ISSN 2349-4506
Impact Factor: 3.799

Global Journal of Engineering Science and Research Management

AN ENHANCED MACHINE LEARNING FRAMEWORK FOR AIR QUALITY PREDICTION

Pawan Yadav*[1], Arpit Solanki*[2]
*[1]Student, Dr. A.P.J. Abdul Kalam University, Indore, M.P., India
*[2]Assistant Professor, Dr. A.P.J. Abdul Kalam University, Indore, M.P., India

KEYWORDS: Air Quality Index, air pollution, machine learning, Random Forest Regression, feature engineering, data preprocessing, regression

ABSTRACT

Reliable prediction of the Air Quality Index (AQI) can support public awareness, environmental monitoring, and timely pollution-management decisions. This paper presents an Enhanced Machine Learning Model (ELMM), a Random Forest Regression-based workflow that combines data preparation with pollutant-derived feature engineering for AQI estimation. The study uses a publicly available Kaggle air-quality dataset containing pollutant measurements and meteorological variables. The workflow includes duplicate and invalid-record handling, missing-value imputation, normalization, feature selection, and derived features representing particulate-matter ratios, combined particulate matter, particulate interaction, and gas-pollutant interaction. The dataset is divided into training and testing subsets. ELMM is compared with Decision Tree Regression (DTR) and standalone Random Forest Regression (RFR) using tolerance-based accuracy, Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE). In the reported experiment, the integrated ELMM pipeline performed best across the evaluated metrics. These findings describe performance on the evaluated data split and should not be treated as proof of generalization to other locations or time periods. The contribution is the integrated prediction workflow, rather than a newly invented Random Forest algorithm.

INTRODUCTION

Air pollution has emerged as one of the most pressing environmental challenges of the twenty-first century, posing serious threats to human health, ecosystems, and long-term development. Rapid industrialisation, urbanisation, population expansion, and increased usage of motor vehicles have all led to a steady rise in air pollution concentrations. Nitrogen dioxide (NO_2), sulphur dioxide (SO_2), carbon monoxide (CO), ozone (O_3), and fine and coarse particulate matter ($PM_{2.1}$ and PM_{1^1}) are significant pollutants; exposure to them is linked to respiratory and cardiovascular problems as well as ecological degradation [1], [2].

By combining the concentrations of many pollutants into a single numerical value, the Air Quality Index (AQI) is a widely used statistic that represents the general quality of ambient air. AQI enables governments, environmental agencies, researchers, and the general public to assess pollution levels and take appropriate preventative measures. Accurate forecasting of AQI is critical for delivering early warnings, supporting policy choices, and raising public awareness about air pollution [3].

Conventional air quality monitoring systems rely mainly on fixed monitoring stations equipped with specialist sensors. Although these systems produce accurate readings, they have high installation and maintenance costs and are generally limited in geographical coverage. Furthermore, typical statistical techniques may fail to accurately simulate the complex and nonlinear interactions that exist between pollutant concentrations, meteorological conditions, and seasonal fluctuations [4], [5].

Recent breakthroughs in machine learning (ML) have opened up new possibilities for constructing intelligent air quality forecast systems. Machine learning algorithms may detect hidden patterns in previous environmental data and make good forecasts without having explicitly specified mathematical correlations. Their capacity to analyse vast datasets and identify nonlinear connections makes them ideal for predicting air quality and assisting in environmental decision-making [5], [6].

The effectiveness of machine learning models depends not only on the learning algorithm but also on the preprocessing stage and the quality of the input data. Environmental datasets commonly contain missing values, noise, duplicate records, and inconsistencies in feature scales, which can lower prediction accuracy. Therefore, in order to improve model performance and produce reliable predictions, thorough data cleaning, missing value management, normalization, and feature selection are essential steps [7], [8].

ISSN 2349-4506
Impact Factor: 3.799

Global Journal of Engineering Science and Research Management

This research proposes the Enhanced Machine Learning Model (ELMM) for Air Quality Index (AQI) prediction using a publicly available Kaggle Air Quality dataset. ELMM integrates data cleaning, missing-value handling, four pollutant-derived features, normalization, and feature selection with Random Forest Regression. Its performance is compared with Decision Tree Regression (DTR) and standalone Random Forest Regression (RFR) using tolerance-based accuracy, Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE). The proposed ELMM demonstrates better overall performance than both DTR and standalone RFR across all three evaluation metrics in the reported experiment.

RELATED WORK

Various researchers have investigated the application of machine learning algorithms for predicting air quality through various datasets, geographical locations, input variables, and prediction aims. Existing research shows that machine learning can provide accurate predictions; nevertheless, variances in datasets, preprocessing methods, algorithms, and assessment criteria make individual approaches context-dependent.

Table 2.1 presents a comparative analysis of existing studies. The comparison is based on the prediction technique used, the primary contribution of the study, and its major limitation. The analysis indicates that although significant progress has been made in air quality forecast or prediction, several challenges still exist, including data quality issues, high computational complexity, and the need for improved prediction accuracy using simple and efficient models.

Ref.	Authors	Technique Used	Major Contribution	Limitation
[1]	Shahbazi et al. (2024)	Machine Learning	Improved air quality forecasting using ML techniques	Further improvement in prediction accuracy is required
[3]	Cabaneros et al. (2019)	Artificial Neural Network	Comprehensive review of ANN models for air pollution prediction	High computational requirement
[4]	Feng et al. (2015)	ANN + Wavelet Transformation	Improved $PM_{2.5}$ prediction accuracy	Complex implementation
[5]	Freeman et al. (2018)	Deep Learning	Time-series air quality forecasting	Longer training time
[6]	Sharma et al. (2005)	Artificial Neural Network	Vehicular pollution modelling	Review-based study
[8]	Narkhede et al. (2023)	Hybrid MIA-LSTM	Improved $PM_{2.5}$ forecasting	High computational complexity
[9]	Alsaber et al. (2021)	Random Forest	Missing value imputation	Focused only on preprocessing
[10]	Zainuri et al. (2015)	Data Imputation	Comparison of missing value handling methods	Prediction model not developed
[11]	Garg & Jindal (2021)	Time-Series Forecasting	Comparison of forecasting models	Dataset-dependent performance
[12]	Kumar & Pande (2023)	Machine Learning	Air pollution prediction for Indian cities	Scope limited to selected datasets
[13]	Maduri et al. (2023)	Machine Learning	Air Pollution Index prediction	Limited comparative analysis
[14]	Sachetti & Mota (2025)	LSTM Review	Review of LSTM-based prediction models	Review study only
[15]	Drewil & Al-Bahadili (2022)	LSTM + Metaheuristics	Improved prediction using optimization	Increased computational cost
[16]	Rybarczyk & Zalakeviciute (2017)	Regression Models	Prediction using affordable environmental data	Moderate prediction accuracy
[17]	Gupta et al. (2023)	Comparative Machine Learning	Comparative analysis of AQI prediction techniques	No preprocessing framework proposed
[19]	Stamenković et al. (2017)	Artificial Neural Network	NO_x emission prediction	Specific to emission prediction

Table 2.1: Comparative Analysis of Existing Studies

ISSN 2349-4506
Impact Factor: 3.799

Global Journal of Engineering Science and Research Management

RESEARCH FRAMEWORK

The research framework presents the systematic approach used to develop and evaluate the proposed Enhanced Machine Learning Model (ELMM) for AQI prediction using a publicly available Kaggle Air Quality dataset. The process includes dataset collection, data inspection, cleaning, missing-value handling, normalization, and feature selection. The preprocessed data is divided into training and testing sets in a 70:30 ratio. ELMM combines feature engineering with Random Forest Regression and is compared with standalone Random Forest Regression (RFR) and Decision Tree Regression (DTR). Model performance is evaluated using tolerance-based accuracy, MAE, and RMSE. The complete research framework is illustrated in Figure 3.1.

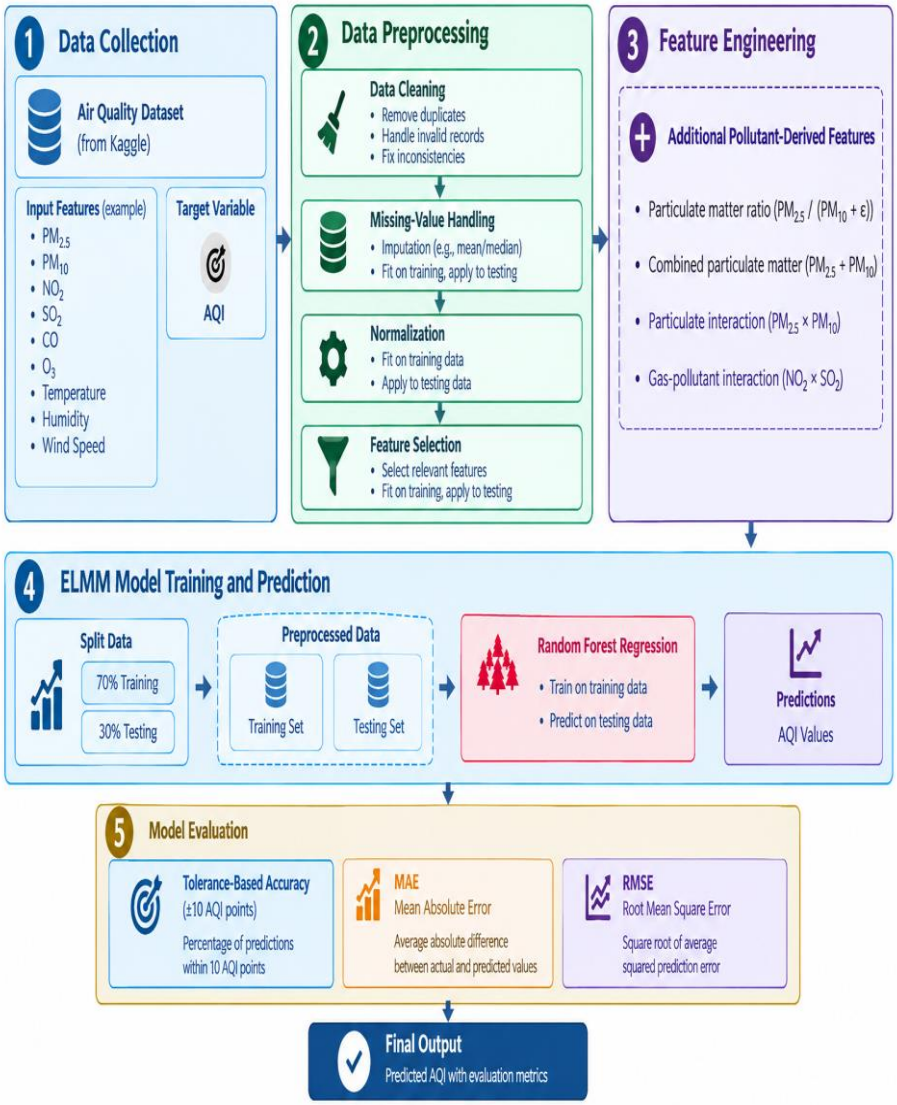

Figure 3.1: Research Framework

RESULTS ANALYSIS

The results indicate that the proposed ELMM outperforms both Decision Tree Regression (DTR) and standalone Random Forest Regression (RFR) across all three evaluation metrics. It achieves higher tolerance-based accuracy and lower MAE and RMSE, demonstrating more accurate AQI predictions with smaller prediction errors. While RFR performs better than DTR, ELMM delivers the strongest overall performance. These improvements may be attributed to its integrated pipeline, which combines data cleaning, missing-value handling, pollutant-derived

feature engineering, normalization, feature selection, and Random Forest Regression. However, the individual contribution of each component requires further investigation.

The graphical comparison of the three models is presented in Figures 4.1, 4.2, and 4.3.
In this experiment, the goal is to expand the number of items in the dataset. The results highlight the comparative performance of our proposed algorithm against the Apriori algorithm.

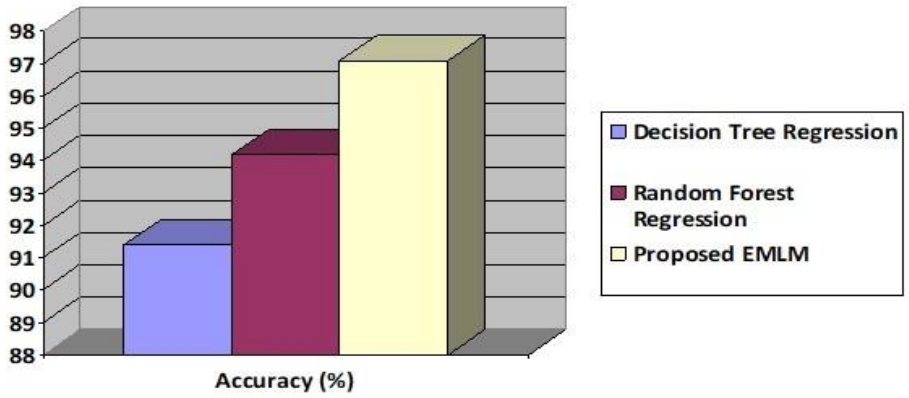

Fig 4.1 Accuracy Comparison

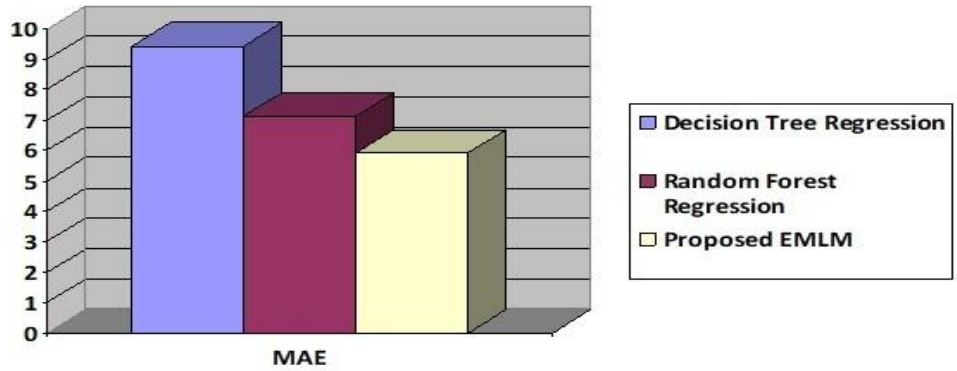

Fig 4.2 MAE Comparison

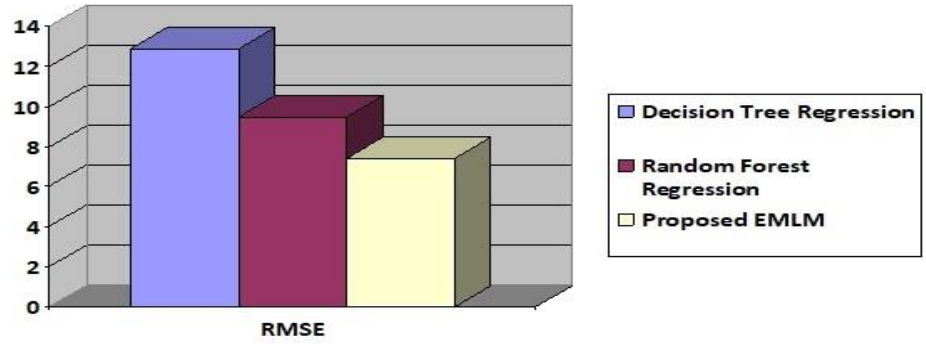

Fig 4.3 RMSE Comparison

CONCLUSION AND FUTURE SCOPE

This research presents the Enhanced Machine Learning Model (ELMM) for Air Quality Index (AQI) prediction using a publicly available Kaggle dataset. The proposed model integrates data cleaning, missing-value handling, pollutant-derived feature engineering, normalization, feature selection, and Random Forest Regression. Experimental results show that ELMM outperforms Decision Tree Regression (DTR) and standalone Random Forest Regression (RFR), achieving higher tolerance-based accuracy and lower prediction errors. The study demonstrates the potential of the proposed approach for air-quality prediction. In the future, the model can be evaluated using diverse datasets, additional environmental factors, and advanced machine learning and deep learning algorithms. Integration with IoT sensors, cloud computing, and web or mobile applications could further support real-time air-quality monitoring and public awareness.

REFERENCES

1. Z. Shahbazi, Z. Shahbazi, and S. Nowaczyk, "Enhancing Air Quality Forecasting Using Machine Learning Techniques," *IEEE Access*, vol. 12, pp. 197290–197299, 2024.
2. I. F. Gheorghe and B. Ion, "The Effects of Air Pollutants on Vegetation and the Role of Vegetation in Reducing Atmospheric Pollution," in *The Impact of Air Pollution on Health, Economy, Environment and Agricultural Sources*, vol. 29, pp. 241–280, 2011, doi:10.5772/17660.
3. S. M. Cabaneros, J. K. Calautit, and B. R. Hughes, "A Review of Artificial Neural Network Models for Ambient Air Pollution Prediction," *Environmental Modelling & Software*, vol. 119, pp. 285–304, 2019, doi:10.1016/j.envsoft.2019.06.014.
4. X. Feng, Q. Li, Y. Zhu, J. Hou, L. Jin, and J. Wang, "Artificial Neural Networks Forecasting of PM2.5 Pollution Using Air Mass Trajectory Based Geographic Model and Wavelet Transformation," *Atmospheric Environment*, vol. 107, pp. 118–128, 2015, doi:10.1016/j.atmosenv.2015.02.030.
5. B. S. Freeman, G. Taylor, B. Gharabaghi, and J. Thé, "Forecasting Air Quality Time Series Using Deep Learning," *Journal of the Air & Waste Management Association*, vol. 68, no. 8, pp. 866–886, 2018, doi:10.1080/10962247.2018.1459956.
6. N. Sharma, K. K. Chaudhry, and C. V. Rao, "Vehicular Pollution Modeling Using Artificial Neural Network Technique: A Review," 2005.
7. H. T. Shahraiyni, S. Sodoudi, A. Kerschbaumer, and U. Cubasch, "New Technique for Ranking of Air Pollution Monitoring Stations in Urban Areas Based upon Spatial Representativity," *Aerosol and Air Quality Research*, vol. 15, no. 2, pp. 743–748, 2015, doi:10.4209/aaqr.2014.12.0317.
8. G. Narkhede, A. Hiwale, B. Tidke, and C. Khadse, "Novel MIA-LSTM Deep Learning Hybrid Model with Data Preprocessing for Forecasting of PM2.5," *Algorithms*, vol. 16, no. 1, Art. 52, 2023, doi:10.3390/a16010052.
9. A. R. Alsaber, J. Pan, and A. Al-Hurban, "Handling Complex Missing Data Using Random Forest Approach for an Air Quality Monitoring Dataset: A Case Study of Kuwait Environmental Data (2012–2018)," *International Journal of Environmental Research and Public Health*, vol. 18, no. 3, Art. 1333, 2021, doi:10.3390/ijerph18031333.
10. N. A. Zainuri, A. A. Jemain, and N. Muda, "A Comparison of Various Imputation Methods for Missing Values in Air Quality Data," *Sains Malaysiana*, vol. 44, no. 3, pp. 449–456, 2015, doi:10.17576/jsm-2015-4403-17.
11. S. Garg and H. Jindal, "Evaluation of Time Series Forecasting Models for Estimation of PM2.5 Levels in Air," in *Proceedings of the 2021 6th International Conference for Convergence in Technology (I2CT)*, pp. 1–8, 2021.
12. K. Kumar and B. P. Pande, "Air Pollution Prediction with Machine Learning: A Case Study of Indian Cities," *International Journal of Environmental Science and Technology*, vol. 20, no. 5, pp. 5333–5348, 2023.
13. P. K. Maduri, P. Dhiman, C. Chaturvedi, and A. Rai, "Air Pollution Index Prediction: A Machine Learning Approach," in *Proceedings of the Second International Conference in Mechanical and Energy Technology*, Singapore, vol. 290, pp. 37–51, 2023.
14. L. L. S. Sachetti and V. F. S. Mota, "Time Series Prediction about Air Quality Using LSTM-Based Models: A Systematic Mapping," 2025.
15. G. I. Drewil and R. J. Al-Bahadili, "Air Pollution Prediction Using LSTM Deep Learning and Metaheuristics Algorithms," *Measurement: Sensors*, vol. 24, Art. 100546, 2022, doi:10.1016/j.measen.2022.100546.

16. Y. Rybarczyk and R. Zalakeviciute, "Regression Models to Predict Air Pollution from Affordable Data Collections," in *Machine Learning – Advanced Techniques and Emerging Applications*, IntechOpen, 2017, doi:10.5772/intechopen.71848.
17. N. S. Gupta, Y. Mohta, K. Heda, R. Armaan, B. Valarmathi, and G. Arulkumaran, "Prediction of Air Quality Index Using Machine Learning Techniques: A Comparative Analysis," *Journal of Environmental and Public Health*, vol. 2023, Art. 4916267, 2023, doi:10.1155/2023/4916267.
18. G. S. Vidya and V. S. Hari, "Gold Price Prediction and Modelling Using Deep Learning Techniques," in *Proceedings of the IEEE Recent Advances in Intelligent Computational Systems (RAICS)*, pp. 28–31, 2020.
19. L. J. Stamenković, D. Z. Antanasijević, M. Đ. Ristić, A. A. Perić-Grujić, and V. V. Pocajt, "Prediction of Nitrogen Oxides Emissions at the National Level Based on Optimized Artificial Neural Network Model," *Air Quality, Atmosphere & Health*, vol. 10, no. 1, pp. 15–23, 2017.
20. G. N. Kouziokas, "The Application of Artificial Intelligence in Public Administration for Forecasting High Crime Risk Transportation Areas in Urban Environment," *Transportation Research Procedia*, vol. 24, pp. 467–473, 2017.
21. B. Mahapatra and M. Irfan, "Asymmetric Impacts of Energy Efficiency on Carbon Emissions: A Comparative Analysis Between Developed and Developing Economies," *Energy*, vol. 227, Art. 120485, 2021.