OPEN ACCESS JOURNALS
**Indiana Journal of Multidisciplinary Research**
*Abbriviate Tittle- Ind J Multi Res*
ISSN *(Online)*- 2583-3820
Journal Homepage Link- https://indianapublications.com/Journals/IJMR
DOI: https://doi.org/10.5281/zenodo.12685945

INDIANA PUBLICATIONS
PRODUCTIVE AND QUALITY RESEARCH

**Research Article**

Volume-04|Issue-03|2024

# Predictive Modelling of Air Quality Index for Environmental Forecasting: Integrating Environmental Data Analysis

**Prof. Samatha R Swamy[1], S Harini*[2], P Khyathi Reddy[3]**

[1]Assistant Professor, Information Science & Engineering Department, RV Institute of Technology and Management, Bengaluru, Karnataka, India.
[2,3]Student in Information Science & Engineering Department, RV Institute of Technology and Management, Bengaluru, Karnataka, India.

**Article History**
Received: 20.05.2024
Accepted: 05.06.2024
Published: 30.06.2024

**Citation**
Swamy, S. R., Harini, S., & Reddy, P. K. (2024) Predictive Modelling of Air Quality Index for Environmental Forecasting: Integrating Environmental Data Analysis. *Indiana Journal of Multidisciplinary Research*, *4*(3), 272-277.

**Abstract:** The Central and State Air Pollution Monitoring Program (NAMP) in India covers 240 cities with 342 monitoring stations. XGBoost was chosen as a cost-effective AQI estimator to gather data from comparison sites. Accurately estimating the air quality index (AQI) is crucial for environmental monitoring and management. Previous studies have overlooked the significance of estimating uncertainties and limiting production during forecasting. To address this, we introduce a new hybrid model, TMSSICX, to predict AQI in various cities. First, we utilized time-varying filtering-based empirical mode decomposition (TVFEMD) to decompose the AQI sequence into its multifunction model (IMF) components. Second, we employed multiscale fuzzy entropy (MFE) to assess the complexity of each IMF component and categorize them into high- and low-frequency domains. Furthermore, to mitigate volatility, we decomposed the frequency component twice using a continuously variable vector (SVMD). Finally, we input six air pollutants (e.g., CO, SO2, PM2.5, PM10, O3, and NO2) into the model.
**Keywords:** Air pollution, Air Quality Index (AQI), Machine learning, XGBoost.



## INTRODUCTION

The environment is greatly affected by air pollution both the environment and human health, resulting in six categories of air quality levels: Good (safe), Fair, Sensitive, Poor, not very good, and Dangerous. Human activities such as business, food production, transportation, and consumption contribute to air pollution. As countries around the world continue to incorporate information and communication technologies and IoT systems to stimulate economic growth, the development of society and infrastructure has also driven the growth of IoT. To protect people and the environment from pollution, all countries strictly control air pollution. The prevalence of respiratory diseases in urban areas is primarily attributed to poor air quality, leading to a significant number of fatalities. Developing countries, in particular, witness 70% of emissions originating from outdated and poorly maintained vehicles, as well as substandard fuel, as reported by the World Health Organization.

With the rapid development and expansion of urban populations, many environmental problems have gained importance. Air pollution negatively affects human health, and air quality sensors have been effective in reducing its impact. The Air Quality Index is a tool used for risk communication and informs the public about the air quality in their surroundings and its health effects, especially for vulnerable populations such as children, the elderly, and those with heart disease or shortness of breath. The Air Quality Index converts the worst air quality data into numbers, text, and values, and is measured on a scale from 0 to 500. The Air Quality Index (AQI) is divided into six categories that focus on the quality of the atmosphere. Various activities such as transportation, industrial processes, and residential heating contribute to air quality, either directly or indirectly. Machine learning-based techniques are employed to recognize patterns based on historical data. For instance, D. Iskandaryan *et al.* [17] used one such method to analyze AQI scores of major cities in Poland. The researchers conducted three main simulations to evaluate forest performance, and the study found a link between air quality and improved air purity, which can also help improve the environment. They adopted a strategy to process climate data and used various machine learning-based models such as decision tree (DT), random forest (RF), and climate gradient boosting (XGBoost) to support AQI prediction. The XGBoost model achieved a remarkable accuracy of 0.03684 in this regard. To tackle this problem, two machine learning algorithms, namely neural network (NN) and support vector machine (SVM), were employed using data sourced from the Central Pollution Control Panel (CPCB). It utilized support vector regressor (SVR) and random forest regressor (RFR) for forecasting AQI and nitrogen dioxide (NO2) levels in Beijing. The SVR-based model

***Corresponding Author:** S Harini

predicted AQI values better than the RFR model with an R2 value of 0.9766 and an RMSE of 7.666. Pant *et al.* [20] developed an accurate prediction of air quality in Dehradun, Uttarakhand using supervised machine learning. The decision tree model had a high accuracy of 98.63%, while logistic regression had the highest probability of 91.78% for air quality. For example, Castelli *et al.* (2020) studied air pollution and small-scale problems using support vector regression (SVR) and radial basis function (RBF) and estimated air quality scores, proving that SVR gives the highest accuracy and reliability; accuracy is 94.1%. Gocheva-Iliev *et al.* [18] reported that a linear regression algorithm and gradient boosting algorithm were used to estimate AQI. The Naive Forecast method is also used, but the Gradient Boost algorithm is more accurate with 96% accuracy. The project helps find the main pollutants causing pollution. Campbell-Lendrum *et al.* [19] demonstrate the use of ML to estimate PM2.5 µm levels from air data in a high-rise non-urban city (Quito, Ecuador).

Autoregression is used to predict PM2.5 µm levels based on historical PM2.5 µm levels. The SVM algorithm and decision tree (DT) can predict PM2.5 µm with approximately 89% accuracy. Balgat *et al.* [4] collected several studies on weather forecasting using SVM, decision trees (DT), and algorithms that perform well on static data. Since static data is used, there is no precise information about where the air quality is, which will make a difference in the actual ground estimates. To solve this challenge, web scraping tools can be used to instantly collect data for training models, making it more accurate and providing better results. To determine the effectiveness of AQI estimation, Nigam *et al.* [22] proposed a method that combines weather data processing methods with machine learning to obtain better results. Three calculation models: DT, RF, and XGBoost were compared with MAE, RMSE, and R2 measurements to find the best model for AQI prediction. The method under consideration was evaluated using two distinct population datasets gathered from various regions of India. The XGBoost technique outperformed alternative AQI estimation methods.. Therefore, XGBoost was again selected as a low-cost method to estimate AQI for fixed-point measurement sites. Similarly, Janarthanan *et al.* [23] used two machine learning algorithms supporting vector machines and neural networks to predict AQI.

# LITERATURE SURVEY

Air quality is a pressing environmental issue that has a significant impact on public health. The Air Quality Index (AQI) is a standard measure of air pollution levels and their effects on health. An accurate estimate of AQI is crucial for environmental forecasting and enables people to take necessary precautions when the weather is poor. The current state of research on the prediction of air quality index is examined in this review, with a focus on machine learning and Deep Learning to predict pollution. Londhe, M *et al.* [1], additionally, meteorological conditions such as wind speed, temperature, and humidity add complexity to weather dynamics. Zhao, X *et al.* [2], Machine learning algorithms are now strong tools for forecasting AQI. They analyze patterns in past air quality data to forecast future AQI values. Support Vector Regression (SVR) is a popular technique that is effective in predicting AQI values, especially in cases of non-linear relationships between input parameters and AQI. Chowdhury, A. S *et al.* [3], Random Forest Regression is a technique that employs multiple decision trees, ensuring both solidity and precision in its results. This method is capable of handling large files and intricate relationships within them. Bhalgat, P *et al.* [4], K-Nearest neighbors (KNN) estimate AQI by examining historical data similar to the current situation. Its simplicity makes it easy to use, but the selection of k (the number of neighbors) is crucial to obtaining accurate results. Deep learning methods, a subset of machine learning, have become popular in predicting AQI because they can recognize complex patterns in big datasets. Notable techniques like LSTM networks excel at understanding time in detailed weather information. Adjustments can be made to match the type of weather prediction, improving the forecast's precision. Additionally, the Convolutional Neural Network (CNN) is effective in extracting spatial features from data. When combined with LSTM (CNN-LSTM model), they can leverage spatial and temporal data to achieve highly accurate AQI predictions, especially in different regions. Kottur, S. V *et al.* [6], Studies have looked at both comparative analysis and hybrid models to understand the pros and cons of different models. Some research has proven the efficiency of deep learning models such as CNN-LSTM for specific data and transmission sources. Castelli, M *et al.* [7], Combining various machine learning or deep learning algorithms in hybrid models can improve prediction accuracy by utilizing the distinct advantages of each method.

# MATERIALS AND METHODS

***Dataset Description:***

This dataset comprises a plethora of essential parameters for forecasting air quality, including the time of observation, year, month, day, hour, and PM2.5 levels. The time serves as a unique identifier for each observation and provides information about when the observation was made. Since year, month, day, and time do not offer useful physical data, it is necessary to search for seasonal and daily patterns in air pollution. PM2.5, which represents the concentration of suspended fine particles, is utilized as the target variable for prediction. XGBoost (a robust gradient boosting algorithm) can effectively model the relationship between time and PM2.5 concentration using this data. XGBoost's capacity to handle nonlinear relationships and process large data sets makes it the ideal choice for creating accurate and scalable climate models that have

***Corresponding Author:** S Harini

significant implications for environmental protection and public health management.

***Load the dataset:***

When stacking datasets from Kaggle to discuss quality, numerous variables ought to be taken under consideration, counting highlights such as time, year, month, day, hour, and PM2.5 concentration. Beyond any doubt, it is reasonable for examinations based on XGBoost. To begin with, getting information from Kaggle is required to explore the stage, distinguish important information utilizing look or look capacities, and select one based on an inquiry about goals. After the establishment record is downloaded, its structure and quality ought to be checked. Beginning steps incorporate dealing with lost values, guaranteeing protest consistency, and parsing the Timestamp protest into a date organized for time investigation. Once the information set is cleaned and organized, it can be stacked into an information examination environment such as Python utilizing pandas. Once the information is stacked, the information is run through the race to encourage assessment of the demonstration, which is vital for execution assessment. The prescient demonstration was prepared utilizing XGBoost, a capable angle-boosting calculation. Procedures such as framework look or arbitrary look can be utilized to tune hyperparameters to optimize the execution of the demonstration. At last, the learned XGBoost show was assessed on test information utilizing suitable parameters to confirm its adequacy in foreseeing PM2.5 concentration. The thorough preparation of extricating information from the assessment show leads to the best climate ponders utilizing XGBoost and Kaggle datasets.

***Data pre-processing:***

This record is pre-processed to guarantee that the dataset is legitimately organized and prepared to prepare the XGBoost demonstration. It covers progressed capacities such as data transformation, extraction of comes about, dealing with lost comes about, and determination of highlights, which shape the premise of great modelling. Since the awful climate observatory is found within the same zone as the climate observatory, the reasonable climate observatory is associated with the closest climate observatory. The input for the adjusted XGBoost demonstrates preparation is the day's climate perception information.

***Converting timestamp to datetime object:***

Timestamp columns are at first put away as a string or information sort. Change over this to a datetime protest utilizing the pd.to_datetime() work. This altar makes it simpler to check and evacuate physical highlights. After changing the Timestamp column you'll effortlessly evacuate custom objects like year, month, day and time utilizing the dt accessor given by pandas. This step breaks down the timestamp into its physical properties that can be utilized for investigation. After bringing in the time properties, utilize the drop() strategy to evacuate the Timestamp field from the dataset along the axis (axis=1). Since their data is now not excess, evacuating them will offer assistance to disentangle the dataset and decrease superfluous highlights. Lost values are common in real-world information and can influence the preparation. Here dropna() is utilized to erase lines with lost values. Depending on the information set and examination targets, other methods such as ascription may be utilized to bargain with lost information.

***Feature selection:***

The select_features variable will contain the list of highlights chosen to prepare the demonstration. Here, this step plans the dataset for the preparing show by isolating highlights and distinctive targets. The selected features variable will contain a list of highlights considered important for foreseeing PM2.5 concentrations. This choice preparation may be based on space information, factual investigation, or need criteria. Include choice can influence demonstration execution. Such unessential or rehashed highlights present clamor and disturb the calculation without making strides in exactness. On the other hand, overlooking critical highlights leads to one-sided or inadequate models, which can lead to destitute forecasts.

***Training and testing data:***

Evaluating the model's performance on hidden data during training is crucial. This helps measure the model's ability to generalize to new, unseen data. The train_test_split function in scikit-learn aids in this process by randomly dividing the dataset into a training set and a testing set.

**X_train, X_test, y_train, y_test**

The yield factors speak to the test network and target vector of the preparing set and test set, separately. X_train and X_test contain the highlights (preselected) utilized to prepare and test the show.

***Featured scaling:***

In many Machine learning algorithms, including XGBoost, the order of execution can impact the performance of the model. Include scaling points to bring all highlights to a comparable scale or extent, which lightens issues caused by the characteristics of distinctive sizes and units. Due to its measure, it guarantees that no single highlight overwhelms the learning handle, hence expanding the speed and steadiness of the calculation. StandardScaler could be a prevalent scaling strategy that standardizes highlights by evacuating the cruel and scaling to unit fluctuation. This change comes about in dissemination with cruel and standard deviation 1 for each include.

***Model Training:***

XGBoost (Extraordinary Slope Boosting) may be a learning method known for its viability in relapse

***Corresponding Author:** S Harini

and classification. Makes prescient models by combining numerous powerless learners (ordinarily choice trees), where each ensuing tree amends the blunder within the past tree. This iterative preparation points to decreasing the foreordained misfortune and moving forward the forecast demonstrates.

```
XGBRegressor
XGBRegressor(base_score=None, booster=None, callbacks=None,
             colsample_bylevel=None, colsample_bynode=None,
             colsample_bytree=None, device=None, early_stopping_rounds=None,
             enable_categorical=False, eval_metric=None, feature_types=None,
             gamma=None, grow_policy=None, importance_type=None,
             interaction_constraints=None, learning_rate=None, max_bin=None,
             max_cat_threshold=None, max_cat_to_onehot=None,
             max_delta_step=None, max_depth=None, max_leaves=None,
             min_child_weight=None, missing=nan, monotone_constraints=None,
             multi_strategy=None, n_estimators=None, n_jobs=None,
             num_parallel_tree=None, random_state=None, ...)
```

**Figure 1:** Model Training

***Best parameters and best score:***

The main purpose of model evaluation is to evaluate the performance of the training model on unseen data. This process provides information about the model's ability to extend to real-world situations and the accuracy of its predictions. Test data (X_test_scaled). The Predict() method calculates the predicted value of the target variable ("PM2.5" concentration) as input. Calculates the mean square difference between the predicted value (y_pred) and the actual value (y_test). The lower the MSE value, the better the model. A perfect model has an MSE of 0.

***Calculate the absolute difference between predicted and actual values:***

Absolute_difference = abs(y_test - y_pred): this line calculates the outright contrast between the real esteem (y_test) and the esteem anticipated by the XGBoost demonstration (y_pred). The abs() work guarantees that all factors are spoken to as positive, overlooking standard deviations.

***Model evaluation:***

The execution of the arbitrary woodland demonstration was assessed with different measurements, counting precision, exactness, review, and F1 score. These estimations give an understanding of the model's capacity to recognize liver infection and its general consistency.

***Dataset splitting for model evaluation:***

Sometime recently preparing and testing the expectation demonstrate, the accessible information must be isolated into subsets for preparing and testing. This method gives a fair-minded evaluation of the model's execution on concealed objects and makes a difference in its capacity to generalize past preparation light rates. By altering this esteem, analysts can control the extent of information put away for testing, in this way influencing the adjustment between preparing and testing expansive datasets. For this case, setting test_size_ratio = 0.2 will designate 20% of the information for testing and take off 80% for preparing. X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = test_size_ratio, random_state = 42). The train_test_split work in scikit-learn is utilized to part the include framework (X) and target vector (y) into preparing and testing. When test_size = test_size_ratio is indicated, the work will naturally distribute a few of the information to the test and the rest to the preparing handle. Also, the random_state = 42 parameter guarantees repeatability by relegating an irregular seed to the information segment.

***Calculating accuracy in percentage:***

Exactness may be a basic degree utilized to assess the execution of an expectation show and speaks to the proportion of redress forecasts to all expectations. Communicating exactness as a rate makes it less demanding to decipher and compare diverse tests or information sets. Genuine. Increase the result by 100 to change over exactness of to rate arrangement.

**print("Correct in ±{}:**
**{:.2f}%".format(threshold, delicate)):**

At that point print the right numbers to the console in a lucid organize. To be clear, the {:.2f} string sort permits the right esteem to be shown up to two decimal places. Furthermore, standard methods are included within the distributed content to supply settings for precision.

## RESULT AND DISCUSSION

The catastrophic climate crisis has prompted major transformations across the globe. Researchers from different nations are working together to create a method for tracking pollution levels by analyzing climate data from Velachery and utilizing the XGBoost model. This model, commonly used in air pollution studies to calculate AQI, has a higher accuracy rate compared to other machine learning algorithms. Additionally, it prioritizes the key factors that contribute to pollution levels. The model produces low RMSE values, making it suitable for real-world AQI prediction. By including other factors, the model can be expanded to predict AQI for a larger area. The inclusion of the XGBoost algorithm has important consequences for forecasting precision, mistake percentage, and understanding. It can forecast the urban air quality index (AQI) based on the current range of 6 pollutants, thereby safeguarding the future of AQI, promoting better practices for people's health, and providing scientific, convenient, and effective decisions for urban air prevention and control.

## CONCLUSION

***Corresponding Author:** S Harini

A detailed study on predicting air quality using the XGBoost algorithm and the proposed TMSSICX hybrid model achieved success in forecasting the AQI level. These models, which incorporate various factors such as climate and pollution data, provide a robust foundation for accurately predicting air pollution levels, and facilitating environmental monitoring and public health management. The results show that the XGBoost model outshines other machine learning algorithms, displaying lower error rates and reduced RMSE values. Moreover, the model evaluates weather parameters by their values, empowering drivers to better grasp fluctuating weather conditions. Additionally, the integration of the XGBoost algorithm has been proven to enhance prediction accuracy, decrease error rates, and improve interpretation. This highlights the efficacy of the technique for immediately estimating AQI, facilitating decision-making and implementing effective measures to mitigate urban pollution. As a result, the suggested model provides a hopeful method for forecasting the AQI level and has important impacts on production methods, advising public health officials, and dealing with issues caused by bad weather. Future research may aim to expand the model to predict AQI over a wider area and incorporate other factors to increase the accuracy and dependability of predictions.

## REFERENCES


1. Londhe, M. (2021). Data mining and machine learning approach for air quality index prediction. *International Journal of Engineering and Applied Physics*, *1*(2), 136-153.
2. Zhao, X., Song, M., Liu, A., Wang, Y., Wang, T., & Cao, J. (2020). Data-driven temporal-spatial model for the prediction of AQI in Nanjing. *Journal of Artificial Intelligence and Soft Computing Research*, *10*(4), 255-270.
3. Chowdhury, A. S., Uddin, M. S., Tanjim, M. R., Noor, F., & Rahman, R. M. (2020, August). Application of data mining techniques on air pollution of Dhaka city. In *2020 IEEE 10th International Conference on Intelligent Systems (IS)* (pp. 562-567). IEEE.
4. Bhalgat, P., Pitale, S., & Bhoite, S. (2019). Air quality prediction using machine learning algorithms. *International Journal of Computer Applications Technology and Research*, *8*(9), 367-370.
5. Soundari, A. G., Jeslin, J. G., & Akshaya, A. C. (2019). Indian air quality prediction and analysis using machine learning. *International Journal of Applied Engineering Research*, *14*(11), 181-186.
6. Kottur, S. V., & Mantha, S. S. (2015). An integrated model using Artificial Neural Network (ANN) and Kriging for forecasting air pollutants using meteorological data. *International Journal of Advanced Research in Computer and Communication Engineering*, *4*(1), 146-152.
7. Castelli, M., Clemente, F. M., Popovič, A., Silva, S., & Vanneschi, L. (2020). A machine learning approach to predict air quality in California. *Complexity*, *2020*.
8. Sivakumar, V., Kanagachidambaresan, G. R., Dhilip Kumar, V., Arif, M., Jackson, C., & Arulkumaran, G. (2022). Energy-efficient Markov-based lifetime enhancement approach for underwater acoustic sensor network. *Journal of Sensors*, *2022*, 1-10.
9. Hoq, M. N., Alam, R., & Amin, A. (2019, February). Prediction of possible asthma attack from air pollutants: Towards a high density air pollution map for smart cities to improve living. In *2019 International Conference on Electrical, Computer and Communication Engineering (ECCE)* (pp. 1-5). IEEE.
10. Pérez, P., Trier, A., & Reyes, J. (2000). Prediction of PM2.5 concentrations several hours in advance using neural networks in Santiago, Chile. *Atmospheric Environment*, *34*(8), 1189-1196.
11. Cannon, A. J. (2012). Neural networks for probabilistic environmental prediction: Conditional density estimation network creation and evaluation (cadence) in R. *Computers & Geosciences*, *41*, 126-135.
12. Hsieh, H. P., Lin, S. D., & Zheng, Y. (2015, August). Inferring air quality for station location recommendation based on urban big data. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 437-446).
13. Sethi, J. K., & Mittal, M. (2019). Ambient air quality estimation using supervised learning techniques. *EAI Endorsed Transactions on Scalable Information Systems*, *6*(22), e8-e8.
14. Swamy, S. R., & KS, N. P. (2022, December). Hybrid machine learning model for early discovery and prediction of polycystic ovary syndrome. In *2022 Second International Conference on Advanced Technologies in Intelligent Control, Environment, Computing & Communication Engineering (ICATIECE)* (pp. 1-8). IEEE.
15. Archana, R., Vaishnavi, C., Priyanka, D. S., Gunaki, S., Swamy, S. R., & Honnavalli, P. B. (2022, May). Remote health monitoring using IoT and edge computing. In *2022 International Conference on IoT and Blockchain Technology (ICIBT)* (pp. 1-6). IEEE.
16. Swamy, S. R., Prasad, K. N., & Tripathi, P. (2020, October). Smart home lighting system. In *2020 International Conference on Smart Innovations in Design, Environment, Management, Planning and Computing (ICSIDEMPC)* (pp. 75-81). IEEE.
17. Iskandaryan, D., Ramos, F., & Trilles, S. (2020). Air quality prediction in smart cities using machine learning technologies based on sensor data: A review. *Applied Sciences*, *10*(7), 2401.
18. Gocheva-Ilieva, S. G., Ivanov, A. V., Voynikova, D. S., & Boyadzhiev, D. T. (2014). Time series analysis and forecasting for air pollution in small urban area: an SARIMA and factor analysis


***Corresponding Author:** S Harini

approach. *Stochastic Environmental Research and Risk Assessment*, *28*, 1045-1060.

19. Campbell-Lendrum, D., & Prüss-Ustün, A. (2019). Climate change, air pollution and noncommunicable diseases. *Bulletin of the World Health Organization*, *97*(2), 160.
20. Haque, A., & Pant, A. B. (2022). Mitigating COVID-19 in the face of emerging virus variants, breakthrough infections and vaccine hesitancy. *Journal of Autoimmunity*, *127*, 102792.
21. Nigam, A., & Srivastava, S. (2023). Hybrid deep learning models for traffic stream variables prediction during rainfall. *Multimodal Transportation*, *2*(1), 100052.
22. Gupta, N. S., Mohta, Y., Heda, K., Armaan, R., Valarmathi, B., & Arulkumaran, G. (2023). Prediction of air quality index using machine learning techniques: A comparative analysis. *Journal of Environmental and Public Health*, *2023*, 1-26.
23. Hukkeri, S., Malage, R. V., Swamy, S. R., & Honnavalli, P. B. (2021). Estimation of engagement of learners in MOOCs using smart visual processing.
24. Singh, N., Shyam, A., Swamy, S. R., & Honnavalli, P. B. (2021). Differential privacy in NoSQL systems. In *Data Science and Security: Proceedings of IDSCS 2021* (pp. 374-384). Springer Singapore.

***Corresponding Author:** S Harini