**Int. J. Sci. R. Tech., 2026 3(5)**
A Multidisciplinary peer-reviewed Journal
www.ijsrtjournal.com [ISSN: 2394-7063]

*Original Article*

# Air Quality Index Prediction By Using Machine Learning

**Sathya U., Monika A. G., Sadhana C., Thamini S., Kanimozhi K.***

*Vivekanandha College of Technology for Women, India.*

## ABSTRACT

One of the biggest environmental problems the world is currently facing is pollution. In cities with rapid industrial development and growing population, pollution is becoming increasingly difficult to control. Pollution from transportation, factories, and buildings has polluted the air to such an extent that it not only poses a threat to the environment but also to people's health. , it is vital to predict the air quality indices to take appropriate measures for tackling this problem. Air Quality Index (AQI) is defined as a measure that determines the amount of primary air pollutants in the air. These pollutants include particulate matters PM2.5 and PM10, as well as carbon monoxide, Sulphur dioxide, nitrogen dioxide, and ozone. Conventional AQI prediction systems rely heavily on the sensors to forecast the upcoming AQI based on manual analysis. But machine learning algorithms offer a great solution to this problem due to their capability of making accurate predictions using data sets.

To enhance data quality and model performance, the procedure includes feature selection, preprocessing, and data collecting. Performance measures like Mea n AbsoluteError (MAE) and Root Mean Square Erro r(RMSE) are used to apply and assess a number of methods, including Linear Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), an Gradient Boosting.

**Keywords**: Air Quality Index (AQI), Machine Learning, Air Pollution Prediction, Random Forest, Support Vector Machine, Environmental Monitoring, IoT, Data Analysis.

## INTRODUCTION

Globally, air pollution has become a serious environmental and public health issue, especially in areas that are quickly industrializing and urbanizing. Air quality has significantly declined as a result of ongoing population growth, increased mobility, industry expansion, and rising energy use. Higher concentrations of dangerous chemicals are found in cities, endangering human health in addition to the environment. Long-term exposure to contaminated air can lead to lung cancer, asthma, respiratory infections, heart problems, and even early death. As a result, governments, scientists, and environmental organizations now consider maintaining high air quality to be crucial. The Air Quality Index (AQI) is frequently used to comprehend and convey the degree of air pollution. AQI is a numerical scale that transforms complicated data on air pollution into a manner that the general public can easily comprehend. Major pollutants include particle matter (PM2.5 and PM10), nitrogen dioxide ($NO_2$), Sulphur dioxide ($SO_2$), carbon monoxide (CO), and ozone ($O_3$) are taken into account. Air quality is classified into several levels, including acceptable, satisfactory, moderate, poor, extremely bad, and dangerous, based on the AQI number. This classification aids people and authorities in taking the appropriate safety measures to reduce exposure to hazardous air conditions.

Machine learning models use past data on air contaminants and meteorological factors including temperature, humidity, wind speed, and rainfall to estimate AQI. The dispersion and concentration of pollutants in the atmosphere are significantly influenced by these factors. ML models may accurately predict future AQI levels by examining these inputs. This aids in giving authorities and people early notice of possible air pollution incidents so they can take preventative measures.

## LITERATURE REVIEW

A method for predicting air quality using machine learning algorithms was presented by Li et al. [1]. In order to forecast the Air Quality Index (AQI) using

**Relevant conflicts of interest/financial disclosures**: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

historical pollutant data, including PM2.5, PM10, $NO_2$, $SO_2$, and CO, their study examined several models, including Linear Regression, Random Forest, and Artificial Neural Networks (ANN). The findings demonstrated that ensemble and neural network models increased prediction accuracy; nevertheless, limited feature engineering had an impact on performance in highly contaminated areas.

A deep learning-based method for predicting air quality using Long Short-Term Memory (LSTM) networks was presented by Zhang et al. [2]. The algorithm was created to forecast short-term AQI by analyzing time-series data on air pollution. Although their approach required a lot of training data and had a significant computing cost, it was successful in capturing temporal correlations with good accuracy.

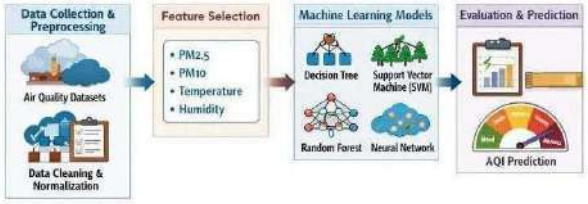
For AQI prediction, Kumar et al. [3] created a hybrid model that used Random Forest and LSTM approaches. This method employed LSTM for temporal prediction and Random Forest for feature selection. Although the hybrid approach increased system complexity and was challenging to implement in low-resource contexts, it did enhance overall prediction performance.

A multi-pollutant-based AQI prediction model utilizing XGBoost was presented by Singh et al. [4]. To improve prediction accuracy, their study included both contaminant and climatic factors. Although the model worked well, it had a propensity to overfit and needed careful hyperparameter tuning to achieve the best outcomes.

## ALGORITHMS USED AND EXPERIMENTAL SETUP

To guarantee accurate and trustworthy results, machine learning-based Air Quality Index (AQI) prediction necessitates both the choice of appropriate algorithms and a clearly defined testing procedure. By examining past air pollution data and environmental factors, the prediction model is constructed, allowing the system to identify trends and predict future AQI values. The algorithms utilized and the experimental setup for model creation and evaluation are described in this section. Because machine learning algorithms can manage massive datasets and recognize intricate correlations between factors, they are essential to AQI prediction. One of the most straightforward prediction algorithms among the different methods is linear regression. It creates a connection between the dependent variable, AQI, and independent variables like pollution concentrations and meteorological conditions. Its efficiency is constrained when the data contains nonlinear patterns, despite its ease of implementation and interpretation. Decision Tree techniques are employed to get around this restriction. A Decision Tree uses conditions applied to input features to split the dataset into smaller subsets. The output is produced at the leaf nodes, and each decision node represents a test on a particular property. This method can simulate nonlinear interactions and is helpful for comprehending the decision-making process. However, if the tree is too complicated, it could experience overfitting. Data collection is the first phase in the experimental setting and is essential to the development of an AQI prediction system. Data is collected from trustworthy sources like IoT-based devices, open-source platforms, and government air quality monitoring stations. In addition to meteorological factors like temperature, humidity, wind speed, and rainfall, the dataset usually contains pollution concentrations like PM2.5, PM10, CO, $SO_2$, $NO_2$, and $O_3$. These factors are important in influencing the quality of the air. Preprocessing is done to enhance the quality of the data once it has been gathered. Model performance may be impacted by missing values, noise, and inconsistencies found in real-world datasets. Methods like interpolation and mean replacement are used to deal with missing values. Duplicates and inaccurate entries are eliminated by data cleansing. To ensure that every feature contributes equally to the model, normalization is used to scale the data within a predetermined range. Another crucial stage is featuring selection, which reduces complexity and increases efficiency by selecting only the most pertinent variables. In conclusion, creating an accurate AQI prediction system requires a combination of several machine learning algorithms and a carefully planned experimental setting. Because they can manage complex relationships and minimize errors, algorithms like Random Forest and Gradient Boosting typically perform better. The system's dependability is guaranteed by the experimental procedure, which includes data preprocessing, model training, and evaluation. This method makes it possible to monitor and forecast air quality

effectively, which improves environmental management and safeguards public health.




Experimental Setup for Air Quality Index Prediction Using Machine Learning

## PROPOSED SYSTEM

The goal of the suggested method is to employ machine learning techniques to create an effective and trustworthy model for forecasting the Air Quality Index (AQI). This system's primary goal is to estimate future air quality levels by analyzing past data on air pollution and meteorological elements. The technology seeks to support environmental monitoring, aid authorities in making decisions, and assist people in taking health- protective precautions by offering precise forecasts.

From data collection to AQI prediction, the system is built as a data-driven model that adheres to a disciplined workflow. To improve accuracy and performance, it combines machine learning algorithms with data pretreatment and evaluation methods. The suggested method guarantees that the model can effectively process real-world data and offer significant insights into patterns in air quality. Data collection is the initial part of the suggested system. The system collects information from dependable sources, including IoT-based sensors, environmental databases, and government air quality monitoring stations. Important pollutants including PM2.5, PM10, carbon monoxide (CO), sulfuric dioxide ($SO_2$), nitrogen dioxide ($NO_2$), and ozone ($O_3$) are included in the dataset. Meteorological variables like temperature, humidity, wind speed, and rainfall are gathered in addition to pollution data. The concentration and dispersion of contaminants in the atmosphere are significantly influenced by these environmental factors. Data preparation is the next stage after data collection and is crucial to raising the dataset's quality. Missing values, noise, and inconsistencies are common in real-world data and can have a detrimental effect on model performance. Techniques like mean imputation and interpolation are used to deal with missing values. Duplicate entries and inaccurate values are eliminated by data cleaning. In order to ensure that every feature contributes equally during model training, normalization is used to scale the data within a consistent range. In order to reduce complexity and increase prediction accuracy, feature selection is also used to find the most pertinent variables.

The dataset is split into training and testing sets following preprocessing. The machine learning models are trained using the training dataset, and their performance is assessed using the testing dataset. The model's ability to effectively generalize to fresh, untested data is ensured by this division. The suggested system makes use of several machine learning methods, including Support Vector Machine (SVM), Random Forest, Decision Trees, Linear Regression, and Gradient Boosting. To assess each algorithm's performance and determine which model is best for AQI prediction, all algorithms are trained on the same dataset.


## METHODOLOGY

The process of utilizing machine learning to forecast the Air Quality Index (AQI) entails a methodical and organized technique that converts unprocessed environmental data into insightful forecasts. Data gathering, preprocessing, feature engineering, model selection, training, evaluation, and deployment are some of the steps in this process. Every step is essential to guaranteeing the prediction system's precision and dependability.

Data collection is the methodology's first stage. Building a successful machine learning model requires accurate and pertinent data. The concentrations of key air pollutants, including particulate matter (PM2.5 and PM10), carbon

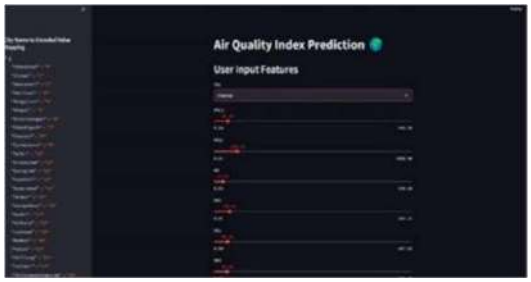
monoxide (CO), sulfur dioxide ($SO_2$), nitrogen dioxide ($NO_2$), and ozone ($O_3$), are usually included in the dataset used for AQI prediction. Meteorological parameters including temperature, humidity, wind speed, and rainfall are also gathered in addition to pollution data since they affect how pollutants disperse and build up in the atmosphere. IoT-based environmental sensors, internet databases like Kaggle, and government organizations like the Central Pollution Control Board (CPCB) can all provide data.

Data preprocessing, which comes after data collection, is crucial to raising the dataset's quality. Inconsistencies, noise, and missing values are common in real-world data. Methods like mean substitution, median filling, or interpolation can be used to deal with missing numbers. Filtering and smoothing techniques are used to reduce noise in the data. To guarantee data consistency, duplicate records and inaccurate entries are found and eliminated. The data is then scaled into a uniform range via normalization or standardization, which improves the performance of machine learning algorithms. The model's accuracy could be greatly impacted by improper preprocessing.

Another crucial phase in the process is featuring engineering. It entails choosing and modifying pertinent characteristics that support AQI forecasting. Feature selection methods like correlation analysis and relevance ranking are used to find the most important qualities because not every variable in the dataset may be valuable. For instance, AQI levels are frequently significantly impacted by pollutants such as PM2.5 and PM10. Certain features, such calculating pollutant ratios or averaging values across time, can be combined to create new ones. The model's capacity to identify significant patterns in the data is enhanced by this procedure.

## RESULT AND DISCUSSION

The outputs of the machine learning models created to forecast the Air Quality Index (AQI) are shown in the results and discussion section, along with a thorough evaluation of their effectiveness. This section's main goal is to assess how well various algorithms can forecast AQI values using meteorological parameters and historical air pollution data. The models' comparative performance, strengths, and limitations are also highlighted in the debate.




Following data preprocessing and model training, the testing dataset was used to assess a number of machine learning techniques, including Linear Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), and Gradient Boosting. Standard assessment criteria such as Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R-squared ($R^2$) were used to assess each model's performance. These measurements aid in comprehending the precision and dependability of the forecasts.


A baseline for comparison was provided by the Linear Regression model. When there was a roughly linear relationship between input attributes and AQI values, it performed fairly well. However, the accuracy of Linear Regression was limited since air pollution data is influenced by numerous nonlinear factors. When compared to more sophisticated algorithms, the model generated larger error values, suggesting that it might not be appropriate for challenging AQI prediction tasks.


Because the Decision Tree model can capture nonlinear interactions between variables, it outperformed Linear Regression. It was able to successfully manage changes in pollution levels and produced predictions that were more accurate. Nevertheless, it was found that the Decision Tree tended to overfit the training set, particularly when the tree's depth was high. As a result, performance on the testing dataset decreased, underscoring the necessity of methods like pruning to enhance generalization. Both the Decision Tree and Linear Regression models were significantly outperformed by Random Forest. Random Forest decreased overfitting and increased prediction accuracy by merging several decision trees. Better performance was shown by the model's lower MAE and RMSE values. Additionally, it was more adept at handling noisy data and missing values. Random Forest was one of the models that performed the best in this investigation because of its resilience and dependability. Additionally, Support Vector Machine (SVM) performed well, especially when nonlinear kernel functions were employed. The model produced precise predictions and was able to identify intricate patterns in the data. However, parameter tuning—including the selection of kernel and regularization parameters—had a significant impact on SVM performance. Of all the models examined, gradient boosting—including sophisticated implementations like XGBoost—produced the best accuracy. The model is gradually improved by this process, which also fixes mistakes from earlier iterations. It thus obtained the greatest R2 score and the lowest MAE and RMSE values. Gradient Boosting is well suited for AQI prediction since it successfully managed the dataset's complicated linkages and variations. Nevertheless, the model needed more training time and careful adjustment. When the models are compared, it is evident that ensemble techniques like Random Forest and Gradient Boosting perform better than individual models. These techniques are better at managing intricate datasets and lowering errors. The findings demonstrate that while more sophisticated models offer greater forecast accuracy and dependability, simpler models are simpler to execute. The models' performance was visualized using graphical analysis in addition to numerical evaluation. Simpler models deviated more from the actual trends, while ensemble models closely matched them, according to line graphs comparing real and anticipated AQI values. Additionally, scatter plots showed that Random Forest and Gradient Boosting predictions were more accurate because they were nearer the ideal line. Error distribution graphs showed that while greater errors were more common in simpler models, the majority of prediction errors for ensemble models were modest. The significance of feature selection and data pretreatment was further emphasized by the analysis. Model performance was greatly enhanced by normalization and appropriate handling of missing variables. It was discovered that characteristics like PM2.5, PM10, and $NO_2$ had a significant impact on AQI results. Prediction accuracy was further improved by incorporating meteorological characteristics like temperature and humidity, which have an impact on pollutant dispersal. The impact of dataset quality and size on model performance is another significant finding from the results. Predictions made by models trained on bigger, cleaner datasets were more accurate. On the other hand, datasets with inconsistent or missing values had more mistakes. This highlights the necessity of trustworthy data gathering and preprocessing methods in AQI prediction systems. The results' practical ramifications are also discussed. Government agencies can employ accurate AQI prediction models to carry out preventive actions including limiting industrial emissions, managing transportation congestion, and issuing health advisories. For instance, officials can take preventative measures to minimize pollution sources if the model forecasts a high AQI level for the following day. Similarly, individuals can utilize this knowledge to organize their activities and prevent exposure to dangerous air conditions.

## CONCLUSION

In recent years, air pollution has emerged as one of the most significant environmental issues, impacting both ecological balance and human health. The deterioration in air quality has been greatly exacerbated by rapid urbanization, industrialization, and rising vehicle emissions. In this regard, forecasting the Air Quality Index (AQI) is crucial for efficient environmental monitoring and for implementing preventative actions to lessen the detrimental consequences of pollution. In contrast to conventional techniques, this work concentrated on creating a machine learning-based method for AQI prediction that is dependable and effective. This work's primary goal was to evaluate air pollution data and develop predictive algorithms capable of accurately estimating AQI values. In addition to climatic variables like temperature, humidity, and wind speed, the study used historical data on important pollutants like PM2.5, PM10, carbon monoxide (CO), sulfur dioxide ($SO_2$), nitrogen dioxide ($NO_2$), and ozone ($O_3$). Machine learning models were able to find trends and connections that affect air quality by combining these variables. A potent and effective method for dealing with air pollution issues is the application of machine learning for AQI prediction. The study demonstrates how sophisticated computers may produce precise and timely forecasts. It is feasible to create intelligent and scalable air quality management solutions by combining data processing methods, machine learning models, and real-time data collection devices. Larger datasets, integration with IoT devices, and the application of deep learning techniques can all be used in the future to increase real-time monitoring and forecast accuracy. All things considered, this strategy greatly contributes to the advancement of sustainable development and the creation of healthy living conditions.

## REFERENCES


1. J. Wang, X. Li, L. Jin, J. Li, Q. Sun, and H. Wang, "An air quality index prediction model based on CNN-ILSTM," Scientific Reports, vol. 12, no. 8373, pp. 1–12, 2022.
2. H. Wu, T. Yang, H. Li, and Z. Zhou, "Air quality prediction model based on mRMR–RF feature selection and ISSA–LSTM," Scientific Reports, vol. 13, no. 12825, pp. 1–15, 2023.
3. N. S. Gupta, Y. Mohta, K. Heda, R. Armaan, B. Valarmathi, and G. Arulkumaran, "Prediction of Air Quality Index Using Machine Learning Techniques: A Comparative Analysis," Journal of Environmental and Public Health, vol. 2023, pp. 1–26, 2023.
4. D. J. S. Naidu and R. Aruna, "Study of Air Quality Detection using Machine Learning Techniques," International Journal of Scientific and Applied Research, vol. 2, no. 8, pp. 1–6, 2022.
5. M. Londhe, "Data Mining and Machine Learning Approach for Air Quality Index Prediction," International Journal of Engineering and Applied Physics, vol. 2, no. 1, pp. 1–10, 2021.
6. G. V. P. S. Sruthi, M. Lokesh, A. Grace, S. L. Reddy, and B. Srinivas Raja, "A novel approach for air quality prediction using machine learning," Journal of Multidisciplinary Research, vol. 4, no. 2, pp. 1–8, 2021.
7. L. Ramesh and S. Gopinathan, "Prediction of air pollution and an air quality index using machine learning techniques," International Journal of Advanced Research in Computer Science, vol. 11, no. 3, pp. 1–7, 2022.
8. S. Bhattacharya and S. Shahnawaz, "Using Machine Learning to Predict Air Quality Index in New Delhi," arXiv preprint arXiv:2112.05753, 2021.
9. K. K. Sidhu, H. Balogun, and K. O. Oseni, "Predictive modelling of Air Quality Index across diverse cities using machine learning," arXiv preprint arXiv:2404.08702, 2024.
10. A. T. Nguyen, D. H. Pham, B. L. Oo, Y. Ahn, and B. T. H.Lim, "Predicting air quality index using hybrid deep learning models," Journal of Big Data, vol. 11, no. 71, pp. 1–20, 2024.
11. IEEE Conference Publication, "Air Quality Prediction Using Machine Learning: A Comparative Study," IEEE Xplore, 2024.


HOW TO CITE: Sathya U., Monika A. G., Sadhana C., Thamini S., Kanimozhi K.*, Air Quality Index Prediction By Using Machine Learning, Int. J. Sci. R. Tech., 2026, 3 (5), 271-276. https://doi.org/10.5281/zenodo.20062391