Advanced International Journal for Research (AIJFR)
E-ISSN: 3048-7641 ● Website: www.aijfr.com ● Email: editor@aijfr.com

Predictive Model for Air Pollutants Concentration Using Machine Learning

Mr. Piyoosh Awthare[1], Devanshu Pote[2], Prerna Hiradkar[3], Manthan Shinde[4]

[1]Assistant Professor, Department of Computer Technology, KITS Ramtek
[2,3,4] U. G. Student, Department of Computer Technology, KITS Ramtek

Abstract
The Air Quality Index (AQI) is widely used by government agencies to communicate air pollution severity to the public and is derived from pollutants such as sulphur dioxide, nitrogen dioxide, ozone, PM_{10} and $PM_{2.5}$. Although several AQI calculation methods have been proposed, no universally applicable approach exists. In this work, standard classification and regression techniques are enhanced by incorporating temporal correlations among sub-models. Historical and meteorological data are utilized to predict pollutant concentrations and forecast AQI over short-term and long-term horizons. The system is trained and evaluated using air pollution data from 2019 to 2024 covering ten Indian cities. Regression-based machine learning models and R-ARIMA time-series forecasting with moving average smoothing are employed to generate daily and monthly AQI predictions. The results demonstrate reliable forecasting performance and support effective air quality monitoring and planning.

Keywords: Air Quality Index, Machine Learning, Time Series Forecasting, Environmental Analytics, ARIMA

1. Introduction

Rapid industrial growth, increasing vehicular density, and unplanned urban expansion have significantly deteriorated ambient air quality in Indian metropolitan regions. Elevated concentrations of gaseous pollutants and particulate matter have been consistently linked to adverse respiratory and cardiovascular outcomes, particularly among vulnerable populations. As a result, air quality monitoring and prediction have become essential components of environmental management and public health planning.

The Air Quality Index (AQI) is widely adopted by regulatory agencies as a standardized indicator to communicate pollution severity. However, conventional AQI systems are primarily limited to reporting current or historical observations, which restricts their usefulness for proactive intervention. With the growing availability of long-term environmental datasets and advances in computational intelligence, predictive modelling techniques have emerged as effective tools for forecasting air quality trends and supporting informed decision-making.

Time-series forecasting techniques, including ARIMA and Triple Exponential Smoothing, are applied to predict daily AQI up to forty-five days ahead and monthly AQI up to one year in advance. Seasonal

variations are analysed to identify significant pollutant fluctuations, providing actionable insights for proactive air quality management and long-term environmental planning.

An intelligent air quality prediction system is developed by combining advanced machine learning algorithms with historical data from major Indian cities. Accurate air quality forecasting and pollutant trend analysis are enabled through the application of state-of-the-art ensemble and time-series methods, supporting informed decision-making and improved public health outcomes.

The objectives of this study include the development of a machine learning–based system for accurate AQI prediction, the application of time-series forecasting techniques for short-term and long-term air quality trend analysis and the generation of AQI-based insights to support public health and urban planning decisions.

2. Literature Review

2.1 Forecasting and Prediction of Air Pollutant Concentrations

In a large-scale study focused on Indian urban environments, machine learning techniques were applied to predict air pollutant concentrations using multi-year meteorological data collected from numerous cities. Comparative evaluation of multiple algorithms indicated that tree-based models demonstrated strong predictive capability, while time-series forecasting methods were effective in capturing seasonal pollution variations over short-term and annual horizons (Sharma et al., 2021).

2.2 Air Pollution Prediction

A data-driven air pollution prediction framework was developed using several years of observations from multiple Indian cities, with emphasis placed on feature relevance and data balance. Ensemble learning techniques achieved superior alignment between observed and predicted AQI values and a distinct reduction in pollutant concentrations was observed during the year 2020, highlighting the impact of large-scale activity changes (Kumar and Pande, 2022).

3. Methodology

Multi-year pollutant and temporal data from several Indian cities were processed from separate CSV files. The recorded attributes included timestamp, $PM_{2.5}$, PM_{10}, NO_2, NH_3, SO_2, CO, O_3 and city identifiers. Data integrity was ensured through systematic aggregation, duplicate record removal, missing-value imputation and validation of date ranges. Following data cleaning, feature engineering was performed to generate seasonal indicators and time-based variables to enhance model learning.

Ensemble-based regression models were trained and evaluated for AQI prediction using the processed dataset. Model performance was assessed against persistence baselines using standard evaluation metrics, including Mean Absolute Error, Root Mean Square Error and the coefficient of determination (R^2). To support temporal prediction, a forecasting module was incorporated to enable multi-day AQI estimation with smoothing techniques applied to improve trend stability.

3.1 System Design and Architecture

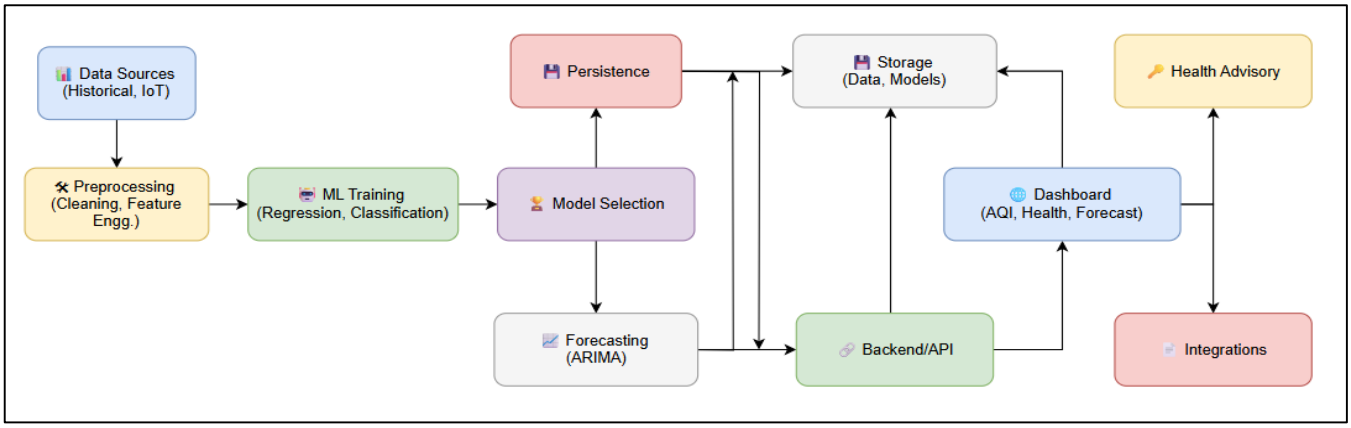
The system architecture is designed to achieve scalability, robustness and functional efficiency across all processing stages. A modular design approach is used to support structured data handling, model training and forecasting operations.

3.1.1 System Architecture

A comprehensive AI- and data mining–based pollution prediction architecture is employed, integrating data acquisition, preprocessing, machine learning and forecasting components. Environmental data from historical records and sensor sources are aggregated and subjected to cleaning, normalization and feature generation to ensure data consistency.

Following preprocessing, regression and classification models are trained to learn pollutant behaviour and AQI patterns. Model selection is performed using defined performance metrics and the selected models along with processed data are preserved through a persistence layer for reuse. In parallel, ARIMA-based forecasting is applied to estimate short-term and long-term AQI trends.

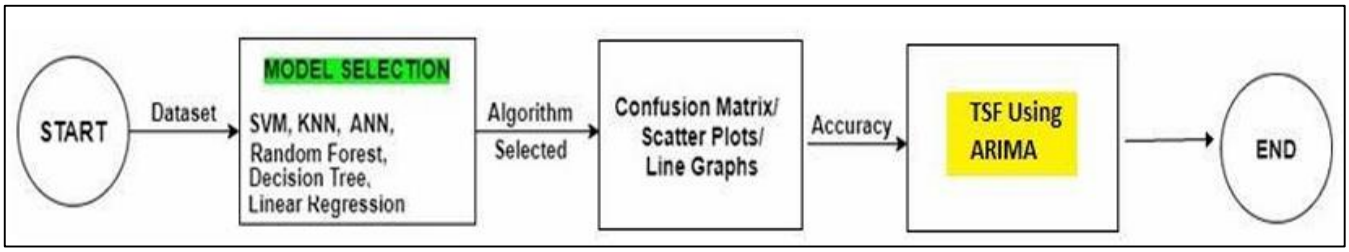
Figure 1: System Architecture



3.1.2 Model Selection

Model selection is conducted using widely adopted machine learning algorithms, including Linear Regression, Support Vector Machine, K-Nearest Neighbour, Decision Tree and Random Forest. Evaluation is performed on a representative subset of the dataset spanning the period from 2016 to 2018. The data are partitioned into training and testing sets using a 70:30 ratio and validation is applied to minimize overfitting and underfitting. Model performance is analysed using confusion matrices and standard evaluation metrics.

Figure 2: Model Selection Flowchart



The proposed methodology is applied to eight major Indian cities, including Delhi, Mumbai, Chennai, Hyderabad, Jaipur, Lucknow, Gwalior and Visakhapatnam. AQI values are predicted over short-term and long-term horizons using machine learning models combined with ARIMA-based forecasting. Moving average smoothing is applied to reduce noise and improve trend estimation, as illustrated in the corresponding flowchart.

4. Results

4.1 Model Score Comparison

Figure 3: MAE and R^2 Score Comparison

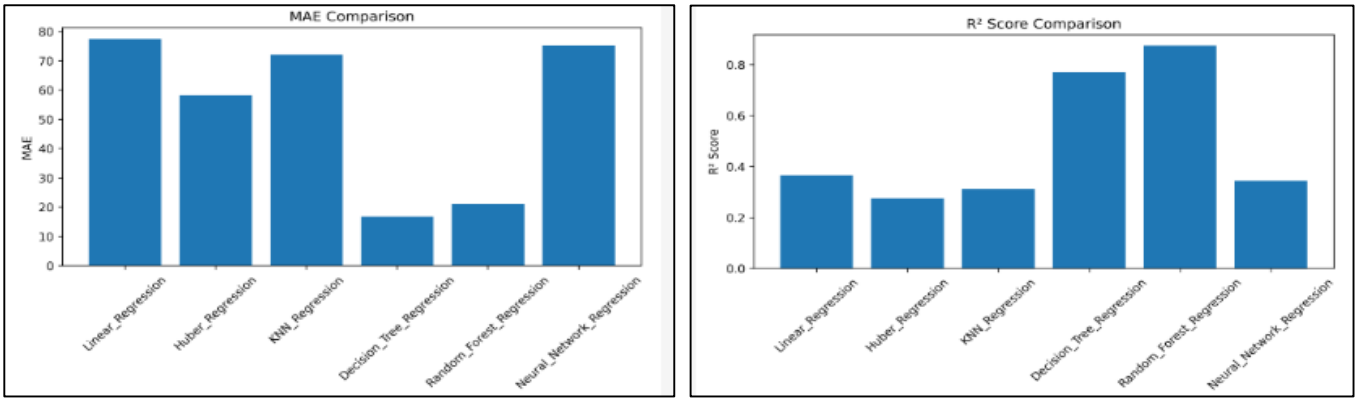
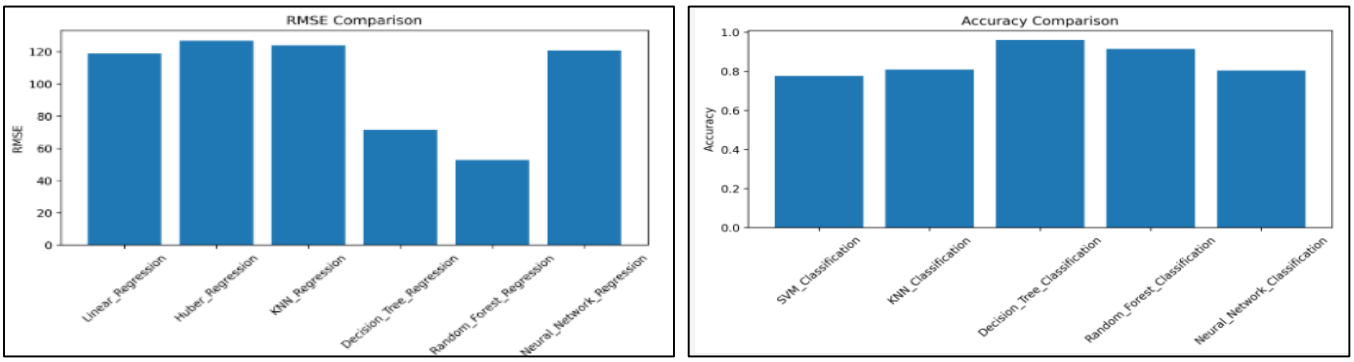


Figure 4: RMSE and Accuracy Comparison



Superior regression performance was demonstrated by tree-based models, as indicated by R^2 scores. The highest value of 0.88 was achieved by Random Forest Regression, followed by Decision Tree Regression at 0.78, while values below 0.4 were recorded by other models, indicating weaker performance. This trend was further confirmed through error analysis. The lowest RMSE of 53 was obtained by Random Forest Regression, with Decision Tree Regression recording 72, whereas RMSE values exceeding 115 were observed for other models, reflecting reduced predictive capability. Mean Absolute Error analysis identified Decision Tree Regression as the lowest-error model with a value of 17, closely followed by Random Forest Regression at 21. Substantially higher error values, ranging from 58 to 77, were observed for the remaining models.

In classification tasks, the highest accuracy of 0.96 was achieved by Decision Tree Classification, followed by Random Forest Classification at 0.91, while accuracy values of approximately 0.81 were attained by other classifiers. These results emphasize the reliability of tree-based classification models for AQI categorization.

4.2 Confusion Matrix of Models

Figure 5: Confusion Matrix (KNN & SVM)

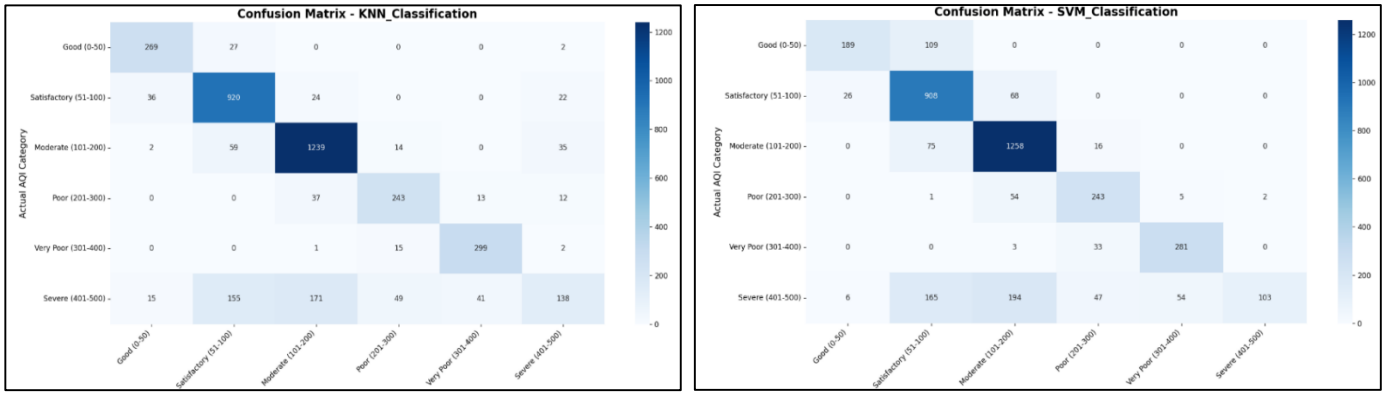
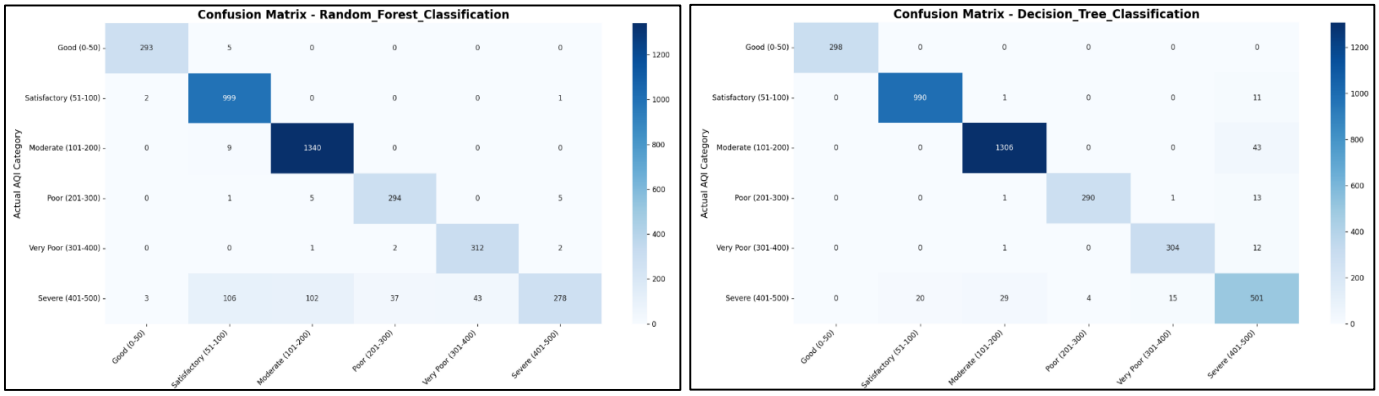


Figure 6: Confusion Matrix (DT & RF)



A final review of the model performance summary confirms the effective application of the machine learning approach for air quality prediction. Random Forest achieved the highest observed accuracy of 99.79%, Followed closely by Decision Tree at 99.77%. The remaining models also showed highly competitive performance, with results ranging from 97.47% to 97.92%. These findings indicate that tree-based ensemble methods outperform traditional approaches and highlight the effectiveness of the adopted preprocessing and feature engineering pipeline.

Model	Algorithm Input/ Accuracy Parameter	Output/Accuracy Result
Multiple linear regression	**Input variables:** SO2, NO2, O3, PM 2.5 & RSPM	**Output variable:** AQI Multilinear model
Artificial Neural Network (ANN)	Confusion Matrix	**Accuracy:** • After Epoch 1 – **92.57** • After Epoch 2 – **96.66** • After Epoch 3 – **97.08** • After Epoch 4 – **97.37** • After Epoch 5 – **97.47**
K-Nearest Neighbour (KNN)	K (Neighbors) = **5 Minkowski distance** Confusion Matrix	Accuracy = **97.92%**
Support Vector Machine (SVM)	Confusion Matrix	Accuracy = **97.91%**
Decision Tree	Confusion Matrix	Accuracy = **99.77%**
Random Forest	Confusion Matrix	Accuracy = **99.79%**

Table 1: Accuracy of Different Models

5. Conclusions

Reliable air quality predictions over both short-term and long-term horizons are achieved through the effective integration of machine learning techniques, ensemble models and ARIMA-based time-series forecasting. Consistent and strong predictive performance is observed across both regression and classification tasks, with tree-based models, particularly Random Forest, demonstrating superior accuracy and robustness. These findings confirm the capability of ensemble learning approaches to model complex, non-linear relationships inherent in air pollution data.

The proposed framework supports detailed pollutant-level analysis and provides interpretable insights into spatial and temporal air quality variations across multiple urban regions. By enabling data-driven assessment and forecasting of air quality trends, this research offers a scalable and reliable solution to support informed decision-making for policymakers, public health authorities and urban planners.

6. Future Work

Future work may focus on the integration of deep learning architectures to enhance long-term predictive accuracy and the incorporation of real-time environmental sensor data to enable continuous monitoring and automated air quality alert systems.

References

1. Moolchand Sharma, Samyak Jain, Sidhant Mittal, Dr. Tariq Hussain Sheikh. (2022) “Forecasting and Prediction of Air Pollutants Using Machine Learning Techniques” IOP Conf. Series: Materials Science and Engineering (ICCRDA 2020)
2. Kalyan Chatterjee, Samla Suraj Kumar, Ramagiri Praveen Kumar, Anjan Bandyopadhyay, Sujata Swain. (2021) “Future Air Quality Prediction Using Long Short-Term Memory” Institute of Electrical and Electronics Engineers (IEEE), Doi: 10.1109/ACCESS.2024.3441109
3. K. Kumar, B. P. Pande (2022) “Air pollution prediction with machine learning” International Journal of Environmental Science and Technology, Doi: https://doi.org/10.1007/s13762-022-04241-5
4. Gokulan Ravindiran, Christian Sonne, Gasim Hayder, Avinash Alagumalai, Karthick Kanagarathinam. (2023) “Air quality prediction (AQI) by machine learning models” 6th International Conference on Engineering, Doi: https://doi.org/10.1016/j.chemosphere.2023.139518.
5. Debopriya Manna, Rohan Mondal, Arpan Sanyal, Ahana Biswas, Hritam Roy and Subhajyoti Barman. (2024) “Air Pollution Prediction System in Smart Cities Using Data Mining Technique” International Research Journal of Engineering and Technology (IRJET)
6. N. K. R, S. Bhumika, S. R, and V. R. (2020) "Air Quality Index Prediction using LSTM" International Research Journal of Engineering and Technology (IRJET), vol. 7, no. 6.
7. N. S. Gupta et al. (2023) "Prediction of air quality index using machine learning techniques: a comparative analysis" Journal of Environmental and Public Health, vol. 2023, pp. 126.
8. M. Imam, S. Adam, S. Dev, and N. Nesa. (2024) "Air Quality Monitoring Using Statistical Learning Models for Sustainable Environment," Intelligent Systems with Applications, vol. 200333.

9. Malhi, G.S., Kaur, M., Kaushik, P. (2021) “Impact of climate change on agriculture and its mitigation strategies” International Conference Sustain, Doi: https://doi.org/ 10.3390/SU13031318.
10. Li, H., Fan, H., Mao, F. (2016) “A visualization approach to air pollution data exploration—a case study of air quality index” Int. Conf. in China, Doi: https://doi.org/10.3390/ATMOS7030035.