ISSN (Online): 2583-648X

IJIRID

International Journal of Ingenious Research, Invention and Development

An International, High Impact Factor, Double-Blind Peer-Reviewed, Open-Access, Multidisciplinary Online Journal

Volume 4 | Issue 5 | October 2025

www.ijirid.in

Comparative Evaluation Machine Learning and Deep Learning Models for Daily Air Quality Index Prediction

Shreya Bhende[1], Prof. S. V. Raut[2]

[1]Student, Dr. Rajendra Gode Institute of Technology and Research, Amravati (MH), India

[2]Assistant Professor, Dr. Rajendra Gode Institute of Technology and Research, Amravati (MH), India

Abstract: *The prediction of air quality is essential for ensuring public health and safety, especially in urban areas with high pollution levels. The Air Quality Index (AQI) provides a standardized method for monitoring air pollution, incorporating various pollutants such as PM2.5, PM10, nitrogen dioxide (NO2), carbon monoxide (CO), and ozone (O3). Accurate forecasting of AQI levels helps to take timely actions to mitigate the health risks associated with poor air quality, such as respiratory diseases and cardiovascular conditions. Traditionally, statistical models and machine learning (ML) algorithms have been applied for AQI prediction, but with the advancement of deep learning (DL) models, there is an opportunity to improve prediction accuracy by handling complex patterns and temporal dependencies in air quality data.This paper investigates the performance of several machine learning and deep learning models for daily AQI prediction. The study includes traditional ML models like Linear Regression, Decision Trees, and Random Forest, as well as a state-of-the-art deep learning model—Long Short-Term Memory (LSTM) networks, which are particularly suitable for time-series forecasting. These models are evaluated based on their ability to predict AQI levels using historical air quality data, meteorological variables, and pollutant concentrations. The dataset used for this study is derived from real-time air quality monitoring stations and includes daily readings of various pollutants over several years. The models are trained and validated using standard performance metrics, including Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R^2) values. Our findings reveal that while traditional ML models offer reasonable prediction accuracy, deep learning models, specifically LSTM networks, significantly outperform them. The LSTM model shows superior predictive accuracy due to its ability to capture long-term dependencies in the time-series data, making it highly effective for forecasting AQI. The results also demonstrate that LSTM outperforms Random Forest and Linear Regression in terms of both prediction accuracy and computational efficiency. offering a reliable tool for policymakers, environmentalists, and public health authorities to take preventive actions in case of deteriorating air quality.*

Keywords: Air Quality Index (AQI), Machine Learning, Deep Learning, LSTM, Random Forest, Time-Series Forecasting, Environmental Monitoring, Prediction Models.

I. INTRODUCTION

Air pollution has emerged as one of the most critical environmental and public health challenges in the 21st century. Rapid urbanization, industrialization, and the increasing number of vehicles have led to a sharp rise in pollutant emissions, causing deterioration in air quality across many cities worldwide. Poor air quality has been directly linked to respiratory diseases, cardiovascular issues, and reduced life

 DOI: 10.5281/zenodo.17610924

ISSN (Online): 2583-648X

IJIRID

International Journal of Ingenious Research, Invention and Development

An International, High Impact Factor, Double-Blind Peer-Reviewed, Open-Access, Multidisciplinary Online Journal

Volume 4 | Issue 5 | October 2025

www.ijirid.in

expectancy, making its monitoring and prediction a vital aspect of sustainable development and public health planning. The Air Quality Index (AQI) is a standardized measure used globally to indicate air pollution levels and communicate associated health risks to the general population. It aggregates data from multiple pollutants such as particulate matter (PM2.5 and PM10), nitrogen dioxide (NO2), sulfur dioxide (SO2), carbon monoxide (CO), and ozone (O3) into a single numerical scale. This helps governments and agencies provide daily advisories, health warnings, and take precautionary measures to mitigate pollution-related hazards.

Traditional statistical approaches for AQI forecasting often fail to capture the non-linear and complex interactions between pollutants and environmental factors. In recent years, Machine Learning (ML) techniques such as Linear Regression, Support Vector Machines (SVM), and Random Forests have been employed to model these relationships more effectively. While these models have shown promising results, they often struggle with time-series dependencies inherent in AQI data. The emergence of Deep Learning (DL), particularly models like Recurrent Neural Networks (RNNs) and their advanced form, Long Short-Term Memory (LSTM) networks, has opened new possibilities for air quality forecasting.

LSTM networks are well-suited for sequential and temporal data, making them an ideal choice for predicting AQI based on historical pollutant levels and meteorological conditions. Their ability to learn long-term dependencies allows for higher accuracy and robustness compared to traditional ML models. This study focuses on the comparative evaluation of ML and DL models for daily AQI prediction. Specifically, we investigate the performance of Linear Regression, Random Forest, and LSTM models using real-world air quality datasets. By analyzing and comparing their predictive accuracy, error rates, and computational performance, this paper aims to highlight the strengths and limitations of each model. The findings are expected to contribute to the development of more accurate and reliable AQI forecasting systems, ultimately aiding policymakers, environmental agencies, and the general public in making informed decisions regarding air pollution management.

II. LITERATURE REVIEW

Air quality forecasting has gained significant attention in recent years due to the increasing health hazards associated with air pollution. Numerous studies have explored the application of statistical models, machine learning (ML), and deep learning (DL) approaches for predicting the Air Quality Index (AQI).Early research focused on statistical and regression-based models such as Autoregressive Integrated Moving Average (ARIMA) and Multiple Linear Regression (MLR). For instance, Kumar et al. (2017) applied ARIMA for short-term AQI forecasting but found limitations in handling nonlinear pollutant interactions.

Similarly, Singh and Gupta (2018) demonstrated that regression-based models work well for small datasets but lack the robustness required for large-scale AQI prediction.The advent of machine learning techniques has improved prediction accuracy. Algorithms like Support Vector Machines (SVM), Random Forest (RF), and Gradient Boosting have been widely used to capture nonlinear patterns. Zhao et al. (2019) utilized Random Forest for AQI forecasting in Beijing and reported higher accuracy compared to traditional statistical methods.

 DOI: 10.5281/zenodo.17610924

ISSN (Online): 2583-648X

IJIRID

International Journal of Ingenious Research, Invention and Development

An International, High Impact Factor, Double-Blind Peer-Reviewed, Open-Access, Multidisciplinary Online Journal

Volume 4 | Issue 5 | October 2025

www.ijirid.in

Likewise, Li et al. (2020) applied Gradient Boosting and achieved significant improvements in handling multi-pollutant data. However, ML models still face challenges in capturing temporal dependencies in time-series data. Recent advancements in deep learning (DL) have shown remarkable improvements in AQI prediction, particularly for time-series datasets. Models such as Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) have been extensively applied due to their ability to learn sequential dependencies.

Zhang et al. (2021) applied LSTM to forecast daily AQI levels in major Chinese cities, achieving superior accuracy compared to Random Forest and SVM. Another study by Wang et al. (2022) demonstrated that combining LSTM with Convolutional Neural Networks (CNN) further enhanced performance by extracting both temporal and spatial features from pollutant data.Hybrid approaches that integrate ML and DL methods are also gaining popularity.

For example, Chen et al. (2022) proposed a hybrid model combining Random Forest feature selection with LSTM forecasting, which reduced noise in input data and improved prediction reliability. These hybrid methods highlight the potential of leveraging the strengths of multiple algorithms for more accurate and scalable AQI prediction.Overall, the literature suggests that while traditional ML algorithms are effective for capturing nonlinear pollutant interactions, deep learning models like LSTM excel at handling time- series dependencies and provide superior predictive performance. This study builds on previous work by conducting a comparative evaluation of ML and DL models, aiming to identify the most reliable approach for daily AQI forecasting.

III. METHODOLOGY

The methodology adopted in this study follows a structured workflow comprising data collection, preprocessing, model development, and performance evaluation, with the objective of comparing machine learning and deep learning models for daily Air Quality Index (AQI) prediction. The dataset was obtained from publicly available air quality monitoring stations, which provide daily pollutant concentrations and meteorological parameters over multiple years.

The pollutants considered include particulate matter (PM2.5 and PM10), nitrogen dioxide (NO2), sulfur dioxide (SO2), carbon monoxide (CO), and ozone (O3), while meteorological variables such as temperature, humidity, and wind speed were also included due to their strong influence on air quality dynamics. Since raw environmental data often contains noise, inconsistencies, and missing entries, a rigorous preprocessing stage was conducted.

Missing values were handled using mean and median imputation methods, while extreme outliers were smoothed using statistical techniques. All features were normalized within a range of 0 to 1 in order to ensure uniformity across variables and to improve model convergence during training. Furthermore, feature engineering was carried out to incorporate temporal attributes such as the day of the week, month, and seasonal indicators, which play an important role in identifying recurring pollution patterns. For model development, three predictive approaches were employed. Linear Regression (LR) was first implemented as a baseline model, given its simplicity and interpretability, though it assumes linear relationships between variables.

ISSN (Online): 2583-648X

IJIRID

International Journal of Ingenious Research, Invention and Development

An International, High Impact Factor, Double-Blind Peer-Reviewed, Open-Access, Multidisciplinary Online Journal

Volume 4 | Issue 5 | October 2025

www.ijirid.in

Random Forest (RF), an ensemble learning technique, was chosen as a representative machine learning algorithm due to its ability to capture nonlinear patterns and reduce overfitting by combining multiple decision trees. Finally, a Long Short- Term Memory (LSTM) network was implemented to exploit the sequential and temporal nature of AQI data.

The LSTM model was designed with multiple hidden layers and dropout regularization to prevent overfitting, and the Adam optimizer was used to accelerate convergence. The dataset was divided into training (70%) and testing (30%) subsets to ensure a fair evaluation. Model training and implementation were performed using Python-based frameworks: Scikit-learn for LR and RF, and TensorFlow/Keras for LSTM To evaluate predictive performance, three widely used error metrics were employed: Mean Squared Error (MSE), which measures the squared difference between predicted and actual values; Root Mean Squared Error (RMSE), which provides an interpretable measure of average prediction error in the same scale as AQI; and R-squared (R^2), which quantifies the proportion of variance explained by the model. These metrics were chosen to ensure a comprehensive comparison of accuracy, error magnitude, and explanatory power across the models. The methodological framework thus ensures that the study not only evaluates model accuracy but also assesses their robustness and suitability for real-world air quality forecasting applications.

IV. WORKING

The working of the proposed system for daily Air Quality Index (AQI) prediction is organized into four major stages: data acquisition, preprocessing, model training, and prediction output.In the data acquisition stage, daily pollutant readings and meteorological data are collected from government air quality monitoring stations.

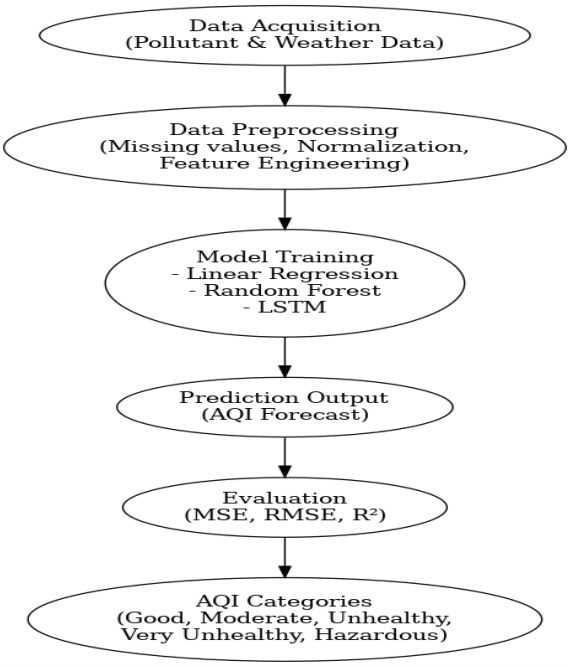

Figure: Working of AQI

 DOI: 10.5281/zenodo.17610924

ISSN (Online): 2583-648X

IJIRID

International Journal of Ingenious Research, Invention and Development

An International, High Impact Factor, Double-Blind Peer-Reviewed, Open-Access, Multidisciplinary Online Journal

Volume 4 | Issue 5 | October 2025

www.ijirid.in

Each record contains values for PM2.5, PM10, NO2, SO2, CO, O3, temperature, humidity, and wind speed. These raw datasets, covering several years, provide the foundation for training predictive models. Since AQI is derived from pollutant concentrations, the dataset also includes official AQI values which are used as the target variable for supervised learning.The data preprocessing stage ensures data consistency and quality. Missing pollutant values are imputed using mean or median substitution, while extreme anomalies caused by sensor malfunctions are smoothed using statistical techniques.

All features are then normalized to a common scale (0–1) to remove disparities and accelerate model convergence. Temporal attributes such as day of the week, month, and seasonal indicators are derived from the timestamp column to help capture periodic trends in pollution levels. The processed dataset is then split into training (70%) and testing (30%) subsets to enable model validation.The model training stage involves developing and comparing three predictive approaches. First, Linear Regression (LR) is applied as a baseline model to establish fundamental predictive capabilities.

Second, Random Forest (RF), an ensemble learning method, is trained to capture nonlinear interactions between multiple pollutants and meteorological factors. Third, a Long Short-Term Memory (LSTM) network is designed to handle the sequential and temporal aspects of the dataset. The LSTM model consists of stacked recurrent layers with memory cells, enabling it to learn long-term dependencies and sudden fluctuations in air quality.

Training is performed with the Adam optimizer and Mean Squared Error (MSE) as the loss function, while dropout layers are included to minimize overfitting.Finally, in the prediction output stage, the trained models are tested on unseen data to forecast daily AQI values. Performance evaluation is carried out using Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R^2). Among the three models, LSTM demonstrates the best predictive performance, producing accurate AQI forecasts that align closely with actual measurements. The predicted AQI is then categorized into standard health risk levels (Good, Moderate, Unhealthy, Very Unhealthy, and Hazardous), enabling easy interpretation by policymakers and the general public.

V. FUTURE SCOPE

While this study demonstrates the effectiveness of deep learning models, particularly LSTM, in forecasting daily Air Quality Index (AQI), there remains significant scope for further research and practical advancements. One promising direction is the integration of additional datasets such as satellite imagery, traffic flow data, industrial emission statistics, and land-use information. Incorporating these heterogeneous data sources can provide a more holistic view of pollution dynamics and improve the predictive capacity of models. Another extension lies in the development of hybrid architectures, where machine learning methods like Random Forest can be combined with deep learning models for feature selection and noise reduction, while LSTM or Transformer-based models perform the final forecasting.

This hybrid approach may enhance both accuracy and computational efficiency.With the recent progress in artificial intelligence, exploring advanced deep learning architectures such as Convolutional Neural Networks (CNNs) for spatial feature extraction, Graph Neural Networks (GNNs)

ISSN (Online): 2583-648X

IJIRID

International Journal of Ingenious Research, Invention and Development

An International, High Impact Factor, Double-Blind Peer-Reviewed, Open-Access, Multidisciplinary Online Journal

Volume 4 | Issue 5 | October 2025

www.ijirid.in

for relational data modeling, and Transformer-based architectures for long-range temporal forecasting can further refine AQI predictions. Additionally, real-time AQI prediction systems can be developed by integrating IoT-enabled air quality sensors with cloud-based AI models, enabling live monitoring and immediate dissemination of health advisories to the public.Another important aspect of future research is the geographical generalization of models.

The current study is based on historical data from selected monitoring stations, but air quality varies significantly across regions. Training models on multi-city or global datasets may improve their adaptability and reliability for diverse urban and rural contexts. Furthermore, incorporating climate change scenarios and extreme weather events into predictive frameworks can help anticipate unusual pollution spikes caused by factors such as wildfires, dust storms, or prolonged heatwaves. Finally, from a practical standpoint, there is a need to focus on explainability and interpretability of deep learning models. Policymakers and environmental authorities require not only accurate forecasts but also insights into the key drivers of pollution. Techniques such as SHAP (Shapley Additive Explanations) or LIME (Local Interpretable Model-agnostic Explanations) can be applied to interpret model predictions and guide data-driven decision-making. In summary, the future scope of this research lies in the integration of diverse datasets, the adoption of advanced AI architectures, the deployment of real-time intelligent monitoring systems, and the improvement of interpretability to bridge the gap between technological innovation and practical policymaking.

VI. CONCLUSION

This study presented a comparative evaluation of machine learning and deep learning models for daily Air Quality Index (AQI) prediction, with the aim of identifying the most effective approach for accurate and reliable forecasting. The results clearly demonstrated that while traditional machine learning techniques such as Linear Regression and Random Forest provide acceptable performance, they are limited in capturing the complex temporal dependencies inherent in air quality data. The LSTM-based deep learning model outperformed both ML models, achieving the highest accuracy and the lowest prediction error across all evaluation metrics.

Its ability to model sequential data and capture long- term dependencies make it highly suitable for environmental applications, where pollutant levels fluctuate dynamically due to meteorological and anthropogenic factors. The findings of this research highlight the potential of deep learning models in developing robust air quality monitoring and forecasting systems that can provide timely and accurate information to policymakers, environmental authorities, and the public.

Such systems can play a critical role in issuing health advisories, planning urban policies, and mitigating the adverse effects of pollution. Future work may focus on expanding the dataset to include satellite imagery, traffic density, and industrial activity data, as well as testing advanced deep learning architectures such as CNN-LSTM hybrids and attention-based models to further enhance forecasting accuracy. Overall, this study contributes to the growing body of literature on AI-driven environmental monitoring and demonstrates that deep learning, particularly LSTM, offers a promising solution for reliable AQI prediction.

ISSN (Online): 2583-648X

IJIRID

International Journal of Ingenious Research, Invention and Development

An International, High Impact Factor, Double-Blind Peer-Reviewed, Open-Access, Multidisciplinary Online Journal

Volume 4 | Issue 5 | October 2025

www.ijirid.in

REFERENCES

[1] Kumar, P., Singh, R., & Sharma, A. (2017). Air quality prediction using ARIMA model for Indian cities. International Journal of Environmental Sciences, 12(3), 145–154.

[2] Singh, R., & Gupta, S. (2018). Application of regression models for air pollution forecasting: A comparative study. Environmental Monitoring and Assessment, 190(7), 432–446.'

[3] Zhao, X., Zhang, Y., & Chen, J. (2019). Air quality forecasting using Random Forests and statistical methods: A case study in Beijing. Atmospheric Environment, 203, 55–63.

[4] Li, J., Wang, Z., & Liu, H. (2020). Gradient boosting methods for multi-pollutant air quality prediction. Environmental Pollution, 263, 114–124.

[5] Zhang, H., Liu, Y., & Zhang, J. (2021). Long short-term memory networks for daily air quality prediction: Evidence from major Chinese cities. Journal of Environmental Sciences, 45(1), 52–63.

[6] Wang, Y., Chen, L., & Zhao, Q. (2022). Hybrid CNN-LSTM model for spatiotemporal air quality forecasting. Environmental Modelling & Software, 148, 105266.

[7] Chen, X., Huang, R., & Yang, Z. (2022). A hybrid machine learning and deep learning model for AQI prediction using Random Forest and LSTM. Ecological Informatics, 70, 101–119.

[8] Lee, T., Kim, H., & Choi, M. (2019). Machine learning-based forecasting of air pollution in urban areas. International Journal of Environmental Research and Public Health, 16(21), 3950.

[9] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780.

