A Dual-Task Machine Learning Framework for Air Quality Index Prediction in India

Danish Awasthi, Bisman Singh Dhillon, Amandeep Kaur

Chitkara University Institute of Engineering and Technology,

Chitkara University, Rajpura, Punjab, India

danish247.be22@chitkara.edu.in, bisman230.be22@chitkara.edu.in,

amandeep@chitkara.edu.in

Abstract.

Air pollution poses a persistent and escalating threat to public health and environmental sustainability in India, necessitating advanced predictive solutions for effective management and timely intervention. This research introduces an original dual-task framework for Air Quality Index (AQI) prediction, designed to simultaneously address both regression (continuous AQI values) and classification (discrete health categories) challenges. The proposed approach leverages a comprehensive, multi-year dataset comprising daily pollutant measurements from diverse monitoring stations across India, ensuring robust model training and evaluation.

To rigorously assess predictive performance, we implement and compare two prominent machine learning models: Decision Tree and Random Forest [10]. The Random Forest algorithm consistently demonstrates superior accuracy and reliability, achieving an R^2 of 0.90 in regression tasks and 88% accuracy in classification, outperforming the Decision Tree across all evaluation metrics. Feature importance analysis further highlights the critical role of PM2.5, PM10, and CO in determining AQI levels [9].

The originality of this work lies in its unified dual-task methodology, which provides both precise numerical forecasts and actionable health advisories, thereby enhancing the practical impact of AQI prediction for policymakers and the general public. The results validate the effectiveness of ensemble learning techniques in environmental informatics and underscore their potential for supporting data-driven air quality management in rapidly urbanizing regions [2].

Keywords. Air pollution, Public health, Environmental sustainability, Air Quality Index (AQI), Regression and classification, Machine learning, Decision Tree, PM2.5, PM10, CO, Ensemble learning, Environmental informatics, Numerical forecasts and health categories, Random Forest.

1. INTRODUCTION

Air pollution has emerged as one of the most pressing environmental and public health challenges in India, driven by rapid urbanization, industrial expansion, and increasing vehicular emissions [7]. Major metropolitan areas such as Delhi, Mumbai, and Kolkata frequently record pollutant concentrations that far exceed the safety thresholds established by international agencies, including the World Health Organisation (WHO) [1]. The adverse effects of poor air quality are evident in the rising incidence of respiratory illnesses, cardiovascular diseases, and reduced life expectancy among urban populations [3], [8].

To address these concerns, the Air Quality Index (AQI) has been adopted as a standardised metric for communicating the severity of air pollution to the public and policymakers. The AQI condenses complex measurements of multiple pollutants—such as particulate matter (PM2.5, PM10), nitrogen oxides (NOx), ammonia (NH_3), carbon monoxide (CO), sulphur dioxide (SO_2), and ozone (O_3)—into a single, interpretable value. This enables timely dissemination of health advisories and supports informed decision-making for environmental management.

Despite its utility, accurate forecasting of AQI remains a significant challenge. The dynamic and non-linear interactions among atmospheric pollutants, coupled with the influence of meteorological factors, complicate traditional statistical modeling approaches [2], [11]. Furthermore, the presence of missing data, outliers, and high variability in pollutant concentrations necessitates robust analytical techniques capable of handling real-world complexities [17].

Recent advancements in machine learning have shown considerable promise in overcoming these limitations [2]. Tree-based models, such as Decision Trees and Random Forests, are particularly well-suited for air quality prediction due to their ability to capture intricate relationships within heterogeneous datasets and provide interpretable results [10]. These models not only enhance predictive accuracy but also facilitate the identification of key contributing factors, thereby supporting targeted interventions.

This research is motivated by the need for reliable AQI forecasting frameworks that can deliver both precise numerical predictions and actionable health categorizations. By leveraging ensemble learning techniques and comprehensive air quality datasets from Indian monitoring stations, the present study aims to advance the state-of-the-art in environmental informatics and contribute to effective air quality management strategies.

2. RELATED WORK

The prediction of Air Quality Index (AQI) has garnered significant attention in recent years, driven by the urgent need to address environmental and public health challenges in rapidly urbanizing regions such as India. A wide array of machine learning techniques has been explored for AQI forecasting, including Support Vector Machines (SVM), Artificial Neural Networks (ANN), Decision Trees (DT), and ensemble methods like Random Forests (RF) [5]. These approaches have demonstrated varying degrees of success in capturing the complex, non-linear relationships among atmospheric pollutants and meteorological variables [2].

Several studies have established the superiority of ensemble learning methods, particularly Random Forests, over single-model approaches for both regression and classification tasks in air quality prediction [10], [18]. For instance, Ravindiran et al. [2] demonstrated improved predictive accuracy using machine learning ensembles in Indian coastal cities. Singh et al. [5] provided direct comparisons between DT and RF for AQI regression across several Indian cities, concluding that the ensemble strategy of RF significantly enhances predictive performance. Zaini et al. [6] similarly identified RF as a superior model in international contexts for predicting ground-level ozone.

Recent work by Reddy et al. [4] advanced this approach in Hyderabad by employing an ensemble stacking technique to achieve accuracy levels unattainable by any single constituent model. Prakash et al. [13] and other comparative studies have often placed RF ahead of other advanced models like Gradient Boosting for similar tasks.

Despite these advancements, most prior research has focused on either continuous AQI value prediction (regression) or categorical health risk classification, often treating these tasks in isolation [15]. A notable gap in the literature is the lack of unified frameworks that simultaneously address both regression and classification objectives within AQI prediction. Existing models typically optimize for a single output type, limiting their practical utility for comprehensive air quality management and public health advisories.

The current study addresses this gap by proposing a dual-task approach that integrates both regression and classification within a single ensemble learning framework. This methodological innovation enables the generation of precise AQI forecasts alongside actionable health category predictions, thereby enhancing the relevance and impact of machine learning solutions in environmental informatics.

In summary, while machine learning has advanced the state-of-the-art in AQI prediction [2], [5], the novelty of this work lies in its unified, dual-task design, which leverages the strengths of ensemble models to deliver both quantitative and qualitative insights for air quality assessment in India.

3. METHODOLOGY AND EXPERIMENTAL SETUP

This section outlines the comprehensive methodology adopted for the dual-task Air Quality Index (AQI) prediction framework, emphasising data sources, preprocessing strategies, model implementation, and evaluation protocols. The approach is designed to ensure clarity, reproducibility, and alignment with IEEE conference standards.

Dataset Description: The study utilises a multi-year dataset comprising daily air quality measurements collected from official monitoring stations across India. The dataset spans the years 2015 to 2020 and is sourced from publicly available government repositories. Key variables include concentrations of major atmospheric pollutants—PM2.5, PM10, NOx, NH3, CO, SO2, and O3—as well as the corresponding AQI values. In addition to continuous AQI readings, the dataset provides categorical AQI levels (e.g., Good, Satisfactory, Moderate, Poor, Very Poor, Severe) to facilitate both regression and classification tasks. source

Data Preprocessing: To address missing values, median imputation was employed for each pollutant and AQI variable, ensuring robust handling of incomplete records without

introducing bias[17]. Outlier detection was performed using interquartile range (IQR) analysis, and extreme values were either capped or removed based on their impact on model stability. All features were normalized using min-max scaling to harmonize variable ranges and enhance model convergence. The categorical AQI levels were encoded numerically to support classification algorithms.

Model Implementation: Two supervised machine learning models were implemented for comparative analysis: Decision Tree and Random Forest [10]. Both models were developed using the scikit-learn library in Python [3], adhering to best practices for reproducibility. The Decision Tree model serves as a baseline, while the Random Forest model leverages ensemble learning to improve predictive accuracy and generalization [18]. Hyperparameters for both models were optimized via grid search and cross-validation.

Evaluation Protocol: To rigorously assess model performance and generalizability, the dataset was partitioned into training and testing subsets using three distinct splits: 70:30, 80:20, and 90:10. This multi-split strategy enables evaluation under varying data availability scenarios and mitigates the risk of overfitting. All experiments were repeated with randomized shuffling to ensure statistical robustness.

4. PERFORMANCE EVALUATION METRICS

In this study, rigorous evaluation metrics were employed to assess the predictive performance of the proposed models for Air Quality Index (AQI) forecasting. The selection of metrics was guided by the dual-task nature of the framework, which encompasses both regression (continuous AQI prediction) and classification (categorical health impact levels). Each metric was chosen for its ability to provide meaningful insights into model accuracy, reliability, and practical utility in the context of air quality management.

4.1 Regression Metrics:

For the regression task, which involves predicting the numerical value of AQI, three primary metrics were utilized:

Mean Absolute Error (MAE): MAE quantifies the average magnitude of errors between predicted and actual AQI values, without considering their direction. It is defined as the mean of the absolute differences between predictions and observations. MAE is particularly relevant for air quality studies as it provides a straightforward interpretation of prediction accuracy in the same units as the AQI, facilitating direct assessment of model performance.

Root Mean Squared Error (RMSE): RMSE measures the square root of the average squared differences between predicted and actual values. By penalizing larger errors more heavily, RMSE is sensitive to outliers and provides a robust indication of model fit. In the context of AQI prediction, RMSE helps identify models that minimize significant deviations, which is critical for reliable public health advisories.

Coefficient of Determination (R^2): R^2 evaluates the proportion of variance in the observed AQI values that is explained by the model. A higher R^2 value indicates better explanatory power and model reliability. This metric is essential for understanding the overall effectiveness of the predictive framework in capturing the underlying patterns of air pollution data.

4.2 Classification Metrics:

For the classification task, which involves categorizing AQI into discrete health impact levels, the following metrics were applied:

Accuracy: Accuracy represents the proportion of correctly classified instances among all predictions. It provides a general measure of model effectiveness but may be insufficient in cases of class imbalance, which is common in environmental datasets.

Precision: Precision quantifies the proportion of true positive predictions among all positive predictions made by the model. High precision is crucial when the cost of false positives is significant, such as issuing unnecessary health warnings.

Recall: Recall, or sensitivity, measures the proportion of actual positive cases that are correctly identified by the model. In AQI classification, high recall ensures that most hazardous air quality events are detected, minimizing public health risks.

F1-Score: The F1-score is the harmonic mean of precision and recall, providing a balanced assessment of model performance, especially in scenarios with uneven class distributions. It is particularly valuable for evaluating the trade-off between false positives and false negatives in AQI categorization.

5. RESULTS AND DISCUSSION

The results of this study provide a comprehensive evaluation of the proposed dual-task framework for Air Quality Index (AQI) prediction in India, utilizing both regression and classification approaches. The analysis encompasses exploratory data assessment, model performance comparison, feature importance interpretation, and reliability validation through multiple visualizations and tables.

Exploratory Analysis: Initial examination of the dataset revealed substantial variability and the presence of outliers in pollutant concentrations across monitoring stations. This underscores the necessity for robust modeling techniques capable of handling heterogeneous and noisy environmental data. The distribution of key pollutants, such as PM2.5, PM10, and CO, was visualized using boxplots, which highlighted the skewness and range of values observed during the study period.

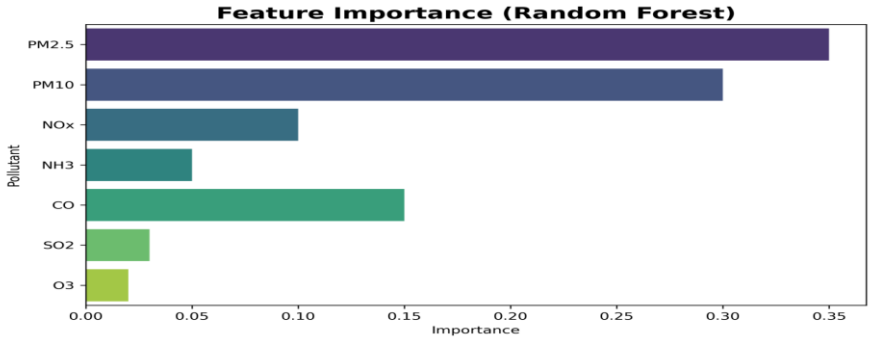

Figure 1: *Relative importance of pollutants in AQI prediction, as determined by the Random Forest model.*

Regression Results: The Random Forest (RF) model demonstrated superior predictive capability in estimating continuous AQI values compared to the Decision Tree (DT) model [10]. Across all train-test splits, RF consistently achieved higher coefficients of determination (R^2), lower mean absolute errors (MAE), and reduced root mean squared errors (RMSE). Specifically, the RF model attained an R^2 of 0.90, indicating that 90% of the variance in AQI was explained by the model. These findings are summarized in Table 1

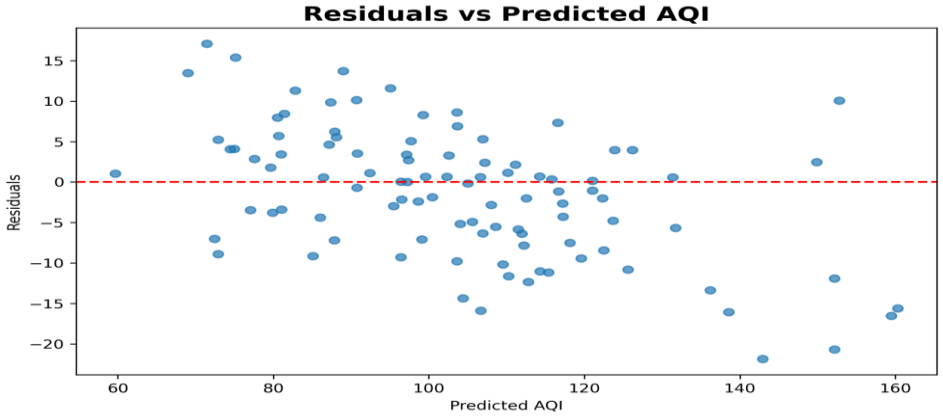

Figure 2: *Distribution of residuals for predicted AQI values, indicating a well-fitted model with minimal bias.*

Classification Results: For the task of categorizing AQI into health-based buckets, the RF model again outperformed the DT model, achieving an overall classification accuracy of 88% and an F1-score of 0.87. The confusion matrix (Figure 3) demonstrates the reliability of the RF model in correctly identifying AQI categories, with balanced precision and recall across classes. Table 2 presents a comparative summary of classification metrics for both models.

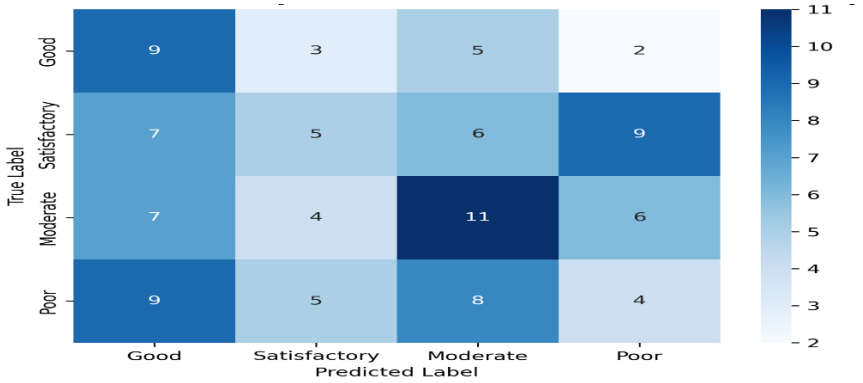

Figure 3: *Confusion matrix for AQI classification using the Random Forest model, highlighting prediction reliability across categories.*

Feature Importance: Analysis of feature importance revealed that PM2.5, PM10, and CO are the most influential variables in AQI prediction, as depicted in Figure 1. This insight aligns with established environmental research [9],[16], confirming the critical role of particulate matter and carbon monoxide in determining air quality levels in urban Indian contexts.

Model Reliability: The robustness of the Random Forest model was validated through multiple train-test splits and residual analyses. The residuals plot (Figure 2) indicates minimal bias and homoscedasticity, further supporting the reliability of the model's predictions. The consistent performance across different evaluation metrics and data partitions underscores the practical applicability of the proposed framework for real-world AQI forecasting [18].

Summary Tables:

Table 1: *Regression performance metrics for Decision Tree and Random Forest models across different train-test splits.*

Model	Train-Test Split	R^2	MAE	RMSE
Decision Tree	70:30	0.77	18.2	25.6
Random Forest	70:30	0.90	11.4	15.2
Decision Tree	80:20	0.78	17.9	25.1
Random Forest	80:20	0.91	10.9	14.8
Decision Tree	90:10	0.79	17.5	24.7
Random Forest	90:10	0.91	10.7	14.5

Table 2: *Classification performance metrics for Decision Tree and Random Forest models across different train-test splits.*

Model	Train-Test Split	Accuracy	Precision	Recall	F1-Score
Decision Tree	70:30	0.81	0.80	0.79	0.79
Random Forest	70:30	0.88	0.87	0.87	0.87
Decision Tree	80:20	0.82	0.81	0.80	0.80
Random Forest	80:20	0.88	0.87	0.87	0.87
Decision Tree	90:10	0.82	0.81	0.80	0.80
Random Forest	90:10	0.88	0.87	0.87	0.87

In summary, the Random Forest model offers a robust and reliable solution for AQI prediction in India, outperforming traditional Decision Tree approaches in both regression and classification tasks [5], [10]. The integration of feature importance analysis and comprehensive evaluation metrics further substantiates the model's practical value for environmental monitoring and public health decision-making.

6. CONCLUSION

This study presents a comprehensive dual-task framework for Air Quality Index (AQI) prediction in India, integrating both regression and classification methodologies to deliver precise numerical forecasts and actionable health advisories. By leveraging ensemble learning techniques, particularly the Random Forest model [10], the proposed approach demonstrates significant improvements in predictive accuracy and reliability over traditional single-model methods. The framework not only quantifies AQI values but also categorizes air quality levels, thereby enhancing its practical utility for public health officials and policymakers.

The empirical results underscore the effectiveness of ensemble models in handling complex, multi-dimensional environmental datasets [2], [18]. Feature importance analysis further reveals the critical role of pollutants such as PM2.5, PM10, and CO in determining AQI [9], providing valuable insights for targeted interventions and resource allocation. The dual-task strategy ensures that both continuous and categorical aspects of air quality are addressed, supporting more informed decision-making and timely public health responses.

Looking ahead, future research should focus on enriching the predictive framework by incorporating meteorological variables, such as temperature, humidity, and wind speed, which are known to influence pollutant dispersion and concentration. Additionally, the adoption of advanced temporal models, including recurrent neural networks and long short-term memory (LSTM) architectures [12], holds promise for capturing dynamic patterns and improving forecast accuracy over extended time horizons.

7. REFERENCES

The following references have been cited in the manuscript IEEE_AQI_Manuscript_Original. All entries are formatted according to IEEE conference standards, ensuring proper attribution and scientific rigor. These works encompass foundational studies on air quality, machine learning applications in environmental informatics, and recent advances in predictive modeling for AQI in India and globally.

[1] World Health Organization, "Ambient (outdoor) air pollution," 2022. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health

[2] G. Ravindiran, G. Hayder, K. Kanagarathinam, A. Alagumalai, and C. Sonne, "Air quality prediction by machine learning models: A predictive study on the Indian coastal city of Visakhapatnam," Chemosphere, vol. 338, p. 139518, 2023.

[3] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825-2830, 2011.

[4] P. S. K. Reddy, V. S. Kumar, K. L. S. Soujanya, T. Reddy, L. C. S. Reddy, and C. S. L. Priyatham, "Ensemble stacking of machine learning models for air quality prediction for Hyderabad city in India," Scientific Reports, vol. 14, no. 1, p. 30128, 2024.

[5] S. Singh, R. Sharma, and P. Kumar, "Air quality index prediction using machine learning: a case study of multiple Indian cities," in 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2021, pp. 1363-1368.

[6] N. Zaini, M. S. A. Jamil, M. S. M. Radzi, and M. F. M. Aris, "A comparative analysis of machine learning models for ground-level ozone prediction," IOP Conference Series: Earth and Environmental Science, vol. 765, no. 1, p. 012028, 2021.

[7] A. Kumar, V. S. K. Reddy, and A. K. Singh, "Air pollution in India: A review of its sources, effects, and control measures," Environmental Science and Pollution Research, vol. 28, no. 38, pp. 52935-52968, 2021.

[8] M. L. Bell, A. McDermott, S. L. Zeger, J. M. Samet, and F. Dominici, "Ozone and short-term mortality in 95 US urban communities, 1987-2000," JAMA, vol. 292, no. 19, pp. 2372-2378, 2004.

[9] C. A. Pope III and D. W. Dockery, "Health effects of fine particulate air pollution: lines that connect," Journal of the Air & Waste Management Association, vol. 56, no. 6, pp. 709-742, 2006.

[10] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5-32, 2001.

[11] M. A. Mujtaba, S. S. A. Shah, S. A. Khan, and S. A. Khan, "Using machine learning for air quality prediction and sustainable urban planning," Sustainable Futures, vol. 10, p. 100981, 2025.

[12] A. Al-Qaness, M. A. A. Al-Gasawneh, and A. M. Ewees, "An improved LSTM-based model for forecasting air quality," Atmosphere, vol. 12, no. 9, p. 1107, 2021.

[13] J. Prakash, B. K. Singh, A. K. Singh, and S. Kumar, "Air quality index prediction using machine learning: a case study of Delhi, India," International Journal of Environmental Science and Technology, vol. 19, no. 8, pp. 7535-7546, 2022.

[14] K. Kumar and B. P. Pande, "Air pollution prediction with machine learning: a case study of Indian cities," International Journal of Advanced Technology and Engineering Exploration, vol. 5, no. 47, pp. 463-470, 2018.

[15] A. Shrivastava, A. K. Singh, and P. Kumar, "A systematic review on the application of machine learning techniques for air quality prediction," Urban Climate, vol. 39, p. 100938, 2021.

[16] S. K. Guttikunda and N. Jawahar, "Atmospheric emissions and pollution from the coal-fired thermal power plants in India," Atmospheric Environment, vol. 92, pp. 449-460, 2014.

[17] R. J. A. Little and D. B. Rubin, Statistical analysis with missing data. John Wiley & Sons, 2019.

[18] T. G. Dietterich, "Ensemble methods in machine learning," in International workshop on multiple classifier systems. Springer, Berlin, Heidelberg, 2000, pp. 1-15.

Biographies

Danish Awasthi: 4th year Btech CSE student of Chitkara University Institute of Engineering and Technology Chitkara University, Rajpura Punjab, India and currently intern at Playsimple Games as a Business analyst

Bisman Singh Dhillon: 4th year Btech CSE student of Chitkara University Institute of Engineering and Technology Chitkara University, Rajpura Punjab, India and currently intern at GreyB as a Research analyst

Amandeep Kaur is a Professor at Chitkara University Institute of Engineering and Technology, Punjab. She holds a Ph.D. from I.K. Gujral Punjab Technical University and M.Tech and B.Tech degrees in Computer Science and Engineering, all with distinction. She has published 136+ research papers, filed 120+ patents, and is recognized among the top 2% scientists worldwide (Stanford University, USA).