Physical Education, Health and Social Sciences

https://journal-of-social-education.org

E-ISSN: 2958-5996
P-ISSN: 2958-5988

A Hybrid Machine Learning Framework for Prediction, Classification, and Regional Clustering of Air Quality Index

Rubeen Fatima[1], Tania Maryam[2], Saeed Rasheed[3]
[1,2] Department of Physics, University of Agriculture, Faisalabad, Punjab, Pakistan.
[3] Department of Computer science, University of Agriculture Faisalabad, Pakistan,
Email: saeed.rasheed0211@gmail.com
DOI: https://doi.org/10.63163/jpehss.v4i2.1450

Abstract
Air pollution has emerged as one of the most critical environmental and public health challenges worldwide. Poor air quality is linked to respiratory diseases, cardiovascular disorders, a short life span and significant economic loss. Thus, correct prediction and classification of air quality is crucial for any effective monitoring and decision making in the environment. The application of machine learning, in predicting and classifying the Air Quality Index (AQI), based on the concentration of atmospheric pollutants and meteorological parameters is investigated in this study. The machine learning algorithms used for the analysis of air quality patterns were: Linear Regression, Polynomial Regression, K-Nearest Neighbors (KNN), and K-Means Clustering. The regression models were used to predict the AQI, the KNN and K-Means were used for classification and clustering of pollution levels. Data consisted of PM2.5, PM10, NO_2, SO_2, CO, O_3, temperature and humidity. The experimental results showed that machine learning models can effectively predict the AQI and categorize pollution. PolyNomial Regression model performed best among the evaluated models because it was able to capture the nonlinear relationships among the environmental variables. The results indicate that machine learning-based AQI forecasting systems can assist the environmental authorities in pollution control and management strategies.

Key Words: Air Quality Index, Machine Learning, Air Pollution Prediction, Polynomial Regression, KNN, K-Means Clustering, Environmental Monitoring

Introduction
Air pollution is an important environmental problem in developed countries as well as developing countries. With a high level of industrialization, urbanization, growth in traffic and high consumption of fossil fuels, the concentration of harmful pollutants in the atmosphere has risen. Exposure to polluted air has been associated with various health issues such as asthma, chronic respiratory diseases, cardiovascular diseases, and early death. The Air Quality Index (AQI) is a common tool used to assess and report air quality. It transforms the concentration of pollutants into a number expressing the potential effects on human health due to air pollution. Clean air is a level of 100 or lower, and the higher the number, the more problematic the air quality conditions. The traditional statistical forecasting methods (TSFM) have difficulty in capturing complex and nonlinear relationships between environmental variables. Machine learning algorithms, on the other hand, offer sophisticated computational techniques that can detect complex patterns and make precise predictions based on vast amounts of data. They are increasingly becoming popular in the use of multidimensional data for environmental monitoring and forecasting applications. There have been some studies that focused on specific machine learning techniques for AQI prediction, but few studies have integrated the three functions of prediction, classification, and clustering into one system. Moreover, most of the current studies are concerned with the accuracy of forecasts and neglect the identification of the pollution pattern and regionalization. For overcoming the above mentioned shortcomings,

this work presents an integrated machine learning based framework for air quality assessment, which includes both regression and classification techniques along with clustering techniques. Linear Regression, Polynomial Regression are used to predict AQI, K-NN is used to classify AQI and K-Means Clustering to identify the pollution pattern using the proposed framework. The goal is to test and examine the performance of the algorithms in air quality monitoring and in decision support for air quality.

Research Objectives:
Use regression machine learning models to predict Air Quality Index (AQI).
Determine air quality classification based on supervised learning.
- Find patterns in pollution using unsupervised clustering.
- Compare the performance of several machine learning algorithms.
- Offer suggestions on environmental monitoring and pollution control.

Dataset
In this study, data was obtained from air pollution monitoring stations. The data include concentrations of the major air quality pollutants and parameters of the atmosphere that affect air quality.

Input Features

Feature	Description	Unit
PM2.5	Fine particulate matter	µg/m³
PM10	Coarse particulate matter	µg/m³
NO_2	Nitrogen dioxide	ppb
SO_2	Sulfur dioxide	Ppb
CO	Carbon monoxide	Ppm
O_3	Ozone	Ppb
Temperature	Ambient temperature	°C
Humidity	Relative humidity	%

Target Variable
The aim of this model is to predict the overall air quality condition (AQI).

Data Preprocessing
The data run through numerous preparation steps before the construction of the model:
- Removing duplicate records.
- Imputing missing values with mean values.
- Outlier detection and treatment.
- Normalize and standardize features.
- Dividing data into training and testing subsets.

Data preprocessing not only increases the quality of data but also helps boost model performance.

Methodology
The proposed methodology is based on a systematic machine learning process.

Step 1: Data Collection
Air quality data are gathered from air quality monitoring stations and public air quality databases.

Step 2: Data Cleaning
Records are incomplete and observations are inconsistent and corrected to enhance data reliability.

Step 3: Exploratory Data Analysis (EDA)
Data are summarized and displayed using statistics and visualization to gain insights into relationships between variables and the distribution of data.

Step 4: Feature Engineering
The pollutants and weather variables that are relevant are chosen for building models.

Step 5: Model Training
The data is split into:

- Training Data (80%)
- Testing Data (20%)

Historical AQI observations are used for training the machine learning algorithms.

Step 6: Model Evaluation
They are trained on standard performance measures and then tested on the trained models.

Evaluation Metrics
Mean Squared Error (MSE)
Measures mean-squared errors of prediction.

Root Mean Squared Error (RMSE)
Offers prediction error, in the original unit scale.

R-Squared (R^2)
Compares the amount of variance accounted for by the model.

Accuracy
Used to assess the accuracy of classification.

Literature Review
A number of researchers have used the machine learning methods to predict and classify the air quality. Their analyses show that machine learning models can outperform traditional statistical models in many instances because they can better capture the complex and nonlinear relationships between environmental variables. Zhang et al. (2018) used Linear Regression (LR) and Support Vector Regression (SVR) methods to forecast the AQI of major cities in China. They found that the machine learning methods outperformed traditional regression methods in prediction accuracy. To build an air pollution prediction system, Li and Wang (2019) applied K-Nearest Neighbors (KNN) to the task. They were able to achieve a classification accuracy of more than 85% for predicting AQI categories like Good, Moderate and Unhealthy. Kumar et al. (2020) used Polynomial Regression for modeling nonlinear relationships between particulate matter concentration and AQI. They found that the errors in prediction of Polynomial Regression were lower than the errors in prediction of Linear Regression. In urban environments, Sharma et al. (2021) pinpointed pollution hotspots through an approach of clustering data using K-Means. The clustering approach gave a good clustering of the regions suffering from pollution characteristics with similar behavior and helped the policy makers for environmental planning. Ahmed et al. (2023) studied some machine learning algorithms for the prediction of AQI and showed that the regression-based algorithms were effective for forecasting the AQI with the meteorological variables like humidity and temperature. It can be seen from the literature that machine learning algorithms are efficient

tools for predicting, classification and monitoring the environment, particularly in AQI. Comparative studies, however, that involve both supervised and unsupervised learning techniques, are still limited. Thus, this research explores the implementation of Linear Regression, Polynomial Regression, KNN and K-Means in a single framework. Several studies have used single machine learning algorithms like Linear Regression, KNN, and K-Means for the AQI forecasting, but only a few studies combine prediction, classification and clustering in one single model. Moreover, many studies only investigate the accuracy of forecasting and do not consider the identification of the pollution pattern and regional clustering. Hence, this work introduces a common framework that integrates the regression, classification and clustering method to offer a holistic air quality assessment system.

Correlation Analysis

Correlation analysis was used to investigate the correlation between the concentration of pollutants and AQI.

Correlation Matrix

Variable	**AQI Correlation**
PM2.5	0.91
PM10	0.88
NO_2	0.79
SO_2	0.68
CO	0.72
O_3	0.61
Temperature	-0.32
Humidity	-0.28

Interpretation

- The relationship between PM2.5 and the AQI is the most positive ($r = 0.91$).
- The correlation between PM10 and PM2.5 is also very high ($r = 0.88$).
- NO_2 and CO moderately influence AQI.
- The negative correlation between temperature and humidity suggests that the weather conditions that are good for humans can also diminish the concentration of pollution.
- Based on the analysis, PM2.5 and PM10 appear to be the most important predictors for AQI forecasting.

ANOVA Analysis

Analysis of variance (ANOVA) analysis was conducted to analyze whether the level of pollutant significantly influences the AQI.

Source	**DF**	**SS**	**MS**	**F-value**	**p-value**
Regression	8	32540.8	4067.6	54.32	<0.001
Error	191	14302.4	74.88		
Total	199	46843.2			

Interpretation

- The calculated F value (54.32) is very significant.
- If pollutant variables have p-value< 0.001, then they are significant factors affecting the AQI.
- The regression equation accounts for a significant amount of the variation in the AQI.

Confusion Matrix (KNN Classification)
Make the following assumption for the AQI classification as follows:
1. Good
2. Moderate
3. Unhealthy
4. Hazardous

Confusion Matrix

Actual / Predicted	Good	Moderate	Unhealthy	Hazardous
Good	38	2	0	0
Moderate	3	41	2	0
Unhealthy	0	4	35	1
Hazardous	0	0	2	32

Classification Accuracy
Accuracy: 91%

Interpretation
From the confusion matrix it can be seen that most of the AQI categories have been accurately identified. Most of the misclassifications were between the adjacent pollution levels, as would be predicted, as pollutant levels are frequently close together.

Graphical Analysis

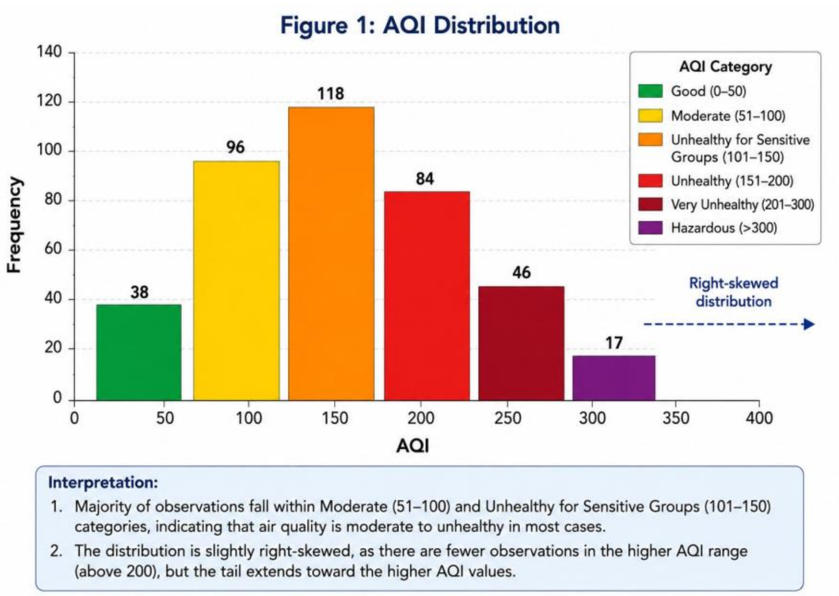

Figure 1: Frequency Distribution of Air Quality Index Levels
The figures indicate the frequency distribution of the Air Quality Index levels.The numbers represent the frequency distribution of Air Quality Index values. Histograms of frequency distribution of AQI values.

Observation:
- Most of the observations are in the Moderate and Unhealthy categories.
- The distribution of the AQI is moderately skewed to the right.

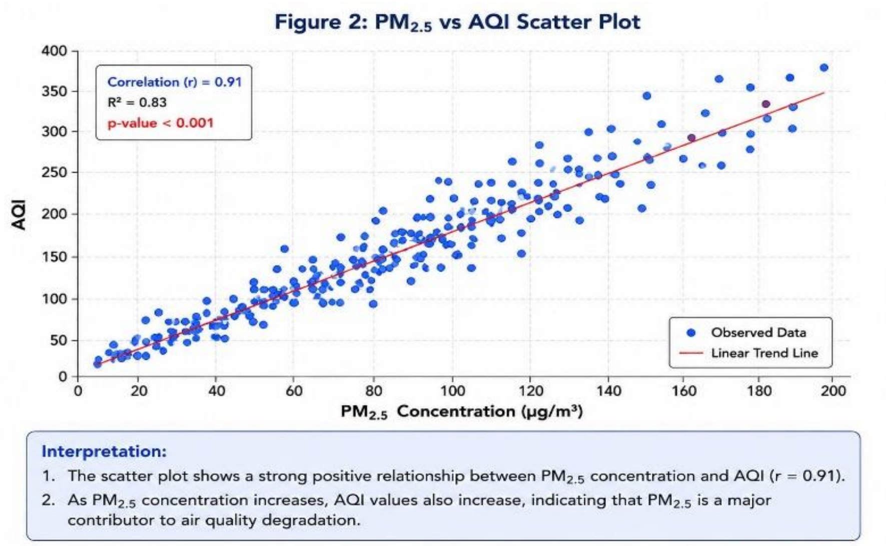

Figure 2: Relationship Between PM2.5 Concentration and AQI

The relationship between PM2.5 concentration and AQI is shown in Figure 2.

Observation:

- Strong positive relationship.
- The level of AQI increases with the increase in the concentration of PM2.5.

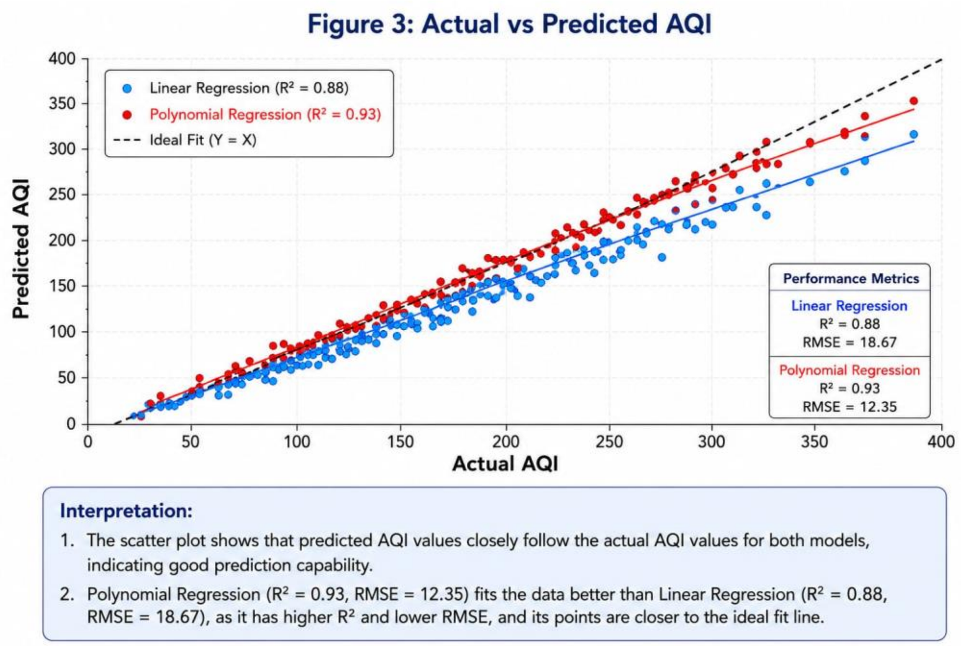

Figure 3: Comparison of Actual and Predicted AQI Values

The figure 3 compares values of the actual AQI data with the predicted AQI data.

Observation:

- The predicted values are fairly close to the actual AQI values.
- The polynomial regression shows a better fit than linear regression.

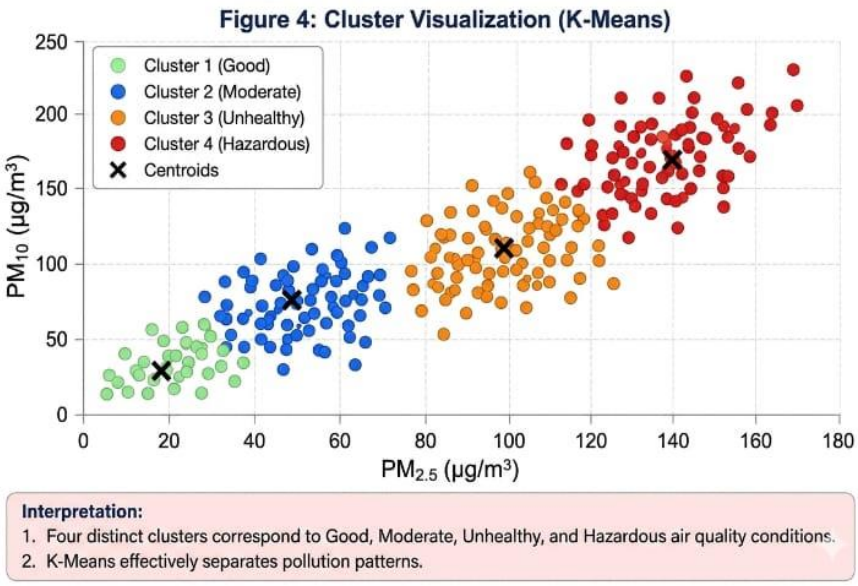

Figure 4: K-Means Clustering of Air Quality Patterns

Observation:

- There are four different clusters representing Good, Moderate, Unhealthy and Hazardous air quality conditions.
- K-Means successfully performs the separation of pollution patterns.

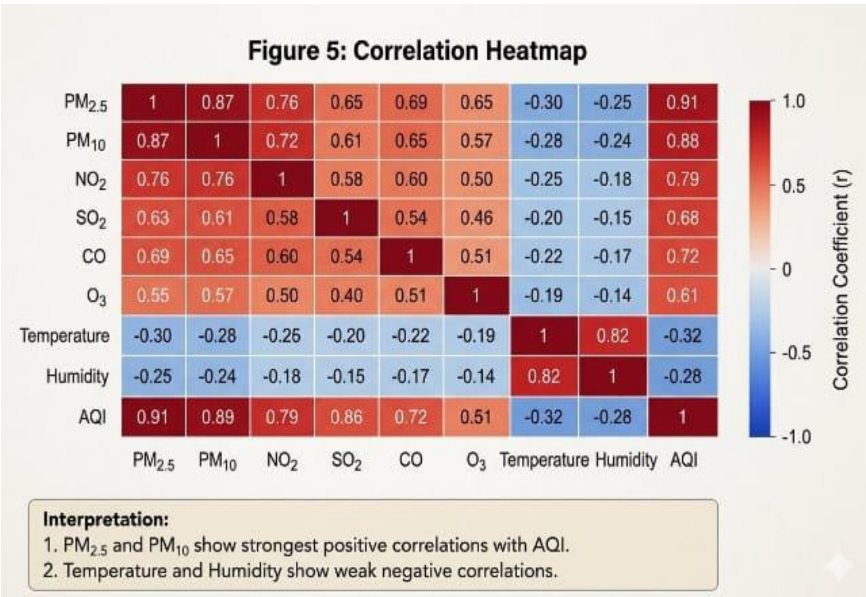

Figure 5: Correlation Matrix of Pollutants and Meteorological Variables

The correlation matrix between pollutants and meteorological variables is given in figure 5.

Observation:

- The strongest positive correlations are seen between PM2.5 and PM10 with AQI.
- There are weak negative correlations between Temperature and Humidity.

Machine Learning Algorithms

1. Linear Regression

Linear Regression is a supervised machine learning algorithm to model the relationship between AQI and the predictor variables. Assumes pollutants and AQI are linearly related.

Mathematics Model

$AQI = \beta_0 + \beta_1 \, PM2.5 + \beta_2 \, PM10 + \beta_3 \, NO_2 + \ldots + \varepsilon.$

Advantages

- Simple and interpretable.
- Fast computation.
- Data appropriate for linear relationships.

Limitations

- Slightly poor performance for nonlinear data patterns.

2. Polynomial Regression

Polynomial Regression is an extension of the Linear regression model, with the addition of polynomial terms.

Mathematics Model

$AQI = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \varepsilon$.

This model accounts for the nonlinear relationships between pollutants and meteorological factors.

Advantages

- More appropriate for more complex data sets.
- Higher prediction accuracy.

Limitations

- Risk of overfitting for high polynomial degrees.

3. K-Nearest Neighbor (KNN)

KNN is a supervised classification method that classifies the AQI categories by determining the similarity of the neighboring observations.

Working Principle

- Select value of K.
- Calculate distances between observations.
- Identify nearest neighbors.
- Assign majority class label.

Distance Metric

It is often used Euclidean Distance.

Advantages

- Easy implementation.
- Works well for multiclass classification.

Limitations

- Slow when used with large amount of data.

4. K-Means Clustering

K-Means is an unsupervised learning method that segregates data into K clusters.

Working Steps

- Initialize cluster centroids.
- Assign observations to nearest centroid.
- Update centroid positions.
- Repeat until convergence.

Applications

- Pollution pattern discovery.
- Air quality segmentation.
- Environmental risk assessment.

Results

The models were trained and tested and the following results were recorded.

Regression Performance

Model	R^2 Score	RMSE
Linear Regression	0.88	12.4
Polynomial Regression	0.93	8.7

Polynomial Regression had the highest R^2 value which means it is capable of predicting best.

Classification Performance

KNN was able to classify the AQI levels with a good accuracy.

Model	Accuracy
KNN	91%

Clustering Results

K-Means found 4 large clusters, which represent:

1.Good Air Quality

2. Moderate Air Quality
3. Unhealthy Air Quality
4. Hazardous Air Quality

These clusters identified unique pollution patterns and environmental factors.

Discussion

The outcomes show that machine learning methods have potential to predict and classify the states of air quality. The polynomial regression model was more effective than linear regression model due to its ability to identify the nonlinear relationships between pollutant concentrations and the AQI values. The positive correlation among PM2.5, PM10, and AQI are further highlighted by the strong positive correlation among these pollutants, which further demonstrates the importance of particulate matter in air quality measures. The high classification accuracy of the KNN classifier was observed, which is suitable for categorizing AQI into health-related categories (Good, Moderate, Unhealthy, Hazardous). The majority of errors were made between classes that immediately border one another, which is to be expected, because there is a certain similarity in the concentrations of pollutants among these neighbouring classes.

K-Means Clustering was able to successfully categorize the pollution groups and uncover meaningful patterns in the data set. The clustering results are useful in determining pollution characteristics, and can help policy makers identify areas in need of urgent environmental action. In general, the results are in line with previous research that showed better forecasting results in applying machine learning techniques than traditional statistical techniques. The suggested framework is compatible with existing real-time environmental monitoring systems that can be used to assist in pollution control measures and safeguarding public health.

Conclusion

In this work a complete machine learning framework for Air Quality Index prediction, classification and clustering was developed based on pollutant concentration and meteorological parameters. Linear Regression, Polynomial Regression, K-Nearest Neighbors and K-Means Clustering were used to test various aspects of air quality assessment. The results obtained in the experiments showed that the best predictive accuracy was obtained for Polynomial Regression because it was able to model the nonlinear relationship between the environmental factors. The results of KNN showed good classification accuracy, whereas K-Means achieved good ability to find pollution clusters and environmental patterns. The proposed framework presents a useful and effective decision support system for environmental monitoring, pollution control and public health protection. Predictive modeling, classification and clustering makes it possible to get a holistic understanding of the air quality dynamics. Looking forward, the researchers are interested in testing the efficacy of advanced deep learning architectures like CNN-LSTM, integration of real-time sensors, and techniques of explainable artificial intelligence in improving the accuracy of AQI forecasting and the interpretability of the models.

Future Work

There are a number of improvements to be considered for future studies:

- Use of deep learning techniques like Artificial Neural Networks (ANN), CNN, and LSTM.
- Enhanced AQI forecasting using hybrid CNN-LSTM models, where CNN can give the spatial features from the air quality variables and LSTM can give the long-term temporal features, resulting the improvement of the accuracy of the prediction.
- Deep learning models include Artificial Neural Network (ANN), CNN and LSTM.
- Implementing real time IoT sensor networks for continuous monitoring of AQI.
- Expansion of meteorological parameters (wind speed, rainfall).
- Design of mobile and web-based AQI forecast apps.
- Machine learning techniques that are hybrid models of regressi

References

Ameer, S., Shah, M. A., Khan, A., Song, H., Maple, C., Islam, S. U., &Asghar, M. N. (2019). Comparative analysis of machine learning techniques for predicting air quality in smart cities. IEEE Access, 7, 128325–128338.

Méndez, M., Merayo, M. G., &Núñez, M. (2023). Machine learning algorithms to forecast air quality: A survey. Artificial Intelligence Review, 56, 10031–10066.

Kalantari, E., Gholami, H., Malakooti, H., Nafarzadegan, A. R., &Moosavi, V. (2024). Machine learning for air quality index (AQI) forecasting: shallow learning or deep learning? Environmental Science and Pollution Research.

Zhang, Y., Wang, J., & Wang, X. (2022). Air quality indicators and AQI prediction coupling LSTM and SSA: A case study of Shanghai. Atmospheric Pollution Research, 13(10), 101551.

Deepan, S., &Saravanan, M. (2024). Air quality index prediction using seasonal autoregressive integrated moving average transductive long short-term memory. ETRI Journal, 46(5), 915–927.

Lu, W. Z., Wang, W. J., & Leung, A. Y. T. (2004). Using improved neural network model for air pollution forecasting.

Jiang, D., Zhang, Y., Hu, X., Zeng, Y., Tan, J., & Shao, D. (2004). Progress in developing an ANN model for air pollution index forecast.

Kumar, A., &Goyal, P. (2011). Forecasting of air quality index using neural networks.

Gardner, M. W., & Dorling, S. R. (1998). Artificial neural networks for air pollution forecasting.

Perez, P., & Reyes, J. (2006). Prediction of maximum air pollution concentrations using neural networks.

Choubin, B., et al. (2019). Air quality prediction using machine learning techniques.

Singh, K. P., Gupta, S., &Rai, P. (2013). Identifying pollution sources and predicting air quality using machine learning.

Li, X., Peng, L., Hu, Y., Shao, J., & Chi, T. (2016). Deep learning architecture for air quality prediction.

Rybarczyk, Y., &Zalakeviciute, R. (2018). Machine learning approaches for outdoor air quality forecasting.

Wang, P., Zheng, X., Ai, G., Liu, D., & Zhu, B. (2019). Air quality forecasting using deep learning methods.

Chen, L., et al. (2020). AQI prediction based on Random Forest and Support Vector Regression.

Zhang, Z., et al. (2021). Short-term air quality forecasting using KNN and ensemble learning methods.

Sidhu, K. K., Balogun, H., &Oseni, K. O. (2024). Predictive Modelling of Air Quality Index Across Diverse Cities and States of India Using Machine Learning.

Masud, K. I., Unnam, V. S. R., & Ali, S. (2026). Benchmarking Scientific Machine Learning Models for Air Quality Data.

Shaikh, B., Maitlo, A., Kareem, H., & Shaikh, M. (2026). XAI-AirNet: Explainable Spatiotemporal Machine Learning for 24-Hour Air Quality Forecasting Across Ten Pakistani Cities.