International Journal of Science and Engineering Applications
Volume 15-Issue 05, 05 – 17, 2026, ISSN:- 2319 - 7560
DOI: 10.7753/IJSEA1505.1002

Comparative Study on Air Pollution Prediction Using Machine Learning Techniques

Kiran Singh, Shefali Madan

Department of Computer Science and Engineering, Echelon Institute of Technology, Faridabad, Haryana

Abstract: The Air Quality Index (AQI) measures how air pollution affects human health. As pollution levels rise in Indian cities, we need reliable prediction models for better environmental management. This paper analyses different machine learning techniques for predicting AQI. We use Support Vector Regression (SVR), Random Forest Regression (RFR), and CatBoost Regression (CR) on data from New Delhi, Bangalore, Kolkata, and Hyderabad. We evaluate model performance using Root Mean Square Error (RMSE) and accuracy.

Experimental results show that RFR performs best in most cities, while CR is most effective in New Delhi. To tackle dataset imbalance, we use the Synthetic Minority Oversampling Technique (SMOTE), which improves prediction accuracy for all models. Additionally, we assess other models, including SARIMA, Support Vector Machine (SVM), and Long Short-Term Memory (LSTM) for Ahmedabad. Among these, SVM with a radial basis function (RBF) kernel shows the best results.

The findings emphasize how combining data balancing methods with machine learning models can improve AQI prediction. This approach can help with pollution control strategies and better decision-making.

Keywords: Air Pollution Prediction, Machine Learning, Deep Learning, NF-VAE, IoT, Time-Series Forecasting, Multivariate Data.

1. Introduction

Air pollution is a major environmental and public health concern across the world, especially in rapidly growing urban areas. Increased industrialisation, urbanisation, and vehicle emissions have led to a significant decline in air quality. Pollutants such as PM2.5, PM10, NO_2, SO_2, CO, and O_3 pose serious health risks, with PM2.5 being particularly harmful due to its ability to penetrate deep into the lungs and bloodstream. Exposure to such pollutants is linked to respiratory diseases, cardiovascular problems, and millions of premature deaths annually [1].

In addition to health impacts, air pollution contributes to environmental issues such as climate change, acid rain, and ecosystem damage. With the development of smart cities, IoT-based sensor networks are widely used to collect real-time air quality data. However, the large and complex nature of this data makes it difficult to analyze using traditional methods, which often fail to capture dynamic and nonlinear patterns [2].

To address this, Machine Learning (ML) and Deep Learning (DL) techniques

are increasingly used for air quality prediction. ML models like Support Vector Machines and Random Forests can identify patterns from historical data, while DL models such as Neural Networks and Recurrent Neural Networks provide better performance for time-series analysis [3]. However, these models often require high computational resources, making real-time implementation challenging.

Optimization techniques such as model pruning and quantization help reduce complexity and enable deployment on resource-limited devices. Additionally, performance evaluation using metrics like MAE and RMSE is essential to identify the most effective models.

Overall, developing efficient and accurate prediction models is crucial for improving air quality management and supporting sustainable urban development [3] [4].

2. Problem Statement

Air pollution has become a significant challenge in modern urban environments, particularly in rapidly developing smart cities where industrialization, vehicular emissions, and population growth continuously degrade air quality. Accurate air quality prediction is essential for enabling timely health advisories, supporting environmental policies, and ensuring public safety. However, predicting air quality remains a complex task due to the dynamic and nonlinear nature of environmental data, which is influenced by multiple factors such as weather conditions, traffic patterns, and geographical variations.

Traditional statistical models have shown limitations in capturing complex patterns and temporal dependencies present in large-scale, real-world datasets. Although Machine Learning (ML) and Deep Learning (DL) techniques have improved prediction accuracy, they often prioritize performance without adequately addressing computational efficiency, scalability, and real-time applicability. This creates a gap between theoretical model performance and practical deployment.

Furthermore, advanced ML/DL models typically require high computational power, memory, and training time, making them unsuitable for deployment on resource-constrained devices such as IoT nodes and embedded systems commonly used in smart city infrastructures. In addition, issues related to data quality, including missing, noisy, or inconsistent data, further reduce the reliability of predictions. External factors such as sudden weather changes and seasonal variations also add to the complexity of accurate forecasting.

Therefore, there is a need to develop an efficient, scalable, and accurate air quality prediction system that not only improves prediction performance but also ensures low computational complexity and real-time applicability. A systematic comparison of various ML and DL models, along with the integration of optimization techniques, is essential to identify suitable approaches for practical deployment in smart city environments.

3. Objectives

The main thing I want to do here is build some machine learning models that can predict air quality pretty well, you know, by trying out different regression methods. It seems like that could be useful for figuring out pollution levels or something.

Then, comparing how these models do against each other makes sense, to see which one handles the data best in certain situations. Some might be quicker, others more accurate, I am not totally sure yet.

For checking them, I plan to use metrics like Mean Absolute Error and Root Mean Square Error, those are the key ones to measure how off the predictions are. RMSE especially stands out because it penalizes bigger errors more.

On top of that, looking at how fast they run is important too, both on a regular setup and when using Apache Spark for distributed stuff. Processing time can vary a lot, and that might make one model way better for real use.

Overall, the goal is finding a model that gets good predictions without taking forever to compute, so it actually works in practice. That balance is tricky, but I think its worth pushing for. Some parts of this might get messy along the way.

4. Literature Review

Air quality prediction has become an important research area because of its direct impact on human health, environmental sustainability, and urban planning. With rapid industrialization and urban growth, accurately predicting the Air Quality Index (AQI) is essential for controlling pollution and supporting smart city development. Over time, different methods have been proposed, ranging from traditional statistical approaches to more advanced machine learning and deep learning techniques, all aiming to improve prediction accuracy.

The work by Armin Mazinani et al. [1] introduces an advanced framework for air quality prediction that is specifically designed for resource-limited devices such as embedded systems and IoT nodes. This study mainly focuses on PM2.5 due to its serious health effects and its importance in determining AQI. Along with PM2.5, other environmental factors like temperature, humidity, and gas sensor readings are also considered to improve performance. The authors experimented with various ML and DL models, including RNN, LSTM, GRU, BiLSTM, BiGRU, and TCN, as well as hybrid models like CNN-LSTM and CNN-BiGRU. Among these, the CNN-BiGRU model performed the best because it effectively captures both spatial and temporal patterns in the data. The models were evaluated using metrics such as RMSE, MAE, MAPE, and R^2. A key contribution of this research is the use of Post-Training Quantization (PTQ), which significantly reduces model size and inference time while maintaining accuracy, making it suitable for real-time applications in smart cities.

Another study titled “A Complete Air Pollution Monitoring and Prediction Framework” [4] presents a comprehensive system that combines real-time monitoring with prediction. It uses IoT sensors to continuously collect data on pollutants like PM2.5, PM10, CO, NO_2, and SO_2, along with weather-related factors. This data is then analyzed using machine learning models such as Linear Regression, Decision Tree, Random Forest, and SVM. Among these, Random Forest showed the best performance due to its ability to handle nonlinear relationships and avoid overfitting. While this framework provides an effective end-to-end solution, its performance may still depend on sensor

quality, data reliability, and computational limitations in large-scale systems.

Chunhao Liu et al. [5] proposed an optimization-based model called GA-KELM for AQI prediction. This model improves the traditional Extreme Learning Machine by using a genetic algorithm to optimize parameters such as weights and hidden nodes. Since standard ELM models can be unstable due to random parameter selection, this approach enhances both stability and accuracy. The model was tested on multiple pollutants and compared with other methods like SVM, CMAQ, and DBN-BP. Results showed that GA-KELM achieves better accuracy and faster training, making it an efficient solution for air quality prediction.

Overall, the literature shows a clear shift from traditional statistical models to more advanced ML and DL approaches. While older models struggled with nonlinear and dynamic data, newer hybrid and optimization-based techniques such as CNN-BiGRU and GA-KELM have significantly improved performance. The integration of IoT with predictive models has also enabled real-time monitoring, which is crucial for smart cities. However, challenges like data quality, model generalization, and computational efficiency still need to be addressed.

Another study proposes a hybrid deep learning model combining CNN and LSTM for predicting PM2.5 levels. In this approach, CNN is used to extract spatial features from data collected across multiple locations, while LSTM captures time-based patterns. By combining these two methods, the model can effectively learn both spatial and temporal relationships. It uses multiple input features, including pollutant levels and weather conditions, which improves prediction accuracy. The results show that this hybrid model performs better than traditional models such as BPNN, RNN, and standalone CNN or LSTM. However, the study is limited to a specific dataset, and further research is needed to test its performance in different regions [6].

Earlier approaches to AQI prediction mainly relied on statistical models like ARIMA, but these were not effective in capturing nonlinear relationships, leading to lower accuracy [7]. To overcome this, machine learning techniques such as SVR, KNN, and Random Forest were introduced, which improved performance by handling nonlinear data more effectively [8], [9]. However, these models still face challenges when dealing with large datasets and high computational requirements [10].

Neural network-based models, including backpropagation networks and nonlinear autoregressive models, further improved prediction accuracy by learning complex data patterns. Despite this, they often suffer from slow convergence and the risk of getting stuck in local minima, which affects performance [11].

More recently, hybrid and ensemble methods have been developed to improve prediction accuracy by combining multiple algorithms. These approaches have shown better robustness and reliability compared to single models [12], [13]. Optimization techniques such as particle swarm optimization and grey wolf optimization have also been used to improve feature selection and reduce computational complexity [14].

Deep learning models like LSTM, GRU, and CNN have gained popularity due to their ability to capture both temporal and spatial dependencies in air quality data. Advanced hybrid models and attention-based techniques further enhance performance, although they require significant computational resources and large datasets [15] [16].

The integration of IoT with machine learning has enabled real-time air quality monitoring and prediction. These systems rely on sensor networks and data fusion techniques to provide timely alerts, but their effectiveness depends on data availability and infrastructure [17].

Air pollution remains a major global issue, with pollutants such as PM2.5, PM10, NO_2, SO_2, CO, and O_3 causing serious health problems. Accurate prediction is essential for effective mitigation strategies [18]. Although many ML and DL models have been proposed, challenges remain due to the complex and interdependent nature of air pollution data. Small changes in input data can lead to large variations in predictions, affecting model stability.

To address these issues, advanced models like variational autoencoders (VAEs) have been introduced. The NF-VAE model, for example, can capture complex relationships in multivariate data by learning structured representations. It has been shown to outperform traditional models such as LSTM and GRU in terms of accuracy and reliability [18].

Traditional air quality monitoring systems rely on fixed sensors, which are expensive and provide limited spatial coverage [19]. To improve this, IoT-based systems using both fixed and mobile sensors have been developed. Mobile sensors, mounted on vehicles or drones, help capture localized variations in air quality, providing more detailed data.

A hybrid system combining fixed and mobile sensors offers better monitoring and prediction by improving spatial coverage and data quality. Machine learning models such as SVR, Random Forest, and Gradient Boosting are used for prediction, with gradient boosting showing strong performance in handling sudden changes in pollution levels.

Overall, integrating IoT with advanced ML models significantly improves air quality prediction. The use of mobile sensors and visualization tools like heatmaps also helps in better understanding pollution patterns, making these systems useful for both researchers and the general public [19].

5. Methodology

The main aim of this approach is to build a comparative framework that evaluates multiple models based on their prediction accuracy as well as their computational efficiency. The methodology is divided into several important stages, including data collection, data preprocessing, model development, training, evaluation, and performance comparison. The models are tested in both standalone systems and distributed computing environments to analyze their efficiency under different conditions.

The overall objective is to identify the most suitable model that can provide high prediction accuracy while requiring less computation time, making it practical for real-world smart city applications.

International Journal of Science and Engineering Applications
Volume 15-Issue 05, 05 – 17, 2026, ISSN:- 2319 - 7560
DOI: 10.7753/IJSEA1505.1002

5.1 Overall Framework

The proposed system follows a well-defined and organized pipeline. It begins with collecting air quality data from reliable sources and then preprocessing the data to improve its quality. After that, different machine learning models are developed and trained using both standalone systems and distributed platforms such as Apache Spark.

Data Collection → Data Preprocessing → Model Development → Training (Standalone & Apache Spark) → Prediction → Performance Evaluation → Comparative Analysis → Optimal Model Selection

Figure 5.1: Proposed model

5.2 Framework Overview

The proposed framework provides a structured approach for processing data and ensures a fair comparison of different machine learning models under the same conditions.

5.3 Data Collection

The process begins with collecting air quality data from reliable sources such as IoT-based sensors and publicly available datasets. The dataset includes important environmental factors like particulate matter (PM2.5 and PM10), carbon monoxide (CO), nitrogen dioxide (NO_2), sulfur dioxide (SO_2), temperature, and humidity. These factors play a key role in determining air quality levels and are used to predict the Air Quality Index (AQI).

5.4 Data Preprocessing

Since raw environmental data often contains missing values and noise, preprocessing is necessary to improve its quality. This step involves cleaning the data by handling missing values through interpolation or averaging methods, removing outliers to reduce noise, and scaling the data using normalization or standardization techniques. Additionally, feature selection is performed to identify the most relevant variables that significantly influence air quality. These steps ensure that the data is well-prepared for model training.

5.5 Model Development

In this study, several machine learning regression models are developed, including Linear Regression, Decision Tree, Random Forest, Support Vector Regression, and Gradient Boosting. These models are widely used for prediction tasks and are suitable for analyzing environmental data. All models are trained using the same dataset to maintain consistency in evaluation.

5.6 Model Training

5.6.1 Standalone Training

Initially, the models are trained on a single machine using standard machine learning tools. This approach helps in establishing baseline results in terms of training time and prediction performance. However, it may not be efficient for handling large-scale datasets.

5.6.2 Apache Spark-Based Training

To overcome the limitations of standalone systems, Apache Spark is used for distributed model training. Spark allows data to be processed in parallel across multiple systems, making it faster and more scalable. The models are implemented using Spark MLlib, enabling a comparison between standalone and distributed environments.

5.7 Prediction

Once trained, the models are used to predict air quality indicators such as AQI and PM2.5 levels based on input features like pollutant concentrations and weather conditions.

5.8 Performance Evaluation

The performance of each model is assessed using standard evaluation metrics such as Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). MAE calculates the average difference between predicted and actual values, while RMSE gives more importance to larger errors. In addition, the time taken for model training and prediction is also measured to evaluate computational efficiency.

5.9 Comparative Analysis

A detailed comparison of all models is carried out based on their accuracy, execution time, and ability to handle large datasets. The results are presented in the form of tables and graphs to make the comparison clear and easy to understand.

5.10 Optimal Model Selection

Finally, the best-performing model is selected based on its accuracy and computational efficiency. The model with the lowest error values and fastest processing time is considered the most suitable for real-time air quality prediction in smart city environments.

6. Results and Discussion

This Section discusses the experimental results obtained from implementing different machine learning regression models for air quality prediction. The models are evaluated based on both their prediction accuracy and computational efficiency. The main goal of this analysis is to compare the performance of these models and determine which one is most effective in predicting air quality parameters, particularly PM2.5 concentration. To assess performance, metrics such as Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Cross-Validation RMSE (CV RMSE), and processing time are used.

6.1 Experimental Setup

The experiments were carried out using a real-world air quality dataset that includes parameters such as PM2.5, PM10, CO, NO_2, SO_2, and temperature. Before training the models, the dataset was preprocessed to handle missing values using mean imputation. It was then divided into training and testing sets in an 80:20 ratio.

Five machine learning models were selected for comparison: Linear Regression, Decision Tree, Random Forest, Gradient Boosting, and XGBoost. To ensure a reliable and unbiased evaluation, 5-fold cross-validation was applied. This approach helps in assessing how well each model performs on unseen data. In addition to accuracy, the execution time of each model was also recorded to evaluate its computational performance.

6.2 Performance Metrics

International Journal of Science and Engineering Applications
Volume 15-Issue 05, 05 – 17, 2026, ISSN:- 2319 - 7560
DOI: 10.7753/IJSEA1505.1002

The effectiveness of the models was measured using several standard evaluation metrics:

6.2.1 Mean Absolute Error (MAE): This metric calculates the average difference between the predicted and actual values, providing a straightforward measure of prediction accuracy.

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |\, x_i - x|$$

6.2.2 Root Mean Square Error (RMSE): RMSE measures the square root of the average squared differences, giving more weight to larger errors and highlighting significant prediction deviations.

$$RMSE = \sqrt{\frac{1}{2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

6.2.3 Cross-Validation RMSE (CV RMSE): This metric offers a more reliable estimate of model performance by evaluating it across multiple data splits.

6.2.4 Processing Time: This measures how long each model takes to train and make predictions, helping assess computational efficiency.

6.2.5 Model Results Comparison Table

Model	MAE	Test RMSE	CV RMSE	Time (sec)
Linear Regression	21.31	33.43	33.63	0.10
Decision Tree	16.12	29.18	27.65	0.66
Random Forest	14.46	25.74	23.91	115.27
Gradient Boosting	15.55	26.38	24.34	46.63
XGBoost	15.08	25.20	23.82	3.09

6.3 Discussion

6.3.1 Accuracy Analysis

The results clearly show that ensemble learning models perform better than traditional machine learning approaches. Linear Regression recorded the highest error values, with an RMSE of 33.43, which suggests that it is not capable of capturing the complex and nonlinear patterns present in air quality data. The Decision Tree model performed better than Linear Regression, but its performance was still limited due to its tendency to overfit the data.

Among all the models, XGBoost achieved the best results, with the lowest Test RMSE (25.20) and CV RMSE (23.82). This indicates that it not only predicts accurately but also generalizes well to unseen data. Random Forest also showed strong performance, particularly with the lowest MAE (14.46), although its RMSE was slightly higher compared to XGBoost. Gradient Boosting delivered reasonably good results but did not match the performance of XGBoost.

6.3.2 Computational Efficiency

Computation time plays an important role, especially in real-time air quality monitoring systems. The results highlight a noticeable variation in execution time across different models. Random Forest required the longest time (115.27 seconds), followed by Gradient Boosting (46.63 seconds). On the other hand, XGBoost completed its execution in just 3.09 seconds, making it highly efficient.

Although Linear Regression and Decision Tree models were faster, their lower accuracy makes them less suitable for practical use. Overall, XGBoost offers the

International Journal of Science and Engineering Applications
Volume 15-Issue 05, 05 – 17, 2026, ISSN:- 2319 - 7560
DOI: 10.7753/IJSEA1505.1002

most effective balance between prediction accuracy and computational speed.

6.3.3 Model Comparison

The overall comparison leads to a few important observations. Traditional models like Linear Regression struggle to handle the complexity of environmental data. Single-tree models such as Decision Trees show some improvement but are prone to overfitting.

In contrast, ensemble methods—including Random Forest, Gradient Boosting, and XGBoost—consistently deliver better performance. Among these, XGBoost stands out due to its advanced boosting technique, built-in regularisation, and efficient computation, making it the most suitable choice for air quality prediction tasks.

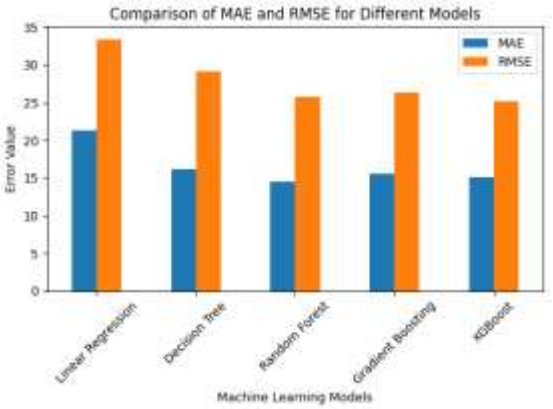

Figure 6.1 MAE and RMSE Comparison

Figure 6.1 presents a comparison of Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) across different machine learning models. From the figure, it is clear that Linear Regression has the highest error values, which suggests that it struggles to capture the complex, nonlinear patterns in air quality data. The Decision Tree model shows better performance than Linear Regression, but its error values are still relatively moderate.

Random Forest stands out by achieving the lowest MAE, indicating strong overall prediction accuracy, while XGBoost records the lowest RMSE, highlighting its ability to handle larger prediction errors more effectively and generalise well to unseen data. Gradient Boosting also performs well and remains competitive, although it does not exceed the performance of XGBoost. Overall, the results demonstrate that ensemble learning models consistently outperform traditional methods in air quality prediction tasks.

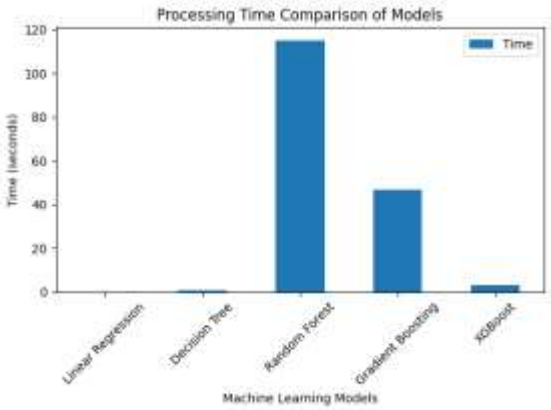

Figure 6.2: Processing Time Comparison

Figure 6.2 shows the comparison of processing time taken by different machine learning models. From the figure, it can be seen that Random Forest and Gradient Boosting take considerably more time to execute, mainly due to their complex ensemble structures. In contrast, Linear Regression and Decision Tree are the fastest models in terms of computation time, but their lower prediction accuracy makes them less practical for real-world applications.

XGBoost, however, achieves a good balance between speed and accuracy. It delivers high prediction performance while keeping the processing time relatively low, making it a suitable choice for real-time air quality prediction systems.

7. Challenges

7.1 Data Quality and Availability

Air quality datasets often contain missing values, noise, and inconsistencies. Handling these issues through preprocessing (such as imputation and outlier removal) can affect the accuracy of the models. Additionally, access to high-quality, real-time data is limited in many regions.

7.2 Nonlinear and Complex Data Patterns

Environmental data is highly complex and influenced by multiple factors such as weather conditions and human activities. Capturing these nonlinear relationships is difficult, especially for simpler models like Linear Regression.

7.3 Model Overfitting

Some models, particularly Decision Trees, tend to overfit the training data. This reduces their ability to generalize well to unseen data, leading to less reliable predictions.

7.4 Computational Cost

Advanced models such as Random Forest and Gradient Boosting require significant computational time and resources, especially when working with large datasets. This can limit their use in real-time applications.

7.5 Parameter Tuning Complexity

Models like XGBoost require careful tuning of hyperparameters to achieve optimal performance. This process can be time-consuming and may require expertise.

Scalability Issues

Handling large-scale datasets efficiently is a challenge in standalone systems. Although distributed frameworks like Apache Spark help address this issue, they introduce additional complexity in implementation.

Real-Time Implementation Constraints

Deploying models in real-time air quality monitoring systems requires a balance between accuracy and speed. Ensuring low latency while maintaining high prediction accuracy remains a challenge.

8.Conclusion

This study explored the use of different machine learning regression techniques for predicting air quality and carried out a comparative analysis to determine the most accurate and efficient model. The models considered in this work include Linear Regression, Decision Tree, Random Forest, Gradient Boosting, and XGBoost. Their performance was evaluated using key metrics such as Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Cross-Validation RMSE (CV RMSE), and processing time.

The results clearly indicate that traditional approaches like Linear Regression are not well-suited for air quality prediction, as they fail to capture the complex and nonlinear nature of environmental data. The Decision Tree model showed some improvement, but its tendency to overfit reduced its reliability. In contrast, ensemble methods such as

Random Forest and Gradient Boosting provided better accuracy by combining multiple models and minimizing prediction errors.

Among all the models, XGBoost delivered the best overall performance. It achieved the lowest RMSE and CV RMSE values, which reflects both high prediction accuracy and strong generalization to unseen data. Although Random Forest recorded a slightly lower MAE, the difference was minor and did not outweigh the overall advantages of XGBoost. Moreover, XGBoost required significantly less computation time compared to Random Forest and Gradient Boosting, making it more suitable for real-time applications.

Based on both accuracy and efficiency, XGBoost can be considered the most effective model for air quality prediction. The findings of this study highlight the importance of advanced ensemble techniques in handling complex and nonlinear environmental data.

6.2 Future Scope

While this study provides valuable insights, there are several directions for future research and improvement. One promising area is the application of deep learning models such as Long Short-Term Memory (LSTM) and Recurrent Neural Networks (RNN), which are well-suited for time-series forecasting and can better capture temporal patterns in air quality data.

Another important extension is the integration of real-time data from IoT-based sensors to develop a dynamic and continuously updating prediction system. This would enhance the practical applicability of the model in smart city environments.

Future work can also include additional influencing factors such as wind speed, traffic density, industrial emissions, and geographical conditions to further improve prediction accuracy. Moreover, advanced hyperparameter tuning techniques like Grid Search and Bayesian Optimization can be applied to optimize model performance.

In addition, deploying the model as a web or mobile-based application can improve accessibility and usability for end users. Researchers can also explore hybrid or ensemble deep learning approaches to combine the strengths of multiple models.

Finally, implementing the system on large-scale distributed platforms can further enhance scalability and efficiency, enabling real-time processing of massive datasets for smart city applications.

References

[1]. Mazinani, A., Antonucci, D., Pau, D. P., Davoli, L., & Ferrari, G. (2025). Air quality prediction via embedded ML/DL and quantized models. *IEEE Access, 13*.

[2]. Ameer, S., Shah, M. A., Khan, A., Song, H., Maple, C., Islam, S. U., & Asghar, M. N. (2019). Comparative analysis of machine learning techniques for predicting air quality in smart cities. *IEEE Access, 7*, 128325–128338.https://doi.org/10.1109/ACCESS.2019.2925082.

[3]. Al-Eidi, S., Amsaad, F., Darwish, O., Tashtoush, Y., Alqahtani, A., &

Niveshitha, N. (2023). Comparative analysis study for air quality prediction in smart cities using regression techniques. *IEEE Access, 11*, 115140–115143.

[4]. J. Kalajdjieski, K. Trivodaliev, G. Mirceva, S. Kalajdziski, and S. Gievska,"A complete air pollution monitoring and prediction framework"*IEEE Access*, vol.11, pp. 88730–88744, year-2023.

[5]. C. Liu, G. Pan, D. Song, and H. Wei, "Air Quality Index Forecasting via Genetic Algorithm-Based Improved Extreme Learning Machine," *IEEE Access*, vol. 11, pp. 67086–67090, 2023.

[6]. Y. Du, Y. Xu, Y. Li, Z. Liu, and L. Wang, "A Novel Combined Prediction Scheme Based on CNN and LSTM for Urban PM2.5 Concentration," *IEEE Access*, vol. 7, pp. 20050–20059, 2019.

[7]. S. K. Natarajan, P. Shanmurthy, D. Arockiam, B. Balusamy, and S. Selvarajan, "Optimized machine learning model for air quality index prediction in major cities in India," *Scientific Reports*, vol. 14, 2024.

[8]. Y. Zhou, S. De, E. G. Perera, and K. Moessner, "Data-driven air quality characterization for urban environments," *IEEE Access*, vol. 6, pp. 77996–78006, 2018.

[9]. Y. Yang, Z. Zheng, K. Bian, L. Song, and Z. Han, "Real-time profiling of fine-grained air quality index distribution using UAV sensing," *IEEE Internet of Things Journal*, vol. 5, no. 1, pp. 186–198, 2018.

[10]. S. Ameer *et al.*, "Comparative analysis of machine learning techniques for predicting air quality in smart cities," *IEEE Access*, vol. 7, pp. 128325–128338, 2019.

[11]. Y.-C. Lin, S.-J. Lee, and C.-H. Wu, "Air quality prediction by neuro-fuzzy modelling approach," Applied Soft Computing, vol. 86, 2020.

[12]. R. Janarthanan, P. Partheeban, and P. Navin Elamparithi, "A deep learning approach for prediction of air quality index in a metropolitan city," *Sustainable Cities and Society*, vol. 67, 2021.

[13]. D. Saravanan and K. Santhosh Kumar, "IoT-based improved air quality index prediction using hybrid FA-ANN-ARMA model," *Materials Today: Proceedings*, vol. 56, 2021.

[14]. C.-Liu, T.-C. Lin, and P.-T. Chiueh, "Spatio-temporal prediction of urban air quality using support vector machine," *Urban Climate*, vol. 41, 2021.

[15]. J. Wang *et al.*, "A hybrid air quality index prediction model based on CNN and attention gate unit," *IEEE Access*, vol. 10, 2022.

[16]. N. Sarkar, R. Gupta, and M. C. Govil, "Air quality index prediction using an effective hybrid deep learning model," *Environmental Pollution*, vol. 315, 2022.

[17]. Y. Hu, X. Chen, and H. Xia, "A hybrid prediction model of air quality based on spatio-temporal feature extraction," *Atmospheric Pollution Research*, vol. 14, 2023.

[18]. P. Dey, S. Dev, and B. S. Phelan, "Predicting multivariate air pollution: A Gaussian-mixture nested factorial variational autoencoder approach," *IEEE Geoscience and Remote Sensing Letters*, vol. 21, 2024, Art. no. 1002805,doi:10.1109/LGRS.2024.3416343.

[19]. D. Zhang and S. S. Woo, "Real-time localized air quality monitoring and prediction through mobile and fixed IoT sensing network," *IEEE Access*, vol. 8, pp. 89584–89594, May 2020, doi:10.1109/ACCESS.2020.2993547.