ADROID Conference Series: Engineering and Technology
Volume 2, Issue 3, pp.141-151, 2026

A Hyperparameter-Optimized LSTM Machine Learning Method for Predicting Air Quality

K. Venkatesh, A. Bhanu Teja, G. Parimala
Department of Computer Science and Engineering, Vignan's Foundation for Science, Technology and Research, Guntur, India
venkateshkoka022@gmail.com, tbhanu838@gmail.com, garnepudi.parimala@gmail.com

Abstract. Urban air pollution is a significant environmental concern that affects both the ecosystem and human health. In this paper, the authors propose a machine learning model for predicting the Air Quality Index (AQI). Various machine learning techniques are employed, including Linear Regression, Decision Tree, K-Nearest Neighbours (KNN), Random Forest, Gradient Boosting, XGBoost, and sequence modelling using Long Short-Term Memory (LSTM) networks for AQI prediction. The performance of these models is evaluated using key performance indicators such as Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), and the coefficient of determination (R^2 score). The experimental results reveal that the XGBoost model outperforms the other models, while hyperparameter tuning further improves the effectiveness of both the Decision Tree and KNN models. The proposed system demonstrates high predictive performance with MAE: 14.637, RMSE: 30.698, MAPE: 9.355%, and R^2 Score: 0.948. In addition, a web interface for real-time AQI monitoring has been developed, making the proposed system useful for public awareness and environmental management.

Keywords: Air Quality Index (AQI), Machine Learning, LSTM, Ensemble Learning, XGBoost, Random Forest, Air Pollution Forecasting, Environmental Monitoring.

1. INTRODUCTION

Globally, air quality is a critical concern, particularly in urban areas where high pollution levels have substantial adverse effects on the environment and human health. The air quality in Dehradun, India, has deteriorated due to rising vehicle emissions and industrial activity. Therefore, implementing efficient strategies for monitoring and managing air quality has become essential. Machine learning techniques can significantly improve the accuracy of Air Quality Index (AQI) prediction, thereby helping reduce the harmful effects of air pollution [1].

Air quality is also a major concern in many urban regions worldwide, where pollution poses serious environmental and public health risks. Fine particulate matter, such as PM2.5, contributes to respiratory and cardiovascular diseases. Consequently, exploring machine learning techniques is essential for accurately predicting PM2.5 levels in Kuwait, enabling more effective air quality management [2].

Air quality directly influences the well-being of urban residents. This study aims to predict the Air Quality Index (AQI) for Visakhapatnam, India, by analysing key pollutants and meteorological factors using machine learning methods. Through this approach, the study seeks to enhance understanding of air quality patterns and ultimately improve the quality of life [3].

Air pollution has become one of the most significant global environmental and health challenges. Industrial growth and rapid urbanization have contributed to a continuous decline in air quality, increasing the need for efficient prediction and management techniques. This research investigates how machine learning methods can be effectively utilized to predict air quality [4].

Forecasting air quality is essential for effective environmental management. This study explores the application of machine learning algorithms using a comprehensive dataset provided by the Environmental Protection Administration in Taiwan. The primary objective is to identify the most suitable prediction techniques for improving air quality management [5].

Environmental and health concerns related to air pollution are particularly severe in urban areas. This research examines how machine learning methods can be applied to enhance air quality management by focusing on critical atmospheric pollutants. The primary objective is to promote better environmental and public health management [6].

Air pollution poses significant risks to both the environment and public health, especially in urban environments. This study addresses the urgent need for efficient air quality monitoring by utilizing machine learning methods to improve the accuracy of Air Quality Index (AQI) forecasting. The research aims to enhance air pollution management through the application of advanced algorithms [7].

As pollutants continue to be released into the atmosphere, air pollution remains a serious threat to public health. Accurate evaluation of air quality is essential for protecting vulnerable populations from pollution-related health problems. This objective can be achieved through the application of XGBoost, which enhances the accuracy of air quality forecasting [8].

This paper evaluates the effectiveness of various machine learning models for forecasting the Air Quality Index (AQI) using an air quality dataset. The primary objective is to develop a highly accurate predictive model capable of estimating the degree of air pollution, thereby assisting both the public and government authorities in taking preventive measures. The outcomes of this study emphasize the importance of advanced machine learning techniques in improving air quality prediction.

2. LITERATURE SURVEY

The importance of accurate air quality prediction for environmental and health monitoring has been emphasized in numerous research studies. Deep Learning (DL) and Machine Learning (ML) methods have proven highly effective in handling the complex data patterns associated with air quality analysis. Although extensive research has been conducted in several major cities, studies focusing on regions such as Dehradun remain limited [1].

To develop effective predictive models, understanding the dynamics of air pollution, particularly PM2.5 concentration, is essential. Existing research provides a basic understanding of air pollution behavior; however, significant research gaps remain, especially in regions such as Kuwait. Recent studies emphasize the importance of addressing region-specific challenges and employing machine learning techniques for accurate air quality forecasting [2].

The adverse effects of air pollution on both environmental sustainability and public health have been widely documented, highlighting the need for efficient monitoring systems. Machine learning techniques have gained considerable attention because of their ability to analyze complex datasets and generate accurate predictions. Studies have demonstrated that algorithms such as Random Forest and XGBoost significantly outperform traditional statistical approaches. Furthermore, air quality is influenced by multiple interacting factors, including pollutant concentrations and meteorological conditions, making advanced computational methods highly suitable for prediction tasks [3].

Considering the environmental and health hazards associated with industrialization and urbanization, research activities in air quality prediction have increased substantially. Intelligent techniques such as Artificial Neural Networks (ANN) and Support Vector Machines (SVM) have been successfully applied to predict pollutants including NO_2, PM2.5, and SO_2. Studies further indicate that integrating satellite observations with ground-based monitoring data enhances prediction accuracy, thereby improving environmental management and awareness [4].

Numerous studies have explored the application of machine learning techniques for predicting air quality and evaluating their capability to process complex environmental datasets. Compared to conventional statistical methods, algorithms such as Support Vector Machines (SVM), Artificial Neural Networks (ANN), and Random Forest (RF) have demonstrated superior predictive performance. In addition, ensemble learning approaches that combine the strengths of multiple algorithms have further improved prediction accuracy. Recent research indicates an increasing trend toward applying advanced machine learning techniques to address air pollution and its associated health impacts [5].

Several researchers have investigated machine learning approaches for forecasting air quality. Lei Liu et al. employed recurrent neural networks to predict pollutant concentrations; however, their work did not directly estimate the Air Quality Index. Similarly, Qitie Xia demonstrated the potential of machine learning techniques for AQI evaluation but did not implement a complete predictive framework. Xiaofeng Song utilized Ridge Regression for air quality forecasting, demonstrating the applicability of regression-based methods. Although considerable progress has been made, there remains a need for a comprehensive predictive model capable of delivering improved forecasting performance [6].

Numerous studies have focused on predicting air pollutant concentrations using machine learning algorithms. Research has demonstrated that classification algorithms can effectively analyze pollutants such as ground-level O_3, CO, SO_2, and NO_2. Various machine learning algorithms, including Support Vector Regression (SVR), Random Forest Regression (RFR), Decision Tree Regression (DTR), and Linear Regression, have been applied for air quality prediction. Among these approaches, Random Forest Regression has frequently produced the most accurate results [7].

Many studies have also investigated the use of Artificial Neural Networks (ANN), Genetic Algorithms, Decision Trees, XGBoost, and Support Vector Machines for air quality forecasting. Recent findings indicate that XGBoost and SVM provide high prediction accuracy, highlighting the importance of selecting the most suitable algorithm for effective air quality monitoring [8].

Air pollution has been extensively studied because of its significant impact on environmental conditions and public health. Gurjar et al. (2016) analyzed air pollution patterns in major Indian cities and emphasized the importance of effective pollution control strategies, while Khillare et al. (2012) examined the health hazards associated with metal particles in Delhi's atmosphere. More recent studies by Li et al. (2016) and Lord et al. (2021) have demonstrated promising results using machine learning techniques for air quality prediction, reflecting the growing adoption of data-driven solutions for environmental management [9].

A review of existing research indicates that many important aspects of air quality prediction have already been investigated. However, previous studies also reveal several knowledge gaps that remain unresolved. Consequently, the present study focuses on introducing new perspectives while building upon existing knowledge to improve air quality prediction methods [10].

Several researchers have further examined the application of machine learning algorithms for air quality prediction. Manisalidis et al. (2020) highlighted the significant impact of air pollution on both environmental quality and human health while emphasizing the importance of monitoring and forecasting air quality. Mahalingam et al. (2019) developed a smart city-based air quality prediction system, demonstrating the role of technology in environmental management. Soundari (2019) applied machine learning techniques for evaluating air quality across India. Additional studies by Pant et al. (2022) and Halsana (2020) reported that algorithms such as Random Forest and Support Vector Machines consistently achieved high prediction accuracy for air pollutant levels [11].

The literature on air quality forecasting presents a wide range of prediction methods and analytical tools. Machine learning techniques such as Random Forest and Support Vector Regression have significantly improved prediction accuracy in various metropolitan areas. Moreover, time-series models including SARIMA and ARIMA have demonstrated effectiveness in analyzing historical air quality trends and forecasting future pollution levels. These findings emphasize the need for continuous improvement of predictive models to address urban air pollution challenges more effectively [12].

3. PROPOSED METHODOLOGY

A. Data Collection

This study focuses on collecting and analyzing air quality data to predict the Air Quality Index (AQI). Air pollutants such as particulate matter (PM), carbon monoxide (CO), nitrogen dioxide (NO_2), sulfur dioxide (SO_2), ozone (O_3), and other environmental parameters are considered because they directly influence air pollution levels. These parameters help in understanding variations in air quality and their impact on pollution levels.

The dataset used in this study contains 29,532 records with 16 variables, representing different environmental conditions and time periods. The availability of a large dataset enables the model to learn more effectively and improve prediction accuracy. The dataset is divided into 80% training data and 20% testing data. This data collection strategy facilitates the development of an efficient and accurate model for AQI prediction and pollution control.

Figure 1. Schematic Representation of the Proposed System

B. LSTM Model Architecture

The proposed system utilizes a Long Short-Term Memory (LSTM) network to forecast the Air Quality Index (AQI) using historical air pollution data. LSTM is a type of recurrent neural network (RNN) specifically designed to capture long-term dependencies in sequential time-series data.

The proposed framework accepts sequential air quality parameters, including pollutant concentrations and environmental variables, as input. These inputs are processed through an LSTM layer consisting of memory cells equipped with input gates, forget gates, and output gates. These gates enable the network to retain important information while discarding irrelevant data, thereby preserving meaningful temporal patterns.

Finally, the processed information is passed through a dense output layer, which generates the predicted Air Quality Index with improved accuracy.

C. Hyperparameter Tuning

The predictive performance of the Air Quality Index (AQI) models is enhanced through hyperparameter optimization. The objective of this process is to identify the optimal combination of model parameters that maximizes prediction accuracy.

For the Decision Tree Regressor, parameters such as:

- Maximum tree depth,
- Minimum samples required for splitting, and
- Minimum samples required at leaf nodes

are adjusted to reduce overfitting.

Similarly, for the K-Nearest Neighbors (KNN) Regressor, hyperparameters including:

- Number of neighbors (k),
- Distance measurement method, and
- Weighting strategy

are optimized.

Multiple parameter combinations are evaluated, and the best-performing configuration is selected based on evaluation metrics. This optimization process improves the overall prediction accuracy of the Air Quality Index.

D. Training Procedure

The training process begins with preparing the air quality dataset. This preparation includes:

- Handling missing values,
- Standardizing the data, and
- Dividing the dataset into training and testing subsets.

The prepared dataset is then used to train the LSTM model, enabling it to learn patterns and relationships between historical pollution levels and AQI values.

During training, an optimization algorithm continuously updates the model parameters to minimize prediction errors. The model is trained in over multiple epochs until stable performance is achieved.

After training, the predictive capability of the model is evaluated using the testing dataset with the following performance metrics:

- Mean Squared Error (MSE),
- Root Mean Square Error (RMSE),
- Mean Absolute Percentage Error (MAPE), and
- Coefficient of Determination (R^2 Score).

E. Web-Based AQI Prediction System

To facilitate real-time Air Quality Index (AQI) forecasting and monitoring, a web-based interface has been developed. The system collects ambient sensor data and processes it to generate AQI predictions.

The web application securely communicates with the backend prediction model using an API key. The prediction model retrieves sensor information, including pollutant concentrations and environmental parameters, to calculate the AQI.

The predicted AQI values are then displayed through an interactive and user-friendly web interface, allowing users to monitor air quality in real time. The integration of machine learning models with the web application enhances usability while providing accurate and timely AQI predictions.

4. ALGORITHM

A. Linear Regression

Linear Regression is a supervised machine learning algorithm that models the relationship between the independent variables (X) and the dependent variable (Y) using a straight-line equation to predict the output.
Equation

$$y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_n x_n$$

where:

- $y$indicates the predicted output.
- $\alpha_0$denotes the intercept.
- $\alpha_0, \alpha_1, \ldots, \alpha_n$represent the regression coefficients.
- $x_1, x_2, \ldots, x_n$represent the input features.

B. K-Nearest Neighbors (KNN) Regressor

The K-Nearest Neighbors (KNN) Regressor predicts the target value of a new data point by calculating the average of the target values of the K nearest data points in the dataset.
Prediction Equation

$$\hat{y} = \frac{1}{k}\sum_{i=1}^{k} y_i$$

where:

- $k$indicates the number of nearest neighbors.
- $y_i$represents the target value of each neighboring data point.

Distance Measurement (Euclidean Distance)

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

The Euclidean distance is commonly used to determine the similarity between data points while identifying the nearest neighbors.

C. Decision Tree Regressor

The Decision Tree Regressor is a machine learning technique that partitions the dataset into multiple branches based on feature values to minimize prediction errors.

Mean Squared Error (MSE) Splitting Criterion

$$MSE = \frac{1}{n}\sum_{i=1}^{n} (z_i - \hat{z})^2$$

where:

- $z_i$represents the actual value.
- $\hat{z}$represents the predicted value.
- $n$indicates the total number of samples.

The Decision Tree uses this Mean Squared Error criterion to determine the optimal split at each node, thereby reducing prediction error.

D. Ensemble Models

Ensemble models are machine learning algorithms that combine multiple individual models rather than relying on a single model. This ensemble learning approach improves prediction accuracy by integrating the strengths of several weaker models into a more robust predictive framework.

For example:

- Random Forest Regressor employs the bagging technique, where multiple decision trees are trained independently using different subsets of the training data. The final prediction is obtained by averaging the outputs of all trees.
- Gradient Boosting Regressor utilizes the boosting technique, where decision trees are trained sequentially. Each successive tree learns from the prediction errors of the previous tree, thereby improving overall prediction accuracy.
- XGBoost Regressor (Extreme Gradient Boosting) is an enhanced implementation of the Gradient Boosting algorithm. It incorporates additional features such as regularization, parallel processing, and efficient handling of missing values, making it faster, more scalable, and more accurate than conventional gradient boosting methods.

5. EXPERIMENTAL RESULTS AND DISCUSSIONS

TABLE I. EVALUATING DIFFERENT MACHINE LEARNING MODELS

Models Used	MAE	RMSE	MAPE	R^2 Score
XGBoost	14.637	30.698	9.355%	0.948
Random Forest	14.415	31.617	9.040%	0.945
Gradient Boosting	17.227	33.171	11.705%	0.939
K-Nearest Neighbors (KNN)	21.168	43.987	13.634%	0.894
Decision Tree	19.162	44.310	11.957%	0.892
Linear Regression	31.238	57.756	22.108%	0.817

From Table I, it is evident that the XGBoost Regressor performed better than all the other models evaluated in this study. It achieved the highest R^2 score of 0.948, along with low error values of 14.637 MAE, 30.698 RMSE, and 9.355% MAPE. The Random Forest Regressor and Gradient Boosting Regressor also demonstrated strong predictive performance, with relatively low error values and high R^2 scores.

The KNN Regressor and Decision Tree Regressor achieved moderate prediction accuracy. In contrast, Linear Regression exhibited the poorest performance, recording the highest prediction errors and the lowest R^2 score of 0.817, indicating its limited effectiveness for AQI prediction.

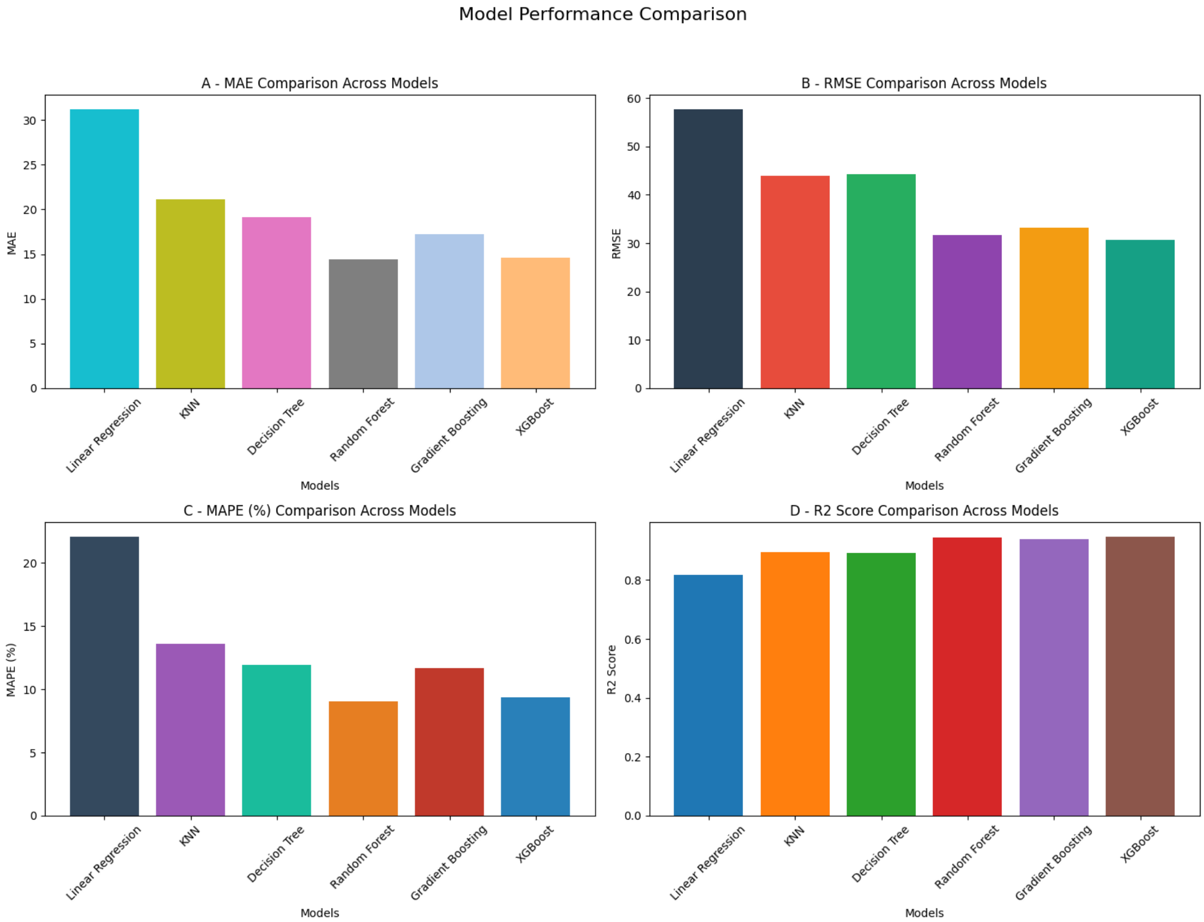

Figure 2. Model Performance Comparison

Figure 2 presents a bar chart comparing the performance of different machine learning models for Air Quality Index (AQI) prediction using evaluation metrics such as Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), and R^2 Score.

The comparison demonstrates that the Random Forest Regressor and XGBoost Regressor, both ensemble learning models, achieved the lowest prediction errors and the highest predictive accuracy. Conventional models such as Linear Regression exhibited significantly higher error rates, indicating comparatively poor prediction capability.

Overall, the results shown in Figure 2 demonstrate that ensemble learning techniques considerably improve AQI prediction performance.

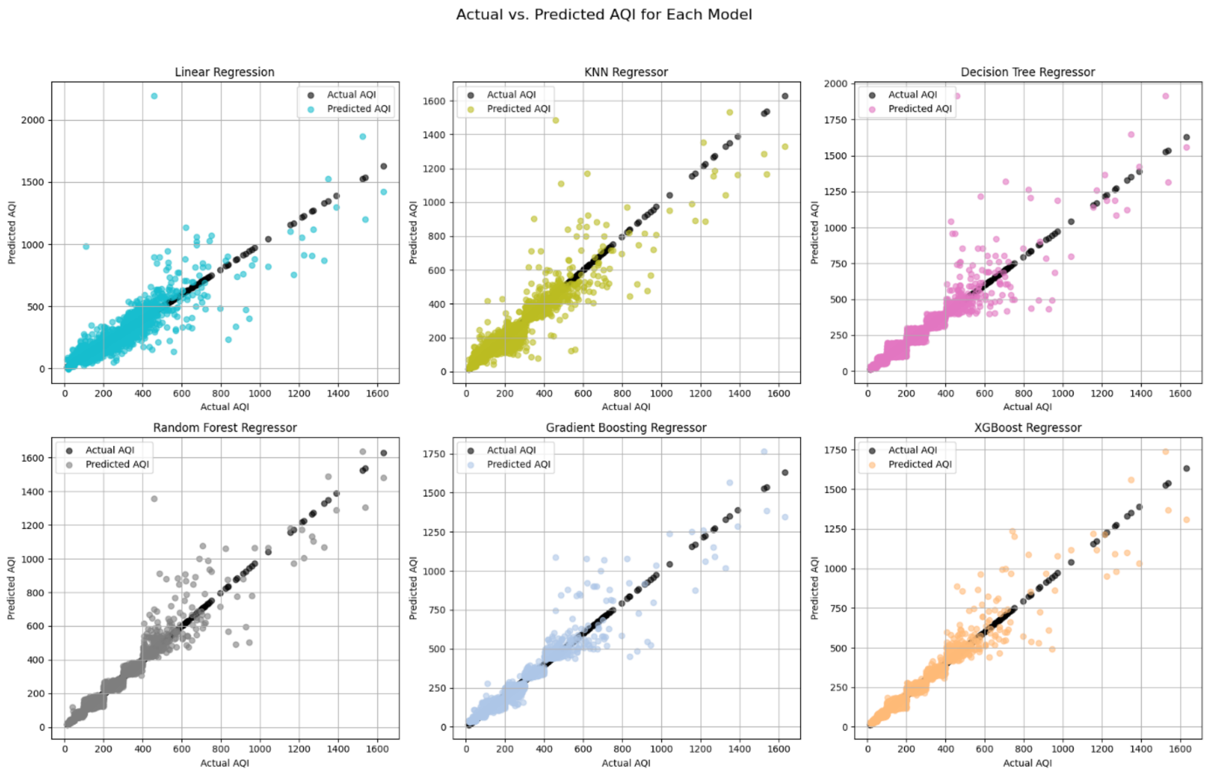

Figure 3. Actual vs. Predicted AQI Values for Different Machine Learning Models

Figure 3 illustrates the comparison between the actual and predicted Air Quality Index (AQI) values generated by different machine learning models.

The figure presents prediction results obtained using:

- Linear Regression
- K-Nearest Neighbors (KNN) Regressor
- Decision Tree Regressor
- Random Forest Regressor
- Gradient Boosting Regressor
- XGBoost Regressor

The accuracy of each model can be assessed by observing how closely its predicted values align with the diagonal reference line. The results indicate that ensemble learning models, particularly Random Forest, Gradient Boosting, and XGBoost, provide predictions that are considerably closer to the actual AQI values than the other machine learning models.

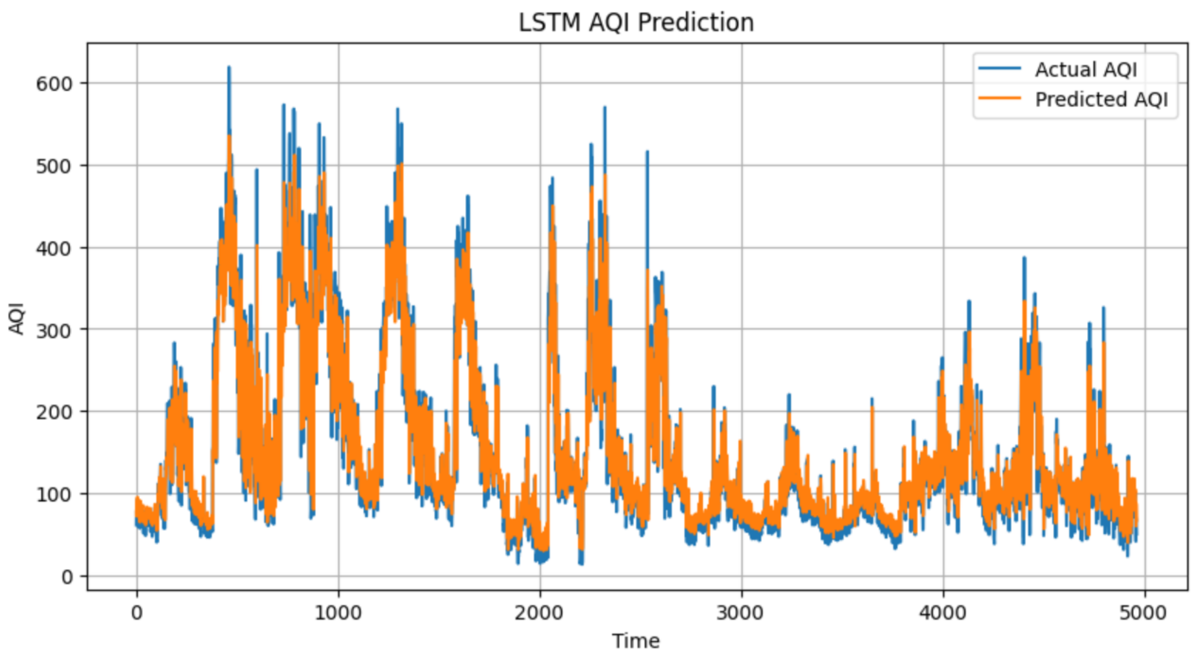

Figure 4. LSTM-Based Actual vs. Predicted AQI Trend

Figure 4 compares the actual Air Quality Index (AQI) values with those predicted by the Long Short-Term Memory (LSTM) model over time.

In the figure:

- The blue line represents the actual AQI measurements.
- The orange line represents the AQI values predicted by the LSTM model.

The close agreement between the two curves demonstrates the ability of the LSTM model to learn temporal patterns and variations in AQI data effectively. Although small discrepancies exist at certain points, the predicted values generally follow the actual AQI trend, indicating satisfactory forecasting performance.

TABLE II. DECISION TREE REGRESSOR PERFORMANCE BEFORE AND AFTER HYPERPARAMETER TUNING

Metric	Before Tuning	After Tuning
MAE	19.162	19.596
RMSE	44.310	39.804
MAPE	11.957%	13.991%
R^2 Score	0.8928	0.9135

TABLE III. KNN REGRESSOR PERFORMANCE BEFORE AND AFTER HYPERPARAMETER TUNING

Metric	Before Tuning	After Tuning
MAE	21.168	19.023
RMSE	43.987	41.485
MAPE	13.634%	11.843%
R^2 Score	0.8943	0.9060

Tables II and III present the performance of the Decision Tree Regressor and K-Nearest Neighbors (KNN) Regressor, respectively, before and after hyperparameter tuning.

As observed in Table II, although slight variations are present in the MAE and MAPE values, hyperparameter tuning improved the Decision Tree Regressor by reducing the RMSE and increasing the R^2 score, thereby enhancing its overall predictive capability.

Similarly, Table III shows that hyperparameter tuning improved the performance of the KNN Regressor, as indicated by reductions in MAE, RMSE, and MAPE, along with an increase in the R^2 score.

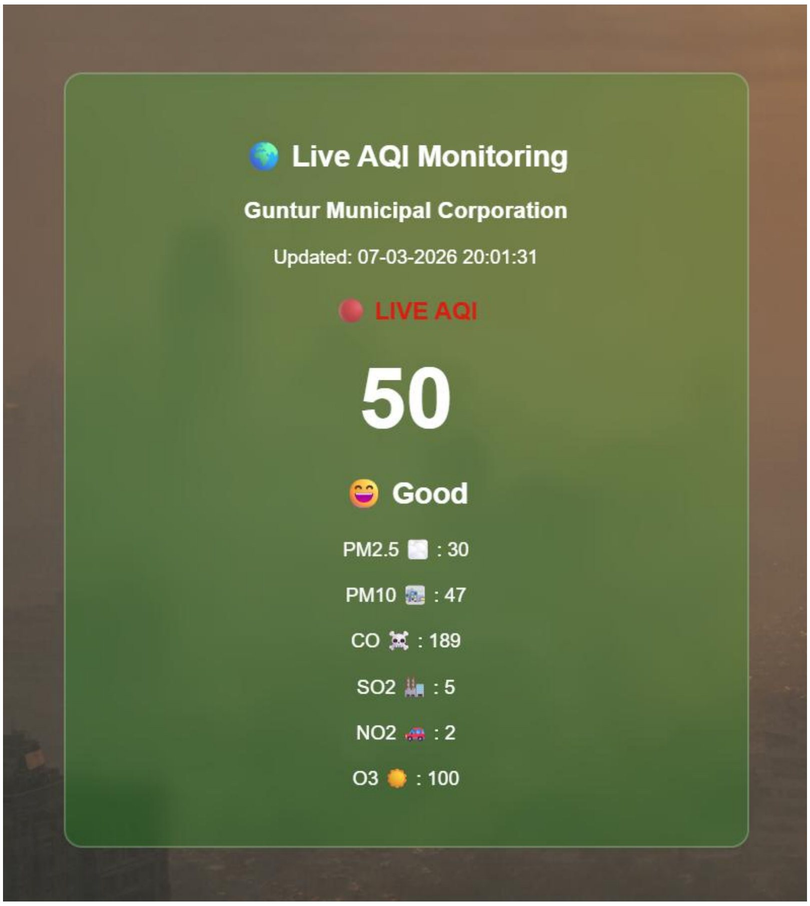

Figure 5. Live AQI Monitoring System Interface

The developed Live AQI Monitoring System displays real-time concentrations of major air pollutants together with the current Air Quality Index (AQI).

The system evaluates the overall AQI by analyzing pollutant concentrations and comparing them against established environmental standards. Furthermore, the interface provides users with real-time updates regarding air quality conditions, enabling continuous monitoring.

The implementation demonstrates that the proposed system can efficiently collect, process, and present air quality information in an intuitive manner, thereby improving user awareness and understanding of air pollution.

6. CONCLUSION

The present study demonstrates the effectiveness of applying machine learning techniques for predicting the Air Quality Index (AQI) by utilizing models such as Linear Regression (LR), Decision Tree (DT), K-Nearest Neighbors (KNN), Random Forest (RF), Gradient Boosting, XGBoost, along with the inclusion of a Long Short-Term Memory (LSTM) model for trend prediction.

Among the evaluated models, ensemble learning techniques, particularly Random Forest and XGBoost, achieved superior prediction accuracy while producing lower prediction errors compared to conventional machine learning models. Furthermore, hyperparameter tuning enhanced the predictive performance of the Decision Tree and KNN models.

In addition, a web-based Air Quality Index monitoring system was developed to provide users with a simple and intuitive interface for monitoring current air quality conditions. Overall, the proposed system has the potential to improve both environmental awareness and effective air quality monitoring.

Conflict of Interest

The authors declare no conflicts of interest regarding the current research.

References

[1] Dawar et al. (2025) explored the use of utilizing machine learning to forecast the Air Quality Index in

Dehradun, a city in the Himalayas, as detailed in their study published in *Natural Hazards*.
[2] Alrashidi H., Sibai F. N., Abonamah A., Alrashidi M., and Alsaber A. examined PM2.5 levels and predicted the Air Quality Index through machine learning based on data from air quality monitoring stations in Kuwait, as published in *Sustainability* in 2025.
[3] Gokulan Ravindiran, Gasim Hayder, Karthick Kanagarathinam, Avinash Alagumalai, and Christian Sonne conducted a predictive study on forecasting air quality in Visakhapatnam through machine learning models, India, published in *Chemosphere*, Volume 338, 2023, Article 139518 (ISSN 0045-6535).
[4] Singh et al. (2022) investigated air quality forecasting using machine learning techniques in their publication in the *SSRN Electronic Journal*.
[5] Liang Y.-C., Maimury Y., Chen A. H.-L., and Juarez J. R. C. published a 2020 study in *Applied Sciences* on utilizing machine learning to predict air quality.
[6] C. Li, Y. Li, and Y. Bao presented their work titled "Research on Air Quality Prediction Based on Machine Learning" at the 2021 2nd International Conference on Intelligent Computing and Human-Computer Interaction (ICHCI) held in Shenyang, China.
[7] B. D. Parameshachari, G. M. Siddesh, V. Sridhar, M. Latha, K. N. A. Sattar, and G. Manjula presented their work titled "Prediction and Analysis of Air Quality Index using Machine Learning Algorithms" at the 2022 International Conference on Data Science and Intelligent Systems (ICDSIS) held in Hassan, India.
[8] M. S. Ram, C. Reshmasri, S. Shahila, and J. V. P. Saketh presented their work titled "Air Quality Prediction using Machine Learning Algorithm" at the 2023 International Conference on Smart Computing and Data Science (ICSCDS) held in Erode, India.
[9] Bhathal, B. S., and Gupta, G. (2024) presented a study on air quality prediction utilizing a hybrid machine learning model from Punjab, featured in the *Journal of Electrical Systems*, Volume 17, Issue 1.
[10] Gupta et al. (2023) conducted a comparative analysis of machine learning techniques for predicting the Air Quality Index, published in the *Journal of Environmental and Public Health*.
[11] Pant, A., Sharma, S., and Pant, K. (2023) evaluated different machine learning algorithms to forecast the Air Quality Index (AQI) in the *Journal of Reliability and Statistical Studies*, Volume 16, Issue 2, pp. 229–242.
[12] Yang, W., Guo, J., and other authors (2024) presented a comparative study on machine learning methods for predicting air quality in Xi'an, China, featured in the proceedings of the Artificial Intelligence and Pattern Recognition Conference, pp. 1170–1176.