

Comparative Evaluation of Machine and Deep Learning Models for Air Quality Index Prediction in Jaipur, India

Parth Banethiwal¹, and Ruchi Sharma^{2*}

¹Student, Environmental Science and Engineering Department, Indian Institute of Technology (Indian School of Mines), Dhanbad, India

²Assistant Professor, Department of Civil Engineering, Malaviya National Institute of Technology, Jaipur, Pin 302017, India

Abstract. Air pollution is an important environmental and public health challenge, therefore, accurate Air Quality Index (AQI) forecasting is important for timely mitigation. This study predicts AQI in Jaipur, India using seven machine learning and deep learning-based models i.e., Multiple Linear Regression (MLR), Random Forest (RF), Support Vector Regression (SVR), XGBoost, Adaboost, Artificial Neural Network (ANN), and Convolutional Neural Network (CNN). For this purpose, two years of hourly data from three monitoring sites were used, with preprocessing to address missing values and outliers. Key pollutant and meteorological variables were selected using Pearson's correlation coefficient values. Models were evaluated under three scenarios: pollutant parameters only (Case 1), meteorological parameters only (Case 2), and a combined dataset (Case 3). Performance was assessed using indices such as R^2 and RMSE. Case 3 consistently produced the most accurate predictions, with Site 2 reflecting the best overall results. Among all models, XGBoost outperformed achieving R^2 values of 0.77–0.95 and RMSE values of 16.96–20.98 across the three sites. The study demonstrates that XGBoost is a reliable approach for AQI forecasting and provides useful insights for air quality management and policymaking in rapidly urbanizing cities like Jaipur.

Keywords: Air Quality Index (AQI) Prediction, Machine Learning Models, XGBoost Algorithm

1 Introduction

Air pollution is described as the change in ambient or indoor air quality through physical, chemical, or biological contaminants originating from human activities or natural events. Sources of air pollution can be household combustion appliances, vehicular emissions, industrial operations, and forest fires. Major pollutants that adversely affect human health include particulate matter (PM), sulfur oxides (SO_x), nitrogen oxides (NO_x), carbon

* Corresponding author: ruchi.ce@mnit.ac.in

Orchid ID: Dr. Ruchi Sharma: <https://orcid.org/0000-0002-8945-2788>

monoxide (CO), ozone (O₃), ammonia (NH₃), and lead (Pb) [1]. Among these pollutants, PM, SO₂, NO₂, and CO are primary pollutants, which are emitted directly from sources such as vehicles, industries, and dust. Conversely, secondary pollutants including ozone, smog, peroxyacetyl nitrate, and photochemical oxidants, which are formed via chemical reactions among primary pollutants in the atmosphere and can be even more harmful [2].

Air pollution causes significant human health impacts. A report by World Health Organization (WHO) reported that breathing polluted leads to approximately 4.2 million premature deaths annually from cardiovascular diseases, respiratory illnesses, lung cancer, and stroke [3]. In India, rapid industrialization and urbanization have intensified the problem, leading to over 1 million premature deaths and over 31 million disability cases [4]. Projections indicate a 24% rise in PM_{2.5} levels and a corresponding increase in premature mortalities by 2050 as compared to 2015 [5]. Thus, effective air quality management (AQM) is needed which involves three core components: (i) identifying and quantifying emission sources, (ii) implementing reduction measures through regulatory and non-regulatory actions, and (iii) monitoring and evaluating ambient air quality. To facilitate such monitoring, the Air Quality Index (AQI) as developed by Central Pollution Control Board (CPCB) and United States Environmental Protection Agency (USEPA) serves as a standardized measure of air pollution levels. AQI integrates multiple pollutant concentrations such as PM₁₀, PM_{2.5}, NO₂, SO₂, O₃, NH₃, CO, and Pb into a single, interpretable indicator where values below 50 show good air quality, while those exceeding 500 indicate severe pollution [6]. AQI has thus become a vital tool for assessing air pollution impacts and guiding policy interventions.

Machine learning (ML), a branch of artificial intelligence, has shown great potential in modeling complex, non-linear environmental processes in recent years. ML algorithms learn patterns from large datasets and make correct predictions without explicit programming [7]. Their ability to handle high-dimensional, unstructured, and incomplete data makes them highly effective for air quality prediction. Previous studies have demonstrated the application of ML as well as deep learning (DL)-based models for both short- and long-term AQI forecasting [8]. Unlike traditional statistical or deterministic models, ML-based approaches can capture intricate interactions among meteorological and pollutant parameters, providing more realistic results. Previous studies in literature have applied ML techniques to AQI prediction worldwide. For example, Liang *et al.* [9] used AdaBoost and stacking ensembles for Taiwan and found them most effective, and Van *et al.* [10] used XGBoost for Indian cities, achieving $R^2 = 0.92$ and RMSE = 29.7. Similarly, Maltare & Vahora [11] compared SVM, SARIMA, and LSTM for Ahmedabad, with SVM achieving 95% accuracy after comprehensive data preprocessing.

In view of the above, this study predicts AQI for Jaipur using pollutant and meteorological parameters through seven ML and DL models i.e., Multiple Linear Regression (MLR), Random Forest (RF), Support Vector Regression (SVR), eXtreme Gradient Boosting (XGBoost), Adaptive Boosting (AdaBoost), Artificial Neural Network (ANN), and Convolutional Neural Network (CNN). For this purpose, hourly data from three Rajasthan State Pollution Control Board (RSPCB) monitoring stations for two years (2018–2019) were analyzed. After data preprocessing (handling missing data and removing outliers), AQI was computed and input features were selected using Pearson's correlation coefficient. Three modeling scenarios were analyzed: Case 1 with pollutant parameters only, Case 2 with meteorological parameters only, and Case 3 with combined parameters. Each model was trained, validated, and tested, with performance assessed using metrics R^2 and RMSE. The results provide useful insights for policy development, urban air quality management, and public health planning, highlighting the effectiveness of data-driven ML approaches for environmental decision-making in rapidly growing urban cities.

2 Material and methods

This Section 2 describes the methodology used in this study to predict AQI for Jaipur city using various ML and DL-based models. The steps included site selection and data acquisition, data preprocessing, AQI computation, input variable selection, model development, performance evaluation, and identification of the best predictive model.

2.1 Site Selection and Data Acquisition

Jaipur, the capital of Rajasthan, is located approximately 260 km southwest of New Delhi in northwestern India on a semi-arid plain (26°46'–27°01'N, 75°37'–76°57'E). This city covers an area of 467 km² with an estimated population of about 4.3 million in 2024. Surrounded by hills on most sides, Jaipur has observed a rapid increase in industrial, commercial, and residential activities, contributing substantially to deteriorating air quality.

For this study, we collected hourly air quality data for the years 2018–2019 from three continuous monitoring stations operated by the RSPCB which is Site 1: Jaipur Psychology Centre, Site 2: Jaipur Police Commissionerate Office, Site 3: Jaipur Science Park. The dataset included pollutant parameters such as PM₁₀, PM_{2.5}, benzene, NO₂, O₃, NH₃, CO as well as meteorological parameters such as temperature (Temp), relative humidity (RH), wind speed (WS), and atmospheric pressure (AP). The locations of the three sites are shown in Figure 1.

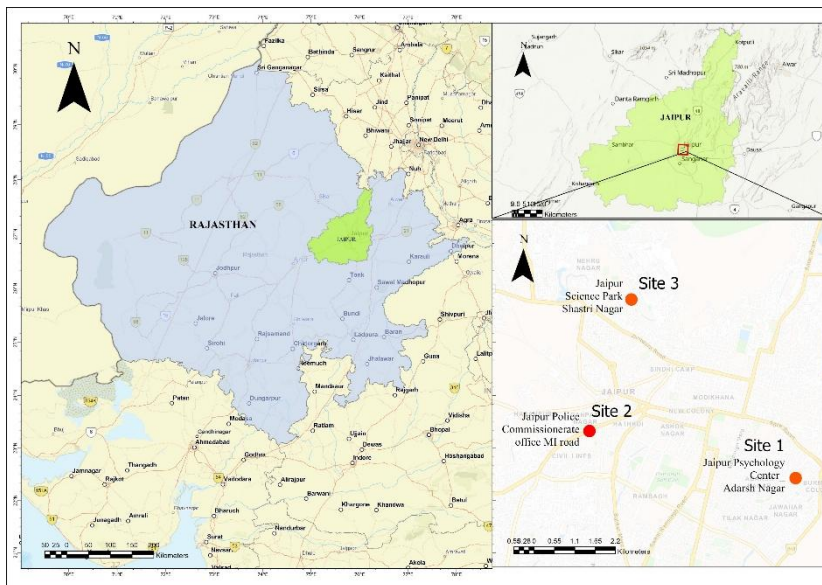


Fig. 1. Map of Jaipur city showing the three RSPCB air quality monitoring stations.

2.2 Data Preprocessing

Initial data analysis was conducted and Table 1 summarizes the general statistics, including mean, and standard deviation for all parameters across the three sites. Each site comprised 17,520 hourly records, corresponding to two years of continuous observations.

Table 1. Summary statistics of pollutant and meteorological parameters for the three sites.

Parameter	Site 1		Site 2		Site 3	
	Mean value	Standard deviation	Mean value	Standard deviation	Mean value	Standard deviation
PM _{2.5}	42.84	29.51	67.82	46.94	52.31	44.90
PM ₁₀	119.95	82.97	130.66	68.68	117.45	107.93
NO ₂	33.39	22.04	42.52	30.15	25.86	19.98
Benzene	1.46	2.72	1.34	1.97	1.21	1.96
O ₃	50.04	33.73	43.19	37.96	43.31	27.40
NH ₃	21.73	16.80	20.86	14.59	34.74	25.65
CO	0.95	0.8	1.05	1.60	0.96	2.04
RH	45.73	23.49	44.18	24.54	43.32	23.42
Temp	26.88	6.63	25.86	7.08	26.83	6.85
WS	1.24	0.68	0.89	0.56	1.03	0.61
AP	748.42	18.65	748.54	25.71	746.72	18.61

All units are in $\mu\text{g}/\text{m}^3$ except for CO (mg/m^3), Temp ($^{\circ}\text{C}$), WS (m/s), and RH (%).

2.3 Handling of Missing Data and Outliers

Missing or null values were identified using Python-based analysis. The percentage of missing values was 15.58% , 22.90%, and 24.06% for Site 1, Site-2 and Site-3, respectively. The missing values were replaced by the mean values in order to maintain the overall statistical consistency of the dataset while minimizing information loss.

Furthermore, outliers were detected as data points that deviated significantly from the average value due to measurement errors, irregular events, or natural fluctuations. Detection was done by visual tools such as box plots and histograms. Outlier removal was carried out using the upper and lower threshold limit method, based on National Ambient Air Quality Standards (NAAQS), 2009 [12]. For example, the 24-hour permissible limit for PM_{2.5} is 60 $\mu\text{g}/\text{m}^3$; hence, the upper threshold was set to three times this value (180 $\mu\text{g}/\text{m}^3$) and the lower limit to 10 $\mu\text{g}/\text{m}^3$. Similar thresholds were applied to other pollutants as well. Since there is no CPCB-defined limits for meteorological parameters, outlier removal was not applied to them.

2.4 AQI Estimation

AQI is an indicator representing overall air quality conditions. In India, AQI is calculated using CPCB standards [6]. It combines eight pollutants i.e., PM_{2.5}, PM₁₀, O₃, NO₂, SO₂, NH₃,

CO, and Pb, and classifies air quality into six categories with 0-50, 51-100, 101-200, 201-300, 301-400 and 401-500 indicating good, satisfactory, moderate, poor, very poor and severe air quality, respectively.

Using the following CPCB formula as shown in Equation 1, a sub-index for each pollutant was computed, and the maximum sub-index among various pollutants denotes the overall AQI:

$$I_p = [(I_{Hi} - I_{Lo}) / (BPH_i - BPL_o)] \times (C_p - BPL_o) + I_{Lo} \tag{1}$$

Where I_p = sub-index of pollutant p , C_p = observed concentration of pollutant p , and I_{Hi} , I_{Lo} = AQI values corresponding to breakpoint concentrations BPL_o , BPH_i

2.5 Selection of Input Parameters Using Pearson Correlation Coefficient (r)

To ensure model efficiency and avoid redundancy, input variables were selected using the Pearson correlation coefficient (r), which quantifies the linear relationship between two variables. It is defined as in Equation 2:

$$r = \frac{Cov(X,Y)}{\sigma_x \cdot \sigma_y} \tag{2}$$

Where, $Cov(X,Y)$ denotes covariance, and σ_x , σ_y denotes the standard deviation of variables X and Y , respectively. Values of r range from -1 to +1 with a positive value representing a direct linear relationship while a negative value indicates an inverse linear relationship, and zero implies no correlation. Parameters showing strong correlation with AQI were selected as input features for model development.

2.6 Machine Learning Models

Several ML models i.e., MLR, RF, SVR, XGBoost, AdaBoost, ANN, and CNN were developed for AQI prediction using pollutant and meteorological variables. The dataset was divided into training, validation, and testing sets in an 80:10:10 ratio. Model-specific details are summarized below:

Multiple Linear Regression (MLR): MLR assumes a linear relationship between multiple independent variables and a dependent variable (AQI). It determines both the collective predictive ability of input variables and their individual significance. The formula for MLR is given in the below Equation 3 [13]:

$$y_i = \alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2} + \dots + \alpha_p x_{ip} + \epsilon \tag{3}$$

Where, x_i = explanatory variable, y_i = dependent variable, α_p = slope value of each explanatory variable, α_0 = y-intercept, and ϵ = error term.

Random Forest (RF): RF is an ensemble learning method which builds multiple decision trees and averages their outputs to reduce overfitting. It efficiently handles non-linear relationships and large datasets.

Support Vector Regression (SVR): SVR forms a hyperplane to fit data within a tolerance margin, effectively handling non-linear and high-dimensional datasets. It minimizes deviation from the true values using kernel functions.

Extreme Gradient Boosting (XGBoost): XGBoost is an efficient ensemble learning method which sequentially builds decision trees to minimize residual errors. It includes L1/L2 regularization to reduce overfitting and supports parallel processing for large datasets. Its iterative learning process is given by the following Equation 4 [14]:

$$\hat{y}_i^t = \sum_{k=1}^t f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i) \quad (4)$$

Where, \hat{y}_i^t = forecasts at the stage t; $f_t(x_i)$ = a learner at stage t; x_i = the input variable; $\hat{y}_i^{(t-1)}$ = forecasts at the stage t-1. In this study, hyperparameters like learning rate, tree depth, and regularization terms were optimized during training.

Adaptive Boosting (AdaBoost): AdaBoost combines multiple weak learners (usually decision trees) sequentially, assigning higher weights to misclassified instances. This iterative reweighting increases classification accuracy and robustness for difficult to predict samples.

Artificial Neural Network (ANN): ANN imitates human brain learning by using interconnected neurons organized into layers. It captures complex, non-linear relationships and patterns in AQI data. Model parameters such as the number of hidden layers, neurons, activation functions, and learning rate were tuned in this work for optimal performance.

Convolutional Neural Network (CNN): CNNs extract spatial and temporal dependencies from multidimensional AQI time series data using convolution and pooling layers. Convolution layers perform feature extraction, while pooling layers decrease dimensionality. Again, various hyperparameters such as filter size, number of layers, and dropout rate were optimized for this model to enhance feature learning and minimize overfitting.

2.7 Model Evaluation

The performance of the developed models was measured with two standard statistical indices i.e., coefficient of determination (R^2) and root mean square error (RMSE). In case of regression models, a higher R^2 (closer to 1) and a lower RMSE values reflect better predictive accuracy and lower error.

Coefficient Of Determination (R^2): The R^2 value calculates the fraction of variance of the dependent variable which can be explained by the independent variables. It gives a estimation of how well the predicted values approximate the observed data [9]. A value of $R^2 = 1$ indicates a perfect prediction, while a value close to zero reflects poor model performance. The R^2 is computed using Equation 5 [7]:

$$R^2 = 1 - \frac{\sum_i^n (y_i - \hat{y}_i)^2}{\sum_i^n (y_i - \bar{y}_i)^2} \quad (5)$$

Where y_i is observed value of the i-th sample, \hat{y}_i is predicted value of the i-th sample, \bar{y}_i denotes mean of observed values, and n denotes total number of samples.

Root Mean Square Error (RMSE): The RMSE measures the average amount of prediction errors by computing the square root of the mean squared difference between predicted and observed values [7]. It displays the model's precision with lower RMSE values indicate higher accuracy and lesser deviations between predicted and actual data. The RMSE is computed using Equation 6 [7]:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (6)$$

Where y_i is observed value of the i-th sample, \hat{y}_i is predicted value of the i-th sample, and n denotes total number of samples. Together, R^2 and RMSE give a comprehensive evaluation of each model's predictive capability. R^2 examines how well the model explains variance, while RMSE computes the average prediction error.

3 Results and discussion

3.1 Variations in AQI Levels

The AQI for each site was calculated as described in Section 2.4. Figure 2 shows the percentage distribution of AQI categories (Good to Severe) for the three monitoring stations. Site 1 most frequently fell in the Moderate category (AQI 101–200), followed by Sites 3 and 2. In contrast, Site 2 showed the highest proportion of Poor (201–300) and Very Poor (301–400) conditions, indicating more frequent pollution episodes. Overall, Site 2 recorded the worst air quality, followed by Sites 3 and 1, respectively. This pattern is likely due to Site 2’s location in dense traffic, commercial activities, and closely spaced tall buildings, which limit pollutant dispersion and promote accumulation in the lower atmosphere.

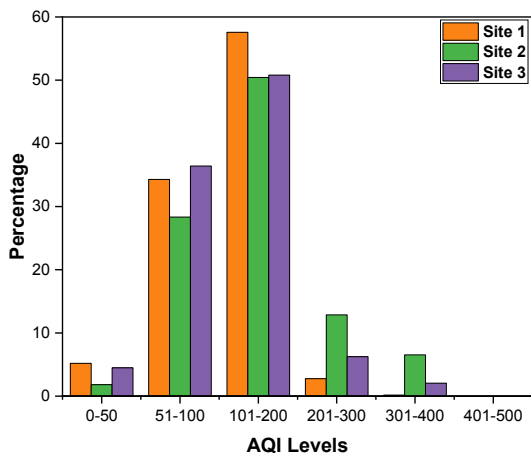


Fig. 2. Percentage distribution of AQI categories across the three sites in Jaipur city.

3.2 Selection of Input Variables using r Values

Pearson correlation coefficient (r) was utilized to assess the relationship between AQI and each pollutant and meteorological variable, and parameters with higher correlation values were selected as model inputs. For pollutants, $PM_{2.5}$ and ozone (O_3) showed strong positive correlations with AQI while for meteorological variables, relative humidity (RH), atmospheric pressure (AP), and temperature (Temp) were selected due to their moderate correlations. Importantly, RH and Temp showed negative correlations which align with known meteorological effects i.e., higher humidity (often after rainfall) enhances particle washout and reduces AQI, while lower winter temperatures promote temperature inversion and trap pollutants near the surface, increasing AQI values. Table 2 presents the correlation coefficients for all the selected variables across the three monitoring sites.

Table 2. Pearson correlation coefficient (r) values between AQI and selected pollutant and meteorological parameters.

Location	$PM_{2.5}$	Ozone	Relative Humidity	Atmospheric pressure	Temperature
Site 1	0.81	0.24	-0.43	0.27	-0.15
Site 2	0.92	0.18	-0.23	0.21	-0.18
Site 3	0.89	0.28	-0.28	0.14	-0.01

3.3 Development and Evaluation of Models

Several ML models i.e., MLR, RF, SVR, XGBoost, AdaBoost, ANN, and CNN were developed to predict AQI across the three sites and model performance was assessed using R^2 and RMSE indices. Three experimental scenarios were considered with Case 1: Pollutant parameters only, Case 2: Meteorological parameters only, and Case 3: Combined pollutant and meteorological parameters and results are discussed below.

Case 1: Using Pollutant Parameters: In this scenario, $PM_{2.5}$ and O_3 were used as input variables. The comparison of R^2 and RMSE values for models is shown in Figures 3a & 3b, respectively. The XGBoost model consistently showed the highest accuracy, with R^2 values ranging from 0.77 to 0.95 and RMSE values between 16.88 and 21.57 across the three sites. The superior performance of XGBoost is attributed to its ability to capture complex non-linear interactions, optimize decision trees through gradient boosting, and fine-tune hyperparameters efficiently. Following XGBoost, the ANN model performed next best, followed by RF, CNN, MLR, AdaBoost, and SVR. Among the sites, Site 2 achieved the highest predictive performance, followed by Site 3 and Site 1.

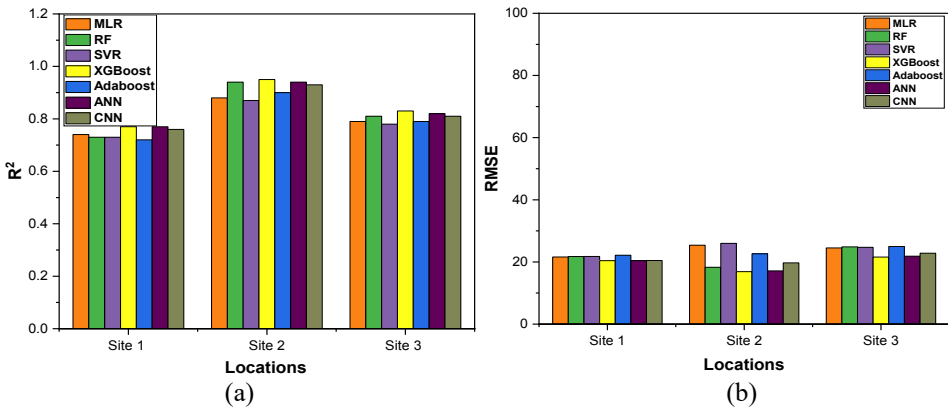


Fig. 3: Comparative performance of (a) R^2 and (b) RMSE for seven models using pollutant parameters.

Case 2: Using Meteorological Parameters: In this case, RH, temperature, and atmospheric pressure were used as predictors and R^2 and RMSE values are illustrated shown in Figures 4a and 4b, respectively.

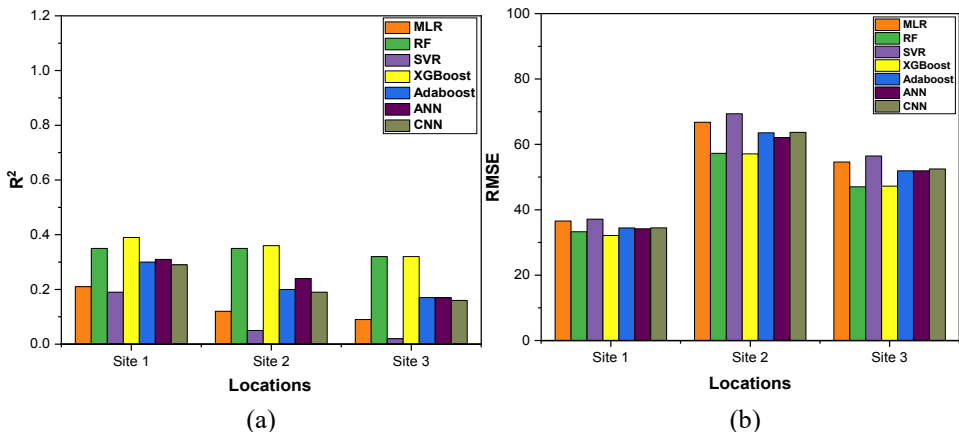


Fig. 4: Comparative performance of (a) R^2 and (b) RMSE for seven models using meteorological parameters.

The results revealed significantly lower R^2 and higher RMSE values compared to Case 1. This suggests that meteorological parameters alone are insufficient to accurately predict AQI, which may be due to they represent environmental conditions rather than pollutant concentrations. The lack of direct emission-related variables limited model performance. Nevertheless, XGBoost again outperformed other models, with R^2 values between 0.32 and 0.39 and RMSE ranging from 32.14 to 57.09.

Case 3: Using Both Pollutant and Meteorological Parameters: In the final scenario, both pollutant and weather variables i.e., $PM_{2.5}$, O_3 , and RH were selected as input parameters and results for R^2 and RMSE values are given in Figures 5a and 5b, respectively. It has been observed that the combined dataset produced significantly improved model performance compared to the previous two cases. This finding suggests that including meteorological conditions alongside pollutants provides a comprehensive understanding of AQI dynamics, allowing the models to capture both emission patterns and atmospheric influences. Again, XGBoost exhibited the best overall performance, with $R^2 = 0.77-0.95$ and $RMSE = 16.96-20.98$ across sites. The improvement confirms XGBoost’s capability to learn complex, non-linear relationships and its robustness against overfitting. Comparative analysis also revealed that Site 2 consistently yielded the most accurate predictions, aligning with its higher pollution levels and data variability. These findings are consistent with previous studies, including Van *et al.* [10], who reported XGBoost’s superior performance for AQI prediction across Indian cities, and Obodoeze *et al.* [15], who found XGBoost most effective in predicting $PM_{2.5}$ levels in Beijing, China.

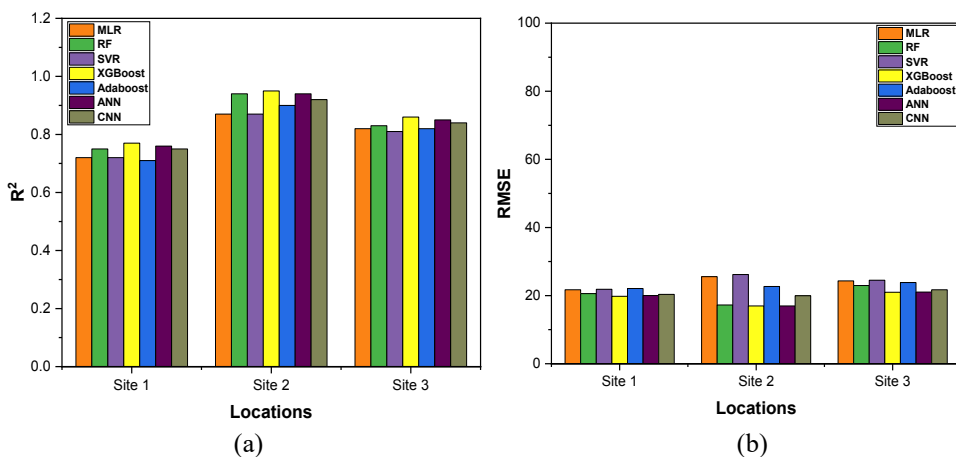


Fig. 5: Comparative performance of (a) R^2 and (b) RMSE for seven models using both pollutant and meteorological parameters.

Overall Discussion: Across all three cases, XGBoost evolved as the most accurate and reliable model for AQI prediction. Its gradient boosting approach, ability to handle feature interactions, and strong generalization make it particularly suited for urban air quality forecasting. Additionally, the results highlight that incorporating both pollutant and meteorological factors yields the most robust predictions, while models relying solely on weather variables perform poorly. This indicates the importance of integrating emission-based and climatic inputs to understand complex air pollution behavior in urban environments like Jaipur. However, despite the promising results, the study is limited by the use of data from only three sites over a two-year period, which may not fully capture spatial or seasonal variations in air quality. So, future work should incorporate longer duration and more monitoring station data for more robust predictions

4 Conclusions

This study provides a thorough analysis of AQI prediction for Jaipur using seven ML and DL-based models (MLR, RF, SVR, XGBoost, AdaBoost, ANN, and CNN) trained on two years of hourly data from three RSPCB monitoring stations. Missing values were handled through mean replacement, and outliers were removed using threshold-based filtering to improve data quality. Three modeling scenarios were developed: (1) pollutant parameters only, (2) meteorological parameters only, and (3) a combination of both. Performance evaluation using R^2 and RMSE showed that Case 3 consistently delivered the highest accuracy. Site 2 recorded the highest AQI levels, indicating a greater need for targeted mitigation in that region. Across all models, ML models outperformed DL-based approaches, with XGBoost achieving the best performance ($R^2 = 0.95$, RMSE = 16.88) at Site 2 because it can model non-linear relationships and optimize features effectively. Overall, the results indicate the strong potential of data-driven ML models, particularly XGBoost, for accurate AQI prediction and informed air quality management in urbanized cities like Jaipur. This is the first AQI prediction study for Jaipur, reflecting the value of advanced predictive analytics in providing public health benefits and policy support.

Acknowledgements & Other Statements: The authors thank the RSPCB for providing access to continuous ambient air quality monitoring data. This study received no external funding. Data will be made available upon request. Dr. Ruchi Sharma conceptualized the study, designed the methodology, and prepared the manuscript. Mr. Parth Banethiwal conducted data collection, literature review, and data analysis. Both authors reviewed and approved the final manuscript.

References

1. N. Manojkumar and B. Srimuruganandam, Age-specific and seasonal deposition of particulate matter in human respiratory tract. *Atmos. Pollut. Res.*, 13 (2), 101298 (2022). <https://doi.org/10.1016/j.apr.2021.101298>
2. I. E. Sitaras and P. A. Siskos, The role of primary and secondary air pollutants in atmospheric pollution: Athens urban area as a case study. *Environ. Chem. Lett.*, 6, 59–69 (2008). <https://doi.org/10.1007/s10311-007-0123-0>
3. World Health Organization, *Ambient air pollution: A global assessment of exposure and burden of disease* (2016). [Online]. Available: <https://www.who.int/publications/i/item/9789241511353>
4. K. Balakrishnan, A. Cohen, and K. R. Smith, Addressing the burden of disease attributable to air pollution in India: the need to integrate across household and ambient air pollution exposures. *Environ. Health Perspect.*, 122 (1), A6–A7 (2014). <https://doi.org/10.1289/ehp.1307822>
5. D. Sharma and D. Mauzerall, Analysis of air pollution data in India between 2015 and 2019. *Aerosol Air Qual. Res.*, 22 (2), 210204 (2022). <https://doi.org/10.4209/aaqr.210204>
6. Central Pollution Control Board, Air Quality Index August 2016 (2016). [Online]. Available: <https://cpcb.nic.in/displaypdf.php?id=bWFudWFsLW1vbml0b3JpbmcvQVFJX05BTVBfUmVwX0F1Z3VzdDIwMTYucGRm>
7. H. Liu, Q. Li, D. Yu, and Y. Gu, Air quality index and pollutant concentration prediction based on machine learning algorithms. *Appl. Sci.*, 9 (19), 4069 (2019). <https://doi.org/10.3390/app9194069>

8. K. Kumar and B. P. Pande, Air pollution prediction with machine learning: a case study of Indian cities. *Int. J. Environ. Sci. Technol.*, 20 (5), 5333–5348 (2023). <https://doi.org/10.1007/s13762-022-04241-5>
9. Y. C. Liang, Y. Maimury, A. H. L. Chen, and J. R. C. Juarez, Machine learning-based prediction of air quality. *Appl. Sci.*, 10 (24), 9151 (2020). <https://doi.org/10.3390/app10249151>
10. N. H. Van, P. Van Thanh, D. N. Tran, and D. T. Tran, A new model of air quality prediction using lightweight machine learning. *Int. J. Environ. Sci. Technol.*, 20 (3), 2983–2994 (2023). <https://doi.org/10.1007/s13762-022-04185-w>
11. N. N. Maltare and S. Vahora, Air Quality Index prediction using machine learning for Ahmedabad city. *Digit. Chem. Eng.*, 7, 100093 (2023). <https://doi.org/10.1016/j.dche.2023.100093>
12. Central Pollution Control Board, *National Ambient Air Quality Standards (NAAQS)* (2009). [Online]. Available: https://cpcb.nic.in/uploads/National_Ambient_Air_Quality_Standards.pdf
13. G. Mani and J. K. Viswanadhapalli, Prediction and forecasting of air quality index in Chennai using regression and ARIMA time series models. *J. Eng. Res.*, 10 (2A), 179–194 (2022). <https://doi.org/10.36909/jer.10253>
14. 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*, 338, 139518 (2023). <https://doi.org/10.1016/j.chemosphere.2023.139518>
15. F. C. Obodoeze, C. A. Nwabueze, and S. A. Akaneme, Comparative Evaluation of Machine Learning Regression Algorithms for PM2.5 Monitoring. *Am. J. Eng. Res.*, 10 (12), 19–33 (2021). <https://www.ajer.org/papers/Vol-10-issue-12/C10121933.pdf>