

Analysis and Comparison of Air Quality Index Prediction using Regression Based Machine Learning Models

Priyanka Goyal*

School of Information and Communication Technology, Gautam Buddha University, Greater Noida, India

E-mail: priyankag@gbu.ac.in

ORCID iD: <https://orcid.org/0000-0003-2512-1629>

*Corresponding author

Utkrisht Patel

School of Information and Communication Technology, Gautam Buddha University, Greater Noida, India

E-mail: utkrisht207patel@gmail.com

ORCID iD: <https://orcid.org/0009-0008-1504-2955>

Received: March 18, 2026; Revised: June 18, 2026; Accepted: July 9, 2026; Published: August 8, 2026

Abstract: This study employs multiple machine learning regression techniques to predict air quality indices and conducts a comparative analysis to identify the optimal model based on data efficiency and processing time. The Mean Absolute Error (MAE), Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE) were used as evaluation measures to compare these regression models. Furthermore, the processing time of each algorithm was evaluated through independent training and hyperparameter tuning to determine the best-fit model regarding computational efficiency and error rates. In this paper, we have calculated the custom score which is sum of MAE, RMSE, MAPE and time processing values. The best model obtained is the custom stacked regression model has custom score of 111.41 which is very less as compared to other regression models.

Index Terms: Air Quality Index (AQI), Regression-Based Models, Mean Absolute Percentage Error (MAPE), Mean Absolute Error (MAE), Root Mean Square Error (RMSE)

1. Introduction

One of the most critical threats to human health is air pollution. According to the World Health Organization (W.H.O) air pollution puts the health of seven million people at risk. Health impacts include asthma, skin infections, cardiovascular issues, respiratory diseases, bronchitis, lung cancer, and ocular conditions [1]. Air pollution is a serious issue that has negative consequences for human health and living situations. The most basic necessity and requirement for human living is air, which is a mixture of several gases such as oxygen and nitrogen. Air is essential for human survival, making its quality a critical public health concern. Air pollution, characterized by the presence of hazardous particulate matter and gases, poses severe threats to human health and environmental sustainability. As a result, constant air pollution analysis is required. Regression algorithms are particularly suitable for predicting continuous air quality metrics. One such category is regression-based machine learning algorithms, including Decision Tree Regression and Random Forest Regression [2].

This research work, discusses the working of how regression models can help to predict the values of the AQI based on the input data. A comparative analysis is conducted to identify the optimal model. All the models that have been implemented in this work for AQI prediction are regression-based models. In this paper, we have implemented the custom stacked regression model.

Custom stacking is an ensemble learning technique that uses meta-regression to combine multiple regression models. An individual regression model is trained using the full training set, and a meta-regression is fitted using the output (meta-features) of the individual regression models in the set. Custom stacked model is used to improve model prediction by mixing the outputs of many models and passing them through meta-learning models, which are machine learning models. The paper is organized as follows: In section 2 literature review is presented which predicts the toxins

in air and how to improve them. Section 3 explains the implementation of the proposed work, and Section 4 presents the calculation of AQI parameters using regression-based models. Section 5 presents the conclusion and future work.

2. Literature Review

Pollution of air in cities has a direct impact on living being health, particularly in developing and industrialized countries where air quality controls are either unavailable or ineffectively implemented or enforced. Recent studies have found substantial evidence that exposure to air contaminants linked to various health problems, including asthma and lung irritation. Zheng et al. [3] suggested a new method that accounts for temporal and spatial correlations. An inflexion predictor, a spatial predictor, a prediction aggregator, and a temporal predictor are four components of this architecture. While separating the wind directions into eight classes, Zheng's approach had one flaw: it did not take into account the topography information. Simultaneously, the wind, which is influenced by terrain, has a substantial impact on PM 2.5 dispersion. As far as social insurance administration costs go, the association between air pollution and human well-being inexorably increases medical clinic confirmations and emergency department visits. Since air quality has a significant impact on human life, W.H.O has developed guidelines to reduce the negative effects of air pollution on public health. The W.H.O has set limits on the concentrations of surface Ozone (O₃), Nitrogen dioxide (NO₂), and Sulphur dioxide (SO₂) in the air [4, 5]. Suh et al. [6] compare the surrounding positions using the nearest neighbour method using the Euclidean distance method ('KNN-ED') and the K nearest neighbour method using the distance DTW method ('KNN-DTWD'), time series at different sites. The model proposed by Suh et al. [6] demonstrated superior accuracy compared to Zheng et al. [3] for short-term forecasting. The 'KNN-DTWD' strategy outperformed the 'KNN-ED' approach when tested experimentally, however in other circumstances, such as flat terrain, the 'KNN-ED' method outperformed the 'KNN-DTWD'. A wireless air quality monitoring framework was proposed in [7], utilizing a General Packet Radio Service (GPRS) modem to transmit collected pollutant levels. An intelligent sensor-based system for air quality monitoring in remote areas was introduced in [8]. Similar to Ong et al. [9], deep learning methods were employed to estimate AQI, specifically using a pre-trained Recurrent Neural Network (RNN) to predict PM_{2.5} levels. The grid-based division method splits the entire area into subsequent grids and then mimics spatial and quality of air data, as the name implies. Shepard's method divides the land into two-dimensional grids for spatial interpolation, assuming that pollutant concentrations vary based on distance from known measurement points. Bai et al. [10] used satellite observation data and uniform grids to interpolate PM_{2.5} concentrations. Tang et al. [11] employed a grid-based method to determine the spatial interpolation of PM_{2.5} concentrations. They gathered data from numerous sources, constructed division grids, and discovered the spatial interpolation of PM 2.5 concentrations. Goodin et al. [12] constructed a three-dimensional grid of wind fields for an urban area. In [13], an optimized machine learning model combining Grey Wolf Optimization with a Decision Tree algorithm was used for accurate AQI prediction in large cities. Further improvements in model architecture could enhance prediction accuracy. In [14], the authors proposed a deep learning framework for AQI prediction; however, it struggled to capture long-term trends during periods of sudden climate change. In [15], a framework utilizing various machine learning techniques was designed for AQI prediction; however, the study was limited to a small, region-specific dataset over a short time span. Complete study shows that various machine learning models have been used to find the AQI of the cities.

Evaluation Criteria

This study predicts AQI based on air pollutant components. Performance is assessed using Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), runtime, and a custom score.

MEAN ABSOLUTE ERROR (MAE)

Mean Absolute Error (MAE) calculates the average absolute difference between predicted and actual values. It is a scale-dependent metric that measures inaccuracy on the same scale as the observations. MAE is a standard metric for evaluating regression models by quantifying the discrepancy between observed and predicted values.

$$MAE = \frac{1}{n} \sum_{i=1}^n y_i - \hat{y}_i \quad (1)$$

ROOT MEAN SQUARE ERROR (RMSE)

Root Mean Square Error (RMSE) is the square root of the mean of squared errors. It penalizes larger errors more heavily than MAE and is expressed in the same units as the target variable.

MEAN ABSOLUTE PERCENTAGE ERROR (MAPE)

Mean Absolute Percentage Error (MAPE) measures the average magnitude of errors in percentage terms,

providing a scale-independent assessment of prediction accuracy.

$$\text{MAPE} = ((|A-F|/A)*100)/N \quad (2)$$

The actual value is A, the anticipated value is F, and the no. of observations is N.

CUSTOM SCORE

Custom Score is the sum of MAE, RMSE, MAPE and runtime/processing time of all the regression-based ML models used in this research. The Custom Score will tell the overall final results of all the regression-based machine learning models.

$$\text{Custom Score} = \text{MAE} + \text{RMSE} + \text{MAPE} + \text{Runtime} \quad (3)$$

3. Novelty of the Proposed Work

In this work, the novelties aim towards improving the accuracy of the proposed model. The first novelty is using the Custom Stacked Model instead of a classic Random Forest Regression model:

Custom Stacked Regression (CSR) is an ensemble technique that combines numerous regression models using a meta-regressor. Individual machine learning regression model can be trained using the whole training set and the "meta-regressor" is fitted using the ensemble's individual machine learning regression models outputs (meta-features). In figure 1, the working of custom stacked model is explained. In this model, two level stacking ensemble regression model is explained. In the first level, Random Forest, XG Boost, Gradient Boosting Regressor, and decision tree predicts AQI. In the second level, Meta Regressor combines the individual value of first level to create the output. The ensemble strategy makes the system more accurate and reliable for AQI prediction.

The second novelty is using Random Search CV instead of Grid Search CV for hyper-parameter tuning: Random Search is a strategy for finding the optimum solution for a created model by using random combinations of hyper-parameters. It is similar to grid search, but it is shown to produce better results when compared.

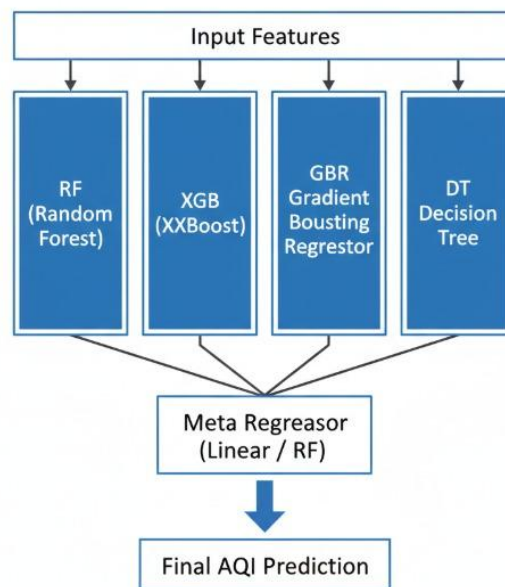


Fig. 1. Architecture of the Proposed Custom Stacked Regression Model

4. Proposed Work Implementation Approach

The Implementation approach of this research work is explained using following four steps to predict the AQI in figure 2.

Step1: Imputation: Data in its raw form contains a large number of missing values (nulls), making it less

trustworthy and valuable for machine learning. ML algorithms, despite their sophistication, have the issue of being unable to utilize data having missing values. Imputation comes into play here. Imputation is the process of using mathematical and statistical approaches such as mean, median and others to fill in missing numbers. Imputation can be done using a variety of procedures, including median, mean, most-frequent and constant. If the count of items in the data is even, the median equals the second element; otherwise, the median equals the third element. $N+1/2$ th element = Median.

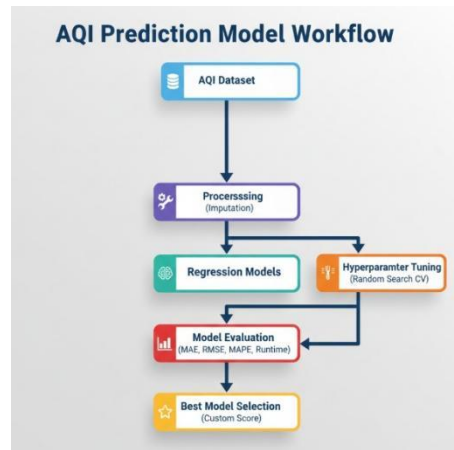


Fig. 2. Flow Chart of the AQI Prediction Model (Proposed System)

Step 2: Model Training: The Machine learning model is provided with data to train it in this step. This is where the learning happens. Consistent training can improve the ML model prediction rate significantly. The weights of the model must be chosen at random. As a result, the algorithms will learn to appropriately change the weights. Figure 3 presents an overview of the AQI dataset and the preprocessing steps applied in the proposed work. The raw dataset consists of multiple air pollutant parameters, which are cleaned using imputation techniques and normalized to generate a reliable dataset before model training.

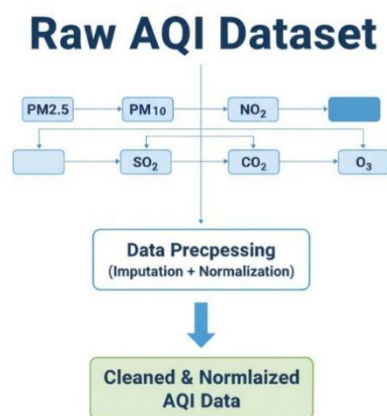


Fig. 3. AQI Dataset Description and Preprocessing Pipeline

Step 3: Hyper-parameter Tuning: When developing machine learning models, hyper-parameter tuning is a means of altering the parameters supplied as tuples. We set these parameters, and the programmer can adjust them as needed. These parameters are never learned by machine learning algorithms. To ensure the success of model, these were fine-tuned. The goal of hyper-parameter tweaking is to identify the parameters that provide the model the best or most results with the lowest error rate.

Randomized Search CV: When there are a lot of parameters to try and the training period is extensive, Randomized Search CV comes in handy.

Step 4: Predictions: The suitable regression models are applied after the imputation. The predictions are generated as the final phase. The predicted AQI values are then compared to the actual AQI values. The model's efficiency is calculated using MAE, RMSE, MAPE, and Runtime. The accuracy of the model is inversely related to the MAE, RMSE and MAPE values, i.e., the lower the values, the higher the accuracy and vice versa.

5. Results and Analysis

For the prediction of AQI, regression based models are used to calculate the values of MAE, RMSE, MAPE and runtime/processing time. The based model is selected on the basis of low custom score. Figure 4, 5, 6, 7 and 8 shows the comparative study of the MAE, RMSE, MAPE, Runtime and Custom Score of all the regression models. In table 2, we observed that the Randomized Search CV Decision Tree Regression model shows the least RMSE whereas Randomized Search CV Random Forest Regression model and Custom Stacked Regression model are showing almost the same RMSE. In table 3, we can observe that the Custom Stacked Regression model and Randomized Search CV Random Forest regression model are showing less MAPE value. Till now, Randomized Search CV Random Forest Regressor and Custom Stacked Regressor are showing good results. To select the best model among these, we examine these regression-based machine learning models on the basis of their runtime and the custom score given in table 4. The custom score of all the regression-based machine learning models used in this study shown in table 5. It is clearly visible that the Custom Stacked Regression model has shown the best results compared to other regression models given in table 6.

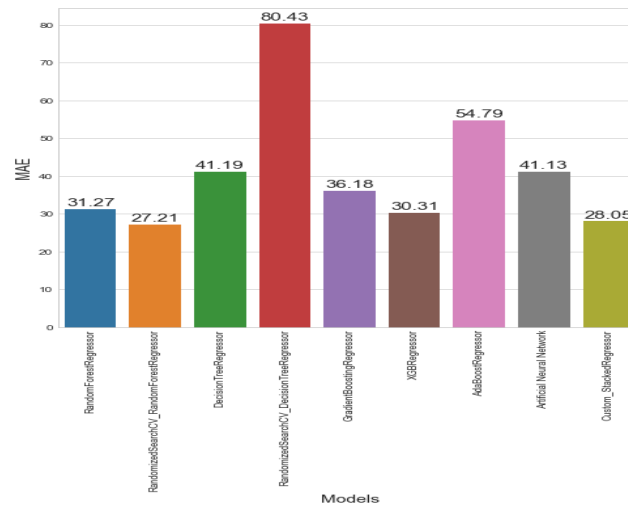


Fig. 4. Mean Absolute Error

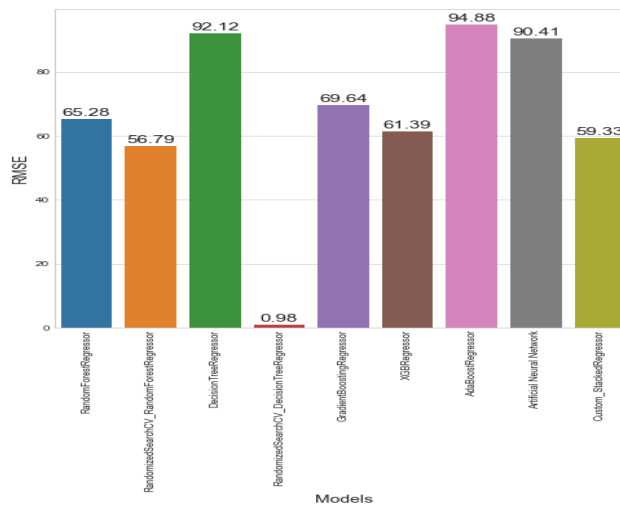


Fig. 5. RMSE of regression models

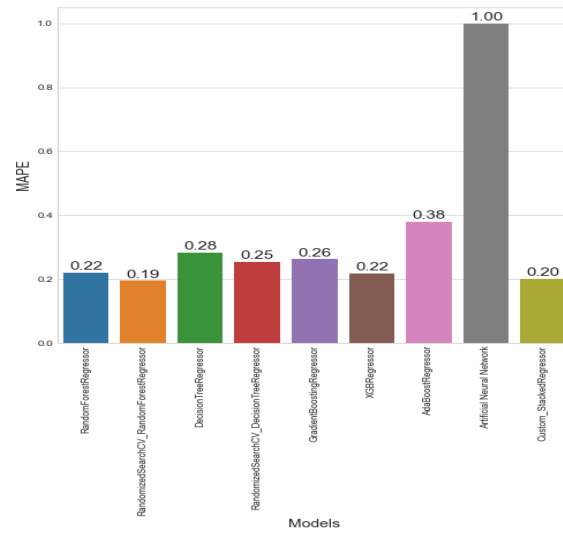


Fig. 6. MAPE of Regression Models

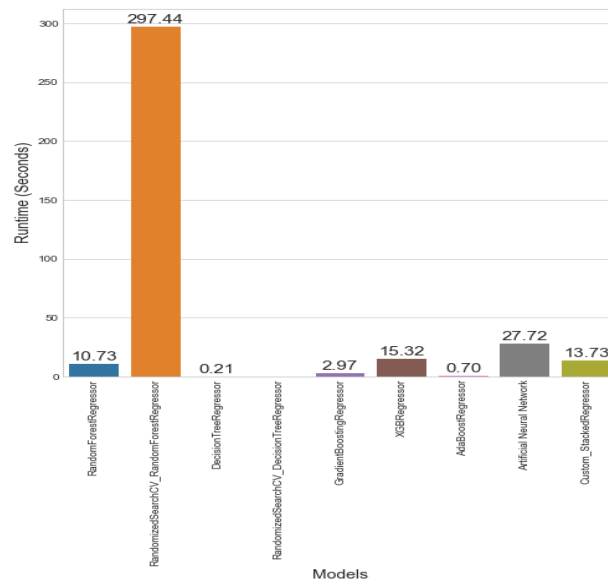


Fig. 7. Runtime of Regression Models

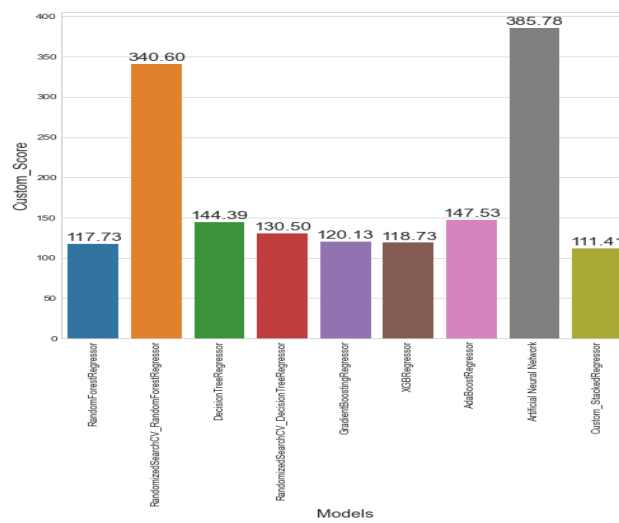


Fig. 8. Custom Score of Regression Models

In table 1, we can observe that Random Forest Regression model (using Randomized Search CV) and Custom Stacked Regression model show less mean absolute error as compared to other models.

Table 1. MAE of Models (Low to High Error)

REGRESSION MODELS	MAE
Randomized Search CV Random Forest Regressor	27.21
Custom Stacked Regressor	28.05
XGB Regressor	30.31
Random Forest Regressor	31.27
Gradient Boosting Regressor	36.18
Artificial Neural Network	41.13
Decision Tree Regressor	41.19
AdaBoost Regressor	54.79
Randomized Search CV Decision Tree Regressor	80.43

Table 2. RMSE of Models (Low to High Error)

REGRESSION MODELS	RMSE
RandomizedSearchCV Decision Tree Regressor	0.98
RandomizedSearchCV Random Forest Regressor	56.79
Custom Stacked Regressor	59.33
XGB Regressor	61.39
Random Forest Regressor	65.28
Gradient Boosting Regressor	69.64
Artificial Neural Network	90.41
Decision Tree Regressor	92.12
AdaBoost Regressor	94.88

Table 3. MAPE of Models (Low to High Error)

REGRESSION MODELS	MAPE
Randomized Search CV Random Forest Regressor	0.19
Custom Stacked Regressor	0.20
XGB Regressor	0.22
Random Forest Regressor	0.22
Randomized Search CV Decision Tree Regressor	0.25
Gradient Boosting Regressor	0.26
Decision Tree Regressor	0.28
Adaboost Regressor	0.38
Artificial Neural Network	1.00

Table 4. Runtime of Models (Low to High)

REGRESSION MODELS	RUNTIME
Randomized Search CV Decision Tree Regressor	0
Decision Tree Regressor	0.21
Adaboost Regressor	0.70
Gradient Boosting Regressor	2.97
Random Forest Regressor	10.73
Custom Stacked Regressor	13.73
XGB Regressor	15.32
Artificial Neural Network	27.72
Randomized Search CV Random Forest Regressor	297.44

Table 5. Custom Score of Models (Low to High)

REGRESSION MODELS	CUSTOM SCORE
Custom Stacked Regressor	111.41
Random Forest Regressor	117.73
XGB Regressor	118.73
Gradient Boosting Regressor	120.13
Randomized Search CV Decision Tree Regressor	130.50
Decision Tree Regressor	144.39
Adaboost Regressor	147.53
Randomized Search CV Random Forest Regressor	340.60
Artificial Neural Networks	385.78

Figure 9 shows the comparison between actual AQI values and predicted AQI values obtained using the proposed Custom Stacked Regression model, indicating close alignment and improved prediction accuracy.

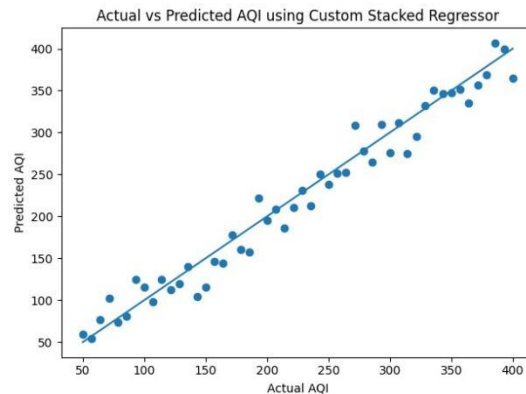


Fig. 9. Predictive AQI vs Actual AQI plot using proposed work

Table 6. Comparison of all the regression-based models

REGRESSION MODELS	MAE	RMSE	MAPE	RUNTIME	Custom Score
Random Forest Regressor	31.27	65.28	0.22	10.73	117.73
Randomized Search CV Random Forest Regressor	27.21	56.79	0.19	297.44	340.60
Decision Tree Regressor	41.19	92.12	0.28	0.21	144.39
Randomized Search CV Decision Tree Regressor	80.43	0.98	0.25	0.0	130.50
Gradient Boosting Regressor	36.18	69.64	0.26	2.97	120.13
XGB Regressor	30.31	61.39	0.22	15.32	118.73
Ada Boost Regressor	54.79	94.88	0.38	0.70	147.53
Artificial Neural Network	41.13	90.41	1.00	27.72	385.78
Custom Stacked Regressor	28.05	59.33	0.20	13.73	111.41

6. Conclusion and Future Work

This research work, deals with the AQI predictions that are obtained by the regression-based machine learning models. Various authors have implemented ML models in order to tackle the same problem that has been handled in this paper. Researchers have implemented methods like grid division, cluster division, and KNN- DTW to predict the values for various AQI components. This study, discussed the application of regression models like Custom Stacked Regressor, decision tree or random forest model etc. also hyper-parameter optimization tool is used that is Random Search CV in the same domain and used to make predictions. The biggest issue that affects any ML model is the presence of null values, which are handled using imputation. The MAE, RMSE, MAPE and Runtime values for all the models used in this paper. The comparison shows that the custom stacked model made predictions with a lesser error rate than other regression models. The comparison between the MAE, RMSE, MAPE, Runtime and Custom Score values for all the models is done.

Future work in this domain can include the application of LSTM models to make predictions, applying CNN to make image based AQI predictions and using hyper-parameter optimization tools like Bayesian search CV to optimize the parameters for the model.

All the Declarations and Statements

Author Contributions Statement

Priyanka Goyal – Conceptualization, Methodology, high quality research and Supervision

Utkrisht Patel – Data Curation and Software Implementation: Collected and preprocessed datasets, handled data organization, and implemented the machine learning and authentication modules used in the proposed system.

All authors have read and agreed to the published version of the manuscript.

Conflict of Interest Statement

The authors declare that there are no conflicts of interest regarding the publication of this manuscript.

Funding Declaration

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Data Availability Statement

The datasets used and/or analyzed during the current study are available from the corresponding author upon reasonable request.

Ethical Declarations

The authors confirm that this research was conducted in accordance with ethical research standards. No human participants or animals were directly involved in the study. Any datasets used for experimentation were anonymized and utilized solely for academic and research purposes. The authors also confirm that the manuscript is original, has not been published previously, and is not under consideration for publication elsewhere.

Acknowledgments

Not applicable.

Declaration of Generative AI in Scholarly Writing

During the preparation of this manuscript, the authors utilized generative AI tools to assist with figures improvement. The authors carefully reviewed and edited all generated content and take full responsibility for the accuracy, originality and integrity of the final submitted manuscript.

Abbreviations

The following abbreviations are used in this manuscript:

Abbreviation - Full Form

AI - Artificial Intelligence

MAE- Mean Absolute Error

RMSE-Root Mean Square Error

MAPE- Mean Absolute Percentage Error

AQI- Air Quality Index

CSR- Custom Stacked Regression

Appendix A/B/C..., with appendix title

None.

References

- [1] Abdelmalek, M.M., Mahmoud, H. & Shokry, H. "Prognosis of air quality index and air pollution using machine learning techniques", *Sci Rep* 15, 25890 (2025). <https://doi.org/10.1038/s41598-025-11260-y>
- [2] Pande, C.B., Radwan, N., Heddarn, S. et al. "Forecasting of monthly air quality index and understanding the air pollution in the urban city, India based on machine learning models and cross-validation", *J Atmos Chem* 82, 1 (2025). <https://doi.org/10.1007/s10874-024-09466-x>
- [3] Y. Zheng et al., "Forecasting Fine-Grained Air Quality Based on Big Data," Aug. 2015. Accessed: Jun. 07, 2022. <https://doi.org/10.1145/2783258.2788573>
- [4] S. E. Kim, "Ordinal time series model for forecasting air quality index for ozone in Southern California," *Environmental Modeling & Assessment*, vol. 22, no. 2, pp. 175–182, Jul. 2016. <https://doi.org/10.1007/s10666-016-9521-7>.
- [5] S. Qin, F. Liu, C. Wang, Y. Song, and J. Qu, "Spatial-temporal analysis and projection of extreme particulate matter (PM10 and PM2.5) levels using association rules: A case study of the Jing-Jin-Ji region, China," *Atmospheric Environment*, vol. 120, pp. 339–350, Nov. 2015. <http://doi.org/10.1016/j.atmosenv.2015.09.006>.
- [6] P.-W. Soh, K.-H. Chen, J.-W. Huang, and H.-J. Chu, "Spatial-Temporal pattern analysis and prediction of air quality in Taiwan," Aug. 2017. <http://doi.org/10.1109/umedia.2017.8074094>
- [7] T. Rakthanmanon et al., "Searching and mining trillions of time series subsequences under dynamic time warping," 2012. Accessed: Jun. 07, 2022. <http://doi.org/10.1145/2339530.2339576>
- [8] D. Shepard, "A two-dimensional interpolation function for irregularly-spaced data," 1968. Accessed: Jun. 07, 2022. <http://doi.org/10.1145/800186.810616>.
- [9] B. T. Ong, K. Sugiura, and K. Zettsu, "Dynamically pre-trained deep recurrent neural networks using environmental monitoring data for predicting PM2.5," *Neural Computing and Applications*, vol. 27, no. 6, pp. 1553–1566, Jun. 2015. <https://doi.org/10.1007/s00521-015-1955-3>.

- [10] Y. Bai, L. Wu, K. Qin, Y. Zhang, Y. Shen, and Y. Zhou, "A Geographically and Temporally Weighted Regression Model for Ground-Level PM_{2.5} Estimation from Satellite-Derived 500 m Resolution AOD," Remote Sensing, vol. 8, no. 3, pp. 262, Mar. 2016, [https://doi: 10.3390/rs8030262](https://doi.org/10.3390/rs8030262).
- [11] M. Tang, X. Wu, P. Agrawal, S. Pongpaichet, and R. Jain, "Integration of Diverse Data Sources for Spatial PM_{2.5} Data Interpolation," IEEE Transactions on Multimedia, vol. 19, no. 2, pp. 408–417, Feb. 2017. [https://doi: 10.1109/tmm.2016.2613639](https://doi.org/10.1109/tmm.2016.2613639).
- [12] W. R. Goodin, G. J. McRae, and J. H. Seinfeld, "An Objective Analysis Technique for Constructing Three-Dimensional Urban-Scale Wind Fields," Journal of Applied Meteorology, vol. 19, no. 1, pp. 98–108, Jan. 1980. [https://doi: 10.1175/1520-0450\(1980\)](https://doi.org/10.1175/1520-0450(1980)).
- [13] Natarajan, S.K., Shanmurthy, P., Arockiam, B. Balamurugun, S. Shitharth, "Optimized machine learning model for air quality index prediction in major cities in India", Scientific Reports, Vol. 14, 6795, 2024. <https://doi.org/10.1038/s41598-024-54807-1>
- [14] Guangyao Ma, Kai Xu, Yue Zhang, Lanhe Zhang, Zicheng Chen, "A novel deep learning model for air quality index prediction integrating time series decomposition and intelligent optimization", Results in Engineering, Vol. 27, 106078, 2025. <https://doi.org/10.1016/j.rineng.2025.106078>
- [15] Madan, T., Sagar, S., Singh, S. Yamini, J. Achin, K. Ondrej, F. Robert, N. Hamidreza. P. Arvind, "Hybrid deep learning model for air quality prediction and its impact on healthcare", Scientific Reports, Vol. 16, 6036, 2026. <https://doi.org/10.1038/s41598-026-36564-5>

Authors' Profiles



Priyanka Goyal, Ph.D, is an assistant professor in Electronics & Communication Engineering Department. She received the Bachelor of Technology from the Kurukshetra University. She is an author of more than thirty research publications. Her area of interest includes IOT, Artificial Intelligence and Machine Learning and Optoelectronics.



Utkrisht Patel, B.Tech., Utkrisht Patel is currently pursuing a bachelor's degree in computer science from Gautam Buddha University. He is an undergraduate researcher with active contributions in the domains of Internet of Things (IoT), Machine Learning (ML) and Reinforcement Learning (RL). He has authored and co-authored three research papers in reputed conferences and journals. His research focuses on developing intelligent, scalable and real-world problem-solving systems, particularly in smart environments and AI-driven applications. His areas of interest include data analytics, embedded systems and adaptive learning models. He is committed to advancing his research capabilities and contributing to innovative technological solutions.

How to cite this paper

Priyanka Goyal, Utkrisht Patel, "Analysis and Comparison of Air Quality Index Prediction using Regression Based Machine Learning Models", International Journal of Education and Management Engineering (IJEME), Vol.16, No.4, pp. 61-70, 2026. DOI:10.5815/ijeme.2026.04.05.