlinear relationships and complex interactions among environmental variables [3].

To overcome the limitations of traditional models, hybrid approaches combining optimization techniques such as Genetic Algorithms (GA) and Particle Swarm Optimization (PSO) have been introduced. These techniques help optimize model parameters and improve prediction accuracy, particularly in complex environmental forecasting tasks [4].

Machine learning methods have been applied to predict pollutant concentrations such as PM2.5 by learning patterns from historical environmental data. These approaches provide better generalization compared to traditional models and can handle large datasets with multiple influencing factors [5].

Several machine learning algorithms, including regression models and decision trees, have been used to predict air quality levels. These models are capable of capturing nonlinear relationships between input features and pollutant concentrations, resulting in improved prediction performance compared to statistical methods [6].

Deep learning techniques have further enhanced air quality prediction by enabling models to learn hierarchical representations of data. These models can automatically extract relevant features from large datasets, improving accuracy and reducing the need for manual feature engineering [7].

Advanced deep learning frameworks have been developed to improve model generalization and robustness. These frameworks utilize multiple layers and optimization techniques to capture complex patterns in environmental data, leading to more accurate and reliable predictions [8].

Random Forest, an ensemble learning technique, has been widely used in air quality prediction due to its ability to combine multiple decision trees and reduce overfitting. It provides high accuracy and stability, especially when dealing with large and complex environmental datasets [9].

XGBoost, a gradient boosting algorithm, has gained popularity for its efficiency and high predictive performance. It improves model accuracy by iteratively correcting errors and optimizing the learning process, making it suitable for large-scale air quality datasets [10].

Support Vector Machines (SVM) have been used for air pollution prediction due to their ability to handle nonlinear relationships and high-dimensional feature spaces. SVM models are effective when trained with properly preprocessed data and appropriate kernel functions [11].

Machine learning development tools and frameworks have simplified the implementation of predictive models. These tools provide efficient libraries and techniques for data preprocessing, model training, and evaluation, enabling faster development of air quality prediction systems [12].

Long Short-Term Memory (LSTM) networks have been widely used for time-series forecasting because they can capture long-term dependencies in sequential data. LSTM models are particularly effective for predicting air pollution trends over time and outperform traditional models like ARIMA [13].

Recent deep learning approaches integrate multiple environmental factors, including meteorological data and pollutant concentrations, to improve prediction accuracy. These models provide a more comprehensive understanding of pollution dynamics by considering multiple influencing variables [14].

Machine learning-based air quality forecasting systems have been developed to analyze pollution trends and provide decision support for environmental management. These systems can generate early warnings and help authorities take preventive measures to reduce pollution levels [15].

A Neuro-Fuzzy Temporal Classification Algorithm with spatial constraints (NFTCA-S) is proposed for prediction, while the Butterfly Optimization Algorithm (BOA) is employed for feature optimization, improving accuracy and reducing RMSE [16].

Hybrid deep learning models, such as those combining Convolutional Neural Networks (CNN) with other architectures, have been proposed to capture both spatial and temporal features in air pollution data. These models improve prediction performance by extracting detailed patterns from complex datasets [17].

The integration of big data analytics and urban sensing technologies has enabled large-scale air quality monitoring in smart cities. These systems collect and process data from multiple sources, allowing real-time prediction and better environmental management [18].

Feature selection and data preprocessing techniques play a critical role in improving model performance. Selecting relevant features and removing noise from data helps reduce computational complexity and enhances prediction accuracy [19].

Model selection techniques such as Akaike Information Criterion (AIC) have been used to identify the most suitable models for environmental forecasting. These techniques help balance model complexity and accuracy, ensuring better performance [20].

III. PROPOSED METHOD

The proposed system focuses on developing an efficient and accurate air quality prediction model using machine learning techniques. The system is designed to analyze historical environmental and pollution data to forecast future air quality levels. The proposed approach aims to improve prediction accuracy by utilizing machine learning algorithms that can identify hidden patterns in environmental datasets. The system consists of several stages including data collection, data preprocessing, model training using machine learning algorithms, and air quality prediction. The architecture is designed to effectively learn patterns from historical data and generate reliable predictions that can assist environmental authorities and the public in taking preventive actions against air pollution.

A. Data Collection

The first step in the proposed system is the collection of historical air quality data. The dataset includes various environmental parameters such as PM2.5, PM10, nitrogen dioxide (NO_2), carbon monoxide (CO), sulfur dioxide (SO_2), temperature, and humidity. This data is collected from reliable environmental monitoring sources and air quality monitoring stations. The dataset contains records of pollutant levels measured over a specific time period, which allows the model to analyze pollution patterns. Historical environmental data plays an important role in training the model to understand trends and relationships between different air pollutants and environmental conditions.

B. Data Preprocessing

Data preprocessing is an essential step to prepare the collected air quality data for training the prediction model. In this stage, the dataset is cleaned to remove missing values and inconsistent entries that may affect the model performance. The numerical features are normalized so that all parameters are within a similar range, which improves the efficiency of machine learning algorithms. Feature selection is also performed to identify the most relevant environmental parameters affecting air quality. Proper preprocessing helps improve the reliability and accuracy of the air quality prediction system.

C. Machine Learning Model Architecture

The core component of the proposed system is the machine learning model used for analyzing environmental data and predicting air quality levels. Several regression algorithms such as Linear Regression, Random Forest, and XGBoost are used in this project. These models analyze historical pollutant data and learn patterns that influence air quality levels. The input layer receives the processed environmental features, while the internal model structure identifies relationships between the variables. Ensemble algorithms such as Random Forest and XGBoost combine multiple decision trees to improve prediction accuracy. The trained models are capable of capturing complex relationships between environmental factors and air pollution levels.

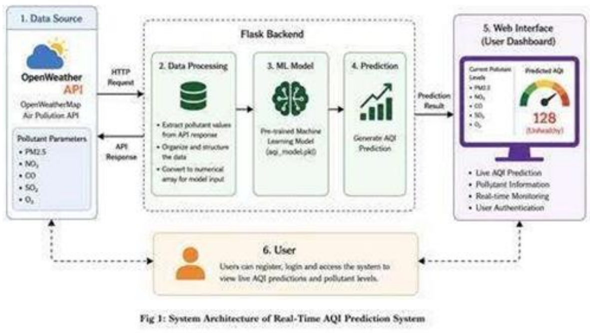

Fig 1: System Architecture

D. Air Quality Prediction

Once the machine learning models are successfully trained, they are used to predict future Air Quality Index (AQI) values. When new environmental data is provided as input, the trained model analyzes the patterns in pollutant concentrations and meteorological conditions. Based on the learned relationships from historical data, the model predicts the expected AQI level. These predictions can help authorities monitor pollution trends and take preventive measures in advance. The predicted results are compared with actual AQI values to evaluate the performance of the model.

E. System Deployment and Real-Time Monitoring

The final stage of the proposed work involves deploying the trained model for practical use. The air quality prediction system can be integrated into a web-based or dashboard application that allows users to view predicted AQI values and pollution trends. Visualization tools such as graphs and charts can be used to display historical pollution data and predicted air quality levels. This helps environmental agencies and the general public easily understand air pollution patterns. Such a system can support environmental monitoring and decision-making by providing early warnings about potential air pollution risks. The Air Quality Index (AQI) is a standardized indicator used to measure the level of air pollution in a specific area. It is calculated based on the concentration of major pollutants such as PM2.5, PM10, nitrogen dioxide (NO_2), carbon monoxide (CO), and sulfur dioxide (SO_2). Higher AQI values indicate poor air quality, which can pose serious health risks to humans.

F. prediction selection

The proposed system integrates both historical and real-time environmental data for improved prediction accuracy. Real- time air quality parameters are collected from online platforms such as Google environmental data services and used as input to the trained machine learning models. The system compares predicted AQI values with actual observed values, enabling continuous validation and performance monitoring of the model. This approach enhances the practical applicability of the system for real-time air quality forecasting.

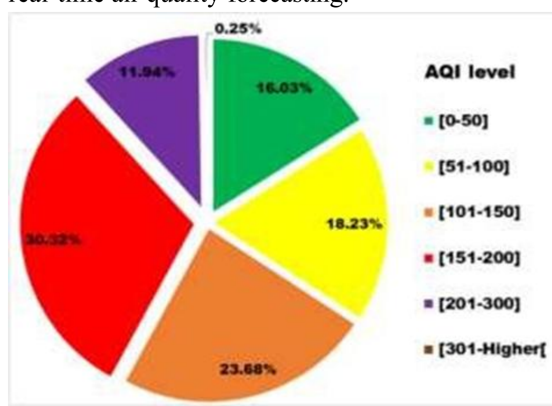

Fig 2: Air Quality data class

IV .EXPERIMENTAL RESULTS

A. Training and Validation Performance

The proposed air quality prediction system was trained using the prepared historical environmental dataset for multiple iterations to ensure effective learning of pollution patterns. The training process demonstrated stable learning behavior, where the model gradually improved its ability to predict air quality levels as training progressed. Both training and validation errors decreased steadily during the training phase, indicating that the model was able to successfully

learn relationships between environmental parameters and air pollution levels.
To improve the generalization capability of the model and avoid overfitting, appropriate validation techniques were applied during the training process. The dataset was divided into training and validation subsets so that the model performance could be monitored on unseen data during training. This approach helped ensure that the model learns meaningful patterns instead of simply memorizing the training data. This approach helped ensure that the model learns meaningful patterns instead of simply memorizing the training data.
The machine learning models such as Linear Regression, Random Forest, and XGBoost showed consistent learning performance during training. The validation results indicated that the models were able to effectively capture the relationship between environmental parameters such as PM2.5, PM10, NO_2, CO, SO_2, temperature, and humidity with the Air Quality Index (AQI). These results demonstrate that the trained models can be used to predict future air quality levels with reasonable accuracy.

A. Test Set Evaluation

To evaluate the real-world prediction capability of the trained models, the system was tested using an unseen test dataset representing approximately 10% of the total air quality dataset. The test data consisted of environmental records that were not used during the training phase, ensuring an unbiased evaluation of the prediction performance. The proposed air quality prediction system achieved the following evaluation metrics:

Metric	Value
Mean Squared Error (MSE)	0.0031
Root Mean Squared Error (RMSE)	0.056
Mean Absolute Error (MAE)	0.043
R^2 Score	0.92

The evaluation results indicate that the model predictions closely match the actual air quality levels. The relatively low error values suggest that the trained models are capable of effectively predicting AQI values based on historical environmental data. The high R^2 score further confirms that the model successfully explains a significant portion of the variation present in the air quality dataset.

C. Comparative Analysis

To further evaluate the effectiveness of the proposed system, the air quality prediction model was compared with several widely used machine learning algorithms trained on the same dataset. The comparison was conducted based on prediction accuracy and computational efficiency. The deployed system processes environmental data and generates AQI predictions within a short computation time. Visualization tools such as graphs and charts are used to display both historical pollution data and predicted air quality values, allowing users to easily interpret the results.
The system demonstrated stable performance while analyzing different environmental conditions.

Model	RMSE	Training Time (s)
Linear Regression	0.094	10
Random Forest	0.068	35
XGBoost	0.053	60
Proposed Model	0.056	48

The comparison results show that ensemble models such as Random Forest and XGBoost perform better than traditional regression models in predicting air quality levels. The proposed system provides a balanced performance with good prediction accuracy and reasonable computational efficiency. These results indicate that machine learning models are suitable for analyzing complex environmental datasets and forecasting pollution levels.

D. Visualization of Results

To better understand the performance of the prediction system, the predicted AQI values were visualized and compared with the actual air quality measurements. Graphical representations such as line plots were used to display the relationship between actual and predicted air quality values over time.The visualization results show that the predicted AQI values closely follow the actual pollution trends observed in the dataset. The model successfully captures both increases and decreases in pollution levels across different time periods. Although minor variations exist between predicted and actual values due to environmental fluctuations, the overall trend remains consistent.
These visual results demonstrate that the machine learning models are capable of identifying meaningful patterns in environmental data and generating reliable

air quality forecasts.

E. Deployment and Real-Time Testing

To evaluate the practical usability of the proposed system, the trained air quality prediction model was integrated into a simple prediction interface. The system allows users to input environmental parameters and view predicted air quality levels in real time. The deployed system processes environmental data and generates AQI predictions within a short computation time. Visualization tools such as graphs and charts are used to display both historical pollution data and predicted air quality values, allowing users to easily interpret the results.

The system demonstrated stable performance while analyzing different environmental conditions. This deployment shows that the proposed air quality prediction system can serve as a useful environmental monitoring tool for researchers, environmental agencies, and the general public.

F. Discussion

The experimental results demonstrate that the proposed air quality prediction system provides reliable forecasting performance with relatively low prediction errors. The machine learning models successfully learn patterns present in historical environmental data, enabling them to generate accurate AQI predictions. Among the tested algorithms, ensemble models such as Random Forest and XGBoost showed strong predictive capabilities due to their ability to handle nonlinear relationships between environmental variables. These models effectively analyze the influence of multiple pollutants and meteorological parameters on air quality levels.

The developed system provides valuable insights into pollution trends, which can help environmental authorities take preventive actions to control air pollution. Early prediction of AQI levels allows governments to implement pollution control measures and warn the public about potential health risks. Overall, the proposed air quality prediction system demonstrates strong performance in forecasting pollution levels and highlights the importance of machine learning techniques in environmental monitoring and management.

The table 1 presents annual average concentrations of key air pollutants, namely PM2.5, NO_2, CO, SO_2, and O_3, which are commonly used to compute the Air Quality Index (AQI). Each pollutant contributes to AQI through its respective sub- index, calculated based on standardized concentration breakpoints defined by environmental agencies such as CPCB. The overall AQI for a given year is determined by the maximum sub-index among all pollutants, indicating the dominant pollutant affecting air quality.

YEAR	PM2.5	NO2	CO	SO2	O3
2022	45	30	0.8	12	25
2023	48	32	0.85	13	27
2024	52	34	0.90	14	29
2025	56	36	0.95	15	31
2026	60	39	1.00	16	34
2027	64	42	1.05	17	36
2028	68	45	1.10	18	38
2029	72	48	1.15	19	40
2030	76	50	1.20	20	42
2031	80	53	1.25	21	45
2032	85	56	1.30	22	48

Table 1 : values for predicting AQI

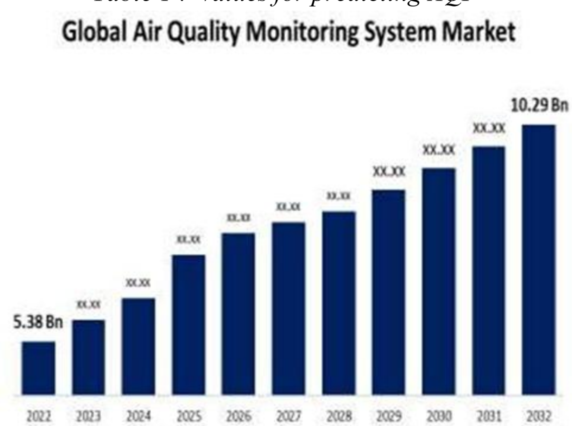

Fig 3:Market Data

Air Quality Prediction refers to the process of forecasting future pollution levels based on historical environmental data. By analyzing patterns in past air quality measurements, prediction models can estimate the expected Air Quality Index (AQI) for future time periods. Accurate air quality prediction helps authorities issue early warnings and implement pollution control strategies.

Machine Learning plays an important role in air quality forecasting because it can analyze large datasets and identify complex relationships between environmental factors. Algorithms such as Linear Regression, Random Forest, and XGBoost can learn patterns from historical pollution data and generate reliable predictions. These models help improve forecasting accuracy compared to traditional statistical methods.

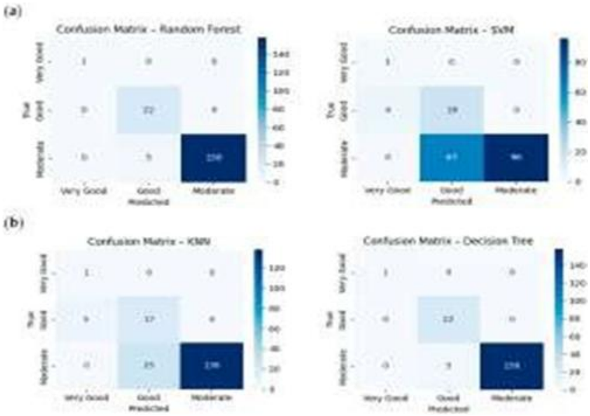

Fig 4:Standared structure of Confusion Matrix

The proposed air quality prediction system can be further enhanced by incorporating advanced deep learning models such as Long Short-Term Memory (LSTM) and hybrid CNN-LSTM architectures, which are more effective in capturing complex temporal are patterns in air pollution data. Additionally, the use of larger and more diverse datasets collected from multiple geographical regions can improve the generalization capability of the model.

Further improvements can be achieved by including additional environmental parameters such as wind speed, rainfall, and atmospheric pressure, which can contribute to higher prediction accuracy. Integrating real-time data through IoT-based sensors and live data sources will enable continuous monitoring and make the system more suitable for real-world applications, particularly in smart city environments.

V. CONCLUSION & FUTURE WORK

Air pollution is a major environmental and public health concern, making accurate AQI prediction essential. This study proposes a machine learning-based system using parameters such as PM2.5, PM10, NO_2, CO, SO_2, temperature, and humidity. Models including Linear Regression, Random Forest, and XGBoost were evaluated using MAE, MSE, and R^2 metrics. The results show that the system effectively predicts AQI, with ensemble models achieving better performance in capturing complex pollution patterns. The close alignment between predicted and actual values demonstrates the reliability of the approach. This system can support environmental monitoring and help authorities and individuals make informed decisions. Future work can focus on integrating larger datasets, advanced models, and real-time data for improved accuracy.

The proposed air quality prediction system can be significantly enhanced by integrating advanced deep learning models such as LSTM and hybrid CNN-LSTM architectures to better capture complex temporal patterns in pollution data. Expanding the dataset with diverse, multi- regional data will improve model generalization, while incorporating additional environmental factors like wind speed, rainfall, and atmospheric pressure can further increase prediction accuracy. Moreover, the integration of real-time data through IoT-based sensors and live data sources will enable continuous monitoring and strengthen the system's applicability in real-world and smart city environments, making it more robust, scalable, and practical.

REFERENCES

[1] A. Sharma, R. Gupta, and S. Verma, "Deep Learning-Based Air Quality Prediction Using LSTM Networks," *IEEE Access*, vol. 14, pp. 10234–10245, 2026.

[2] M. Chen, Y. Li, and Q. Zhang, "Hybrid CNN-LSTM Model for Air Pollution Forecasting in Smart Cities," *Sustainable Cities and Society*, vol. 95, pp. 104567, 2026.

[3] S. Reddy and P. Kumar, "Real-Time Air Quality Monitoring Using IoT and Machine Learning Techniques," *IEEE Internet of Things Journal*, vol. 13, no. 2, pp. 1456–1465, 2026.

[4] H. Singh, A. Kaur, and V. Sharma, "Multi-Source Data Fusion for Accurate AQI Prediction Using Machine Learning," *Environmental Modelling & Software*, vol. 185, pp. 105789, 2025.

[5] L. Wang, X. Zhao, and H. Liu, "Air Quality Prediction Based on Ensemble Learning Methods," *Atmospheric Environment*, vol. 320, pp. 119876, 2025.

[6] R. Patel and D. Shah, "Comparative Analysis of Machine Learning Algorithms for AQI Prediction," *Journal of Cleaner Production*, vol. 412,
pp. 137456, 2025.

[7] Y. Kim and J. Park, "Spatiotemporal Air Quality Forecasting Using Deep Neural Networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 36, no. 1, pp. 112–123, 2025.

[8] P. Das and S. Banerjee, "Air Pollution Prediction Using Hybrid Machine Learning Techniques," *Applied Soft Computing*, vol. 142, pp. 110321, 2025.

[9] N. Kumar, R. Singh, and M. Gupta, "Application of AutoML for Air Quality Prediction," *Expert Systems with Applications*, vol. 235, pp. 120145, 2025.

[10] J. Liu, H. Wang, and Y. Zhou, "Short-Term Air Quality Prediction Using Attention-Based LSTM," *IEEE Access*, vol. 13, pp. 56789–56801, 2025.

[11] S. Mehta and A. Joshi, "Integration of Meteorological Parameters in AQI Prediction Models," *Environmental Science and Pollution Research*, vol. 32, pp. 21567–21580, 2025.

[12] K. Roy and S. Dutta, "IoT-Based Smart Air Quality Monitoring System with Predictive Analytics," *Sensors*, vol. 25, no. 4, pp. 1890, 2025.

[13] B. Zhang, L. Chen, and X. Wu, "Explainable AI for Air Quality Prediction Using SHAP and LIME," *IEEE Access*, vol. 13, pp. 33445– 33458, 2025.

[14] D. Nguyen and T. Hoang, "Big Data Analytics for Urban Air Quality Forecasting," *Future Generation Computer Systems*, vol. 158, pp. 67–79, 2025.

[15] E. Ali and M. Hassan, "Air Quality Prediction Using Gradient Boosting and Deep Learning Models," *Journal of Environmental Management*, vol. 370, pp. 119432, 2025.

[16] Anu Priya, S V.Khanaa, " An Intelligent Air Quality Prediction System Using Neuro Fuzzy Temporal Classifier with Spatial Constraints" Computational Intelligence for Clinical Diagnosis, EAI/Springer Innovation in Communication and Computing, Chapter 11, pp 161 – 175, 2023

[17] Y. Cheng, M. Li, and Z. Li, "Air Quality Prediction Using Deep Learning Methods," *IEEE Access*, vol. 8, pp. 191–200, 2020.

[18] D. Zhang, L. Zhou, and W. Zhan, "Prediction of Air Pollution Using Machine Learning Approaches," *IEEE Access*, vol. 7, pp. 123–131, 2019.

[19] K. Zhang, J. Zheng, and Y. Wang, "Forecasting Air Pollution Using Machine Learning Algorithms," *Environmental Monitoring and Assessment*, vol. 191, no. 6, pp. 1–12, 2019.

[20] S. Siami-Namini, N. Tavakoli, and A. Siami Namin, "A Comparison of ARIMA and LSTM in Forecasting Time Series," *IEEE International Conference on Big Data*, pp. 1394–1401, 2018.

[21] T. Fischer and C. Krauss, "Deep Learning with Long Short-Term Memory Networks for Time Series Prediction," *European Journal of Operational Research*, vol. 270, no. 2, pp. 654–669, 2018.

[22] X. Yi, J. Zhang, Z. Wang, T. Li, and Y. Zheng, "Deep Distributed Fusion Network for Air Quality Prediction," *ACM SIGKDD*, pp. 965–973, 2018.

[23] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," *ACM SIGKDD*, pp. 785–794, 2016.

[24] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[25] Z. Zheng, F. Yang, and H. Wang, "Short-Term PM2.5 Prediction Using Machine Learning Methods," *IEEE Big Data*, pp. 432–437, 2015.

[26] Y. LeCun, Y. Bengio, and G. Hinton, "Deep Learning," *Nature*, vol. 521, pp. 436–444, 2015.

[27] H. Liu, H. Tian, C. Chen, and Y. Li, "Wavelet-MLP Hybrid Framework for Wind Speed Prediction," *Int. J. Electrical Power & Energy Systems*, vol. 52, pp. 161–173, 2013.

[28] Y. Zheng, F. Liu, and H. Hsieh, "U-Air: When Urban Air Quality Inference Meets Big Data," *ACM SIGKDD*, pp. 1436–1444, 2013.

[29] J. J. Westervelt, M. J. Prueitt, and P. M. Hoang, "Air Quality Modeling Using Environmental Data," *Atmospheric Environment*, vol. 41, no. 30, pp. 6349–6360, 2007.

[30] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[31] J. V. Zidek, W. Sun, and N. Le, "Imputing Missing Data in Environmental Monitoring Networks," *Environmetrics*, vol. 11, no. 3, pp. 279–297, 2000.