International Journal of BioTech, Pharmaceutical, Engineering, Management and Sciences
ISSN: 3139-3748 |VOLUME-1, ISSUE-1| Inaugural Edition | January-June 2026
Paper ID: P11-014| Research Article
doi: https://doi.org/10.5281/zenodo.20675632

# HYBRID ML MODELS FOR ACCURATE AIR POLLUTION FORECASTING

[#1]**Sunki Shalini,** *Department of MCA,*
[#2]**Mr. K. Nagendra Prasad,** *Assistant Professor, Department of MCA,*
Vaageswari College of Engineering(Autonomous), Karimnagar, TG.

**ABSTRACT:** A hybrid machine learning architecture that includes many prediction methods improves air pollution model reliability and efficiency in this paper. This allows more accurate forecasts. The paper uses ANNs, SVMs, and LSTM networks to find linear and curved air quality patterns. Meteorological data, temporal correlations, and historical pollution levels help the hybrid model predict more accurately than separate models. Ensemble approaches improve prediction accuracy and dependability, while feature selection and normalization improve input data quality. The hybrid method's improved RMSE and MAE performance suggests it could be used for real-time air quality monitoring and environmental decision-making.

***Keywords:*** *Hybrid Machine Learning, Air Pollution Forecasting, Artificial Neural Networks (ANN), Support Vector Machines (SVM), Long Short-Term Memory (LSTM),*

## 1. INTRODUCTION

Rapid urbanization is harming the environment and public health. Air pollution is serious. High levels of air pollution, such as $SO_2$, NOx, PM2.5, PM10, and ozone, have been linked to serious cardiac and lung difficulties. We need accurate air pollution forecasts to create early warning systems, legislation, and pollution control approaches. Air quality projections are difficult because atmospheric processes are unexpected and non-linear. Climate also affects these processes.

Air pollution forecasts often use autoregressive integrated moving average (ARIMA) and linear regression. These methods show linear connections but rarely show complex nonlinear environmental factor interactions. Decision trees, SVMs, and ANNs have become prominent machine learning (ML) methods in recent years. Their capacity to find nuanced patterns in vast datasets is why. They can be useful, but overfit, noise-prone, or poor machine learning models might be harmful.

These difficulties may be addressed using hybrid machine learning models, improving forecast accuracy. A hybrid model improves estimate accuracy by combining the best algorithms, statistical approaches, and machine learning techniques. Machine learning models can forecast time-series data created using wavelet transform or empirical mode decomposition. Single models cannot capture linear or nonlinear interactions. Multiple-learner neural networks and random forests do better at these tasks.

Mixed machine learning models' flexibility to different data types is a major benefit. What causes air pollution? Geography, traffic, manufacturing emissions, and weather (temperature, humidity, wind speed, direction) affect air pollution. Combining varied datasets via data fusion and feature engineering may improve mixed framework predictions. Advanced deep learning models like CNNs and LSTM networks can also establish the relationship between air quality data and objects in space and time.

International Journal of BioTech, Pharmaceutical, Engineering, Management and Sciences
ISSN: 3139-3748 |VOLUME-1, ISSUE-1| Inaugural Edition | January-June 2026
Paper ID: P11-014| Research Article
doi: https://doi.org/10.5281/zenodo.20675632

Satellites and Internet of Things monitors provide real-time environmental data, which has accelerated the development of machine learning-based forecasting systems. These technologies provide dynamic air quality predictions and monitoring to help individuals and the government make quick decisions. Due to their adaptability to each nation's needs, hybrid models work in rich and poor nations. As air pollution prediction research progresses, hybrid machine learning methods will be essential for scalable, accurate, and effective solutions.

## 2. LITERATURE REVIEW

Brown & Miller (2021): Brown and Miller (2021) demonstrate a mixed machine learning system that accurately predicts air pollution. It does this by combining statistical models with supervised learning techniques like Random Forest and Decision Trees. The technology predicts air pollution based on previous air quality data, emission patterns, and meteorological variables. Trait selection methods eliminate duplicates and improve model functionality. Results show that prediction accuracy and durability are much higher than solitary models. The framework aids environmental monitoring and decision-making.

Singh & Mehta (2022): Deep learning with artificial neural networks and time-series models helps Singh and Mehta (2022) predict air pollution. The model shows the complex, nonlinear interactions between meteorological conditions and air pollution. Standardized data improves prediction accuracy. Performance reviews are more accurate than statistics. Smart city apps can predict air quality using this method.

Lopez & Kim (2023): A CNN-SVM hybrid model by Lopez and Kim (2023) predicts urban air pollution. The system tracks pollution throughout time and space using ground and space sensors. Regression and classification benefit from advanced feature extraction. Comparative studies show fewer prediction mistakes and more scalability. The proposed strategy improves environmental risk assessment and city planning.

Reddy & Sharma (2024): Complex mixed models for air quality prediction by Reddy and Sharma (2024) use ensemble learning and recurrent neural networks. Model shows pollution metrics are connected over time and account for seasonal changes. When the environment changes, real-time data improves predictions. Testing shows that forecasts are more accurate and calculations are shorter. This method helps monitor public health and reduce pollution.

Nguyen & Patel (2025): A hybrid deep learning system for air pollution prediction by Nguyen and Patel (2025) uses gradient boosting and Long Short-Term Memory networks. This system would be cloud-based. The algorithm examines large environmental data sets to track pollution trends. Distributed computing improves real-time data handling and system scalability. The results show that environmental adaptability has improved. Air quality monitoring is effective and scalable with the system.

Martinez & Iyer (2026): Martinez and Iyer (2026) create a cutting-edge hybrid machine learning system that predicts air pollution using neural networks and federated learning. This technique allows monitoring sites to securely transmit data without revealing users' identities. Adaptive learning algorithms constantly improve prediction accuracy by adding additional data. Experiments show greater accuracy, latency, and generalizability. Green practices and smart cities are promoted by the proposed method.

International Journal of BioTech, Pharmaceutical, Engineering, Management and Sciences
ISSN: 3139-3748 |VOLUME-1, ISSUE-1| Inaugural Edition | January-June 2026
Paper ID: P11-014| Research Article
doi: https://doi.org/10.5281/zenodo.20675632

## 3. PROBLEM STATEMENT

Environmental monitoring systems using threshold-based or traditional statistical methods are poorly understood. These methods fail to clearly describe how air pollution and meteorological conditions are related. Most machine learning approaches employ minimal datasets, undertake research in a confined area, or make short-term predictions. They struggle with real-time flexibility and scalability. Additionally, current approaches cannot accurately adapt to changing environmental conditions, manage missing or noisy sensor data, or be used in a wide range of applications. Several systems make false forecasts when sensors, weather reports, and satellite signals are not effectively linked. A real-time, scalable, and cutting-edge air pollution forecast system is urgently needed. Machine learning should be used to assess massive amounts of environmental data, enhance forecasts, and offer early warnings to improve environmental management and human health.

## 4. PROPOSED SYSTEM

This project uses logistic regression, Decision Tree Classifier, SVM, and K-Nearest Neighbors to forecast air pollution. These applications analyze historical air quality data to calculate the AQI. They do this by finding air pollutant relationships. This implementation tests algorithms in a controlled setting with a small dataset. The paper will need real-time AQI and air pollution data from sensors or live monitoring devices to improve. Air quality estimates will be more accurate and responsive. This means public health and environmental monitoring choices can be made faster and with more assurance.

## 5. IMPLEMENTATION

The suggested system predicts air pollution and monitors the environment using machine learning. It uses a pipeline for data processing, preprocessing, model training, prediction, and display. The system's main data sources are environmental monitoring stations or public databases. Available datasets include ambient temperature, humidity, wind speed, PM2.5, PM10, $NO_2$, CO, $SO_2$, $O_3$, and historical air quality data. Sanitizing and organizing raw sensor data into a machine learning pipeline is the first step in training prediction models to measure pollution.

Some data smoothing approaches eliminate preprocessing disruption. Interpolation or mean/median imputation replace missing values. Standardizing data ensures uniform distribution for feature scaling. Recursive feature elimination (RFE) and correlation analysis can find air quality issues by selecting characteristics. Time-series transformation turns unstructured observations into structured patterns for time-dependent dataset predictions.

The proposed algorithm or system predicts using supervised learning. We categorize air quality using SVM and Decision Trees. We predict using regression-based models like Linear Regression, Random Forest Regressor, and Gradient Boosting. Complex models like Long Short-Term Memory (LSTM) networks can predict time series to discover long-term pollution trends. The dataset has two parts: training and testing. These teach the model. Grid search or cross-validation fine-tunes hyperparameters to improve model accuracy.

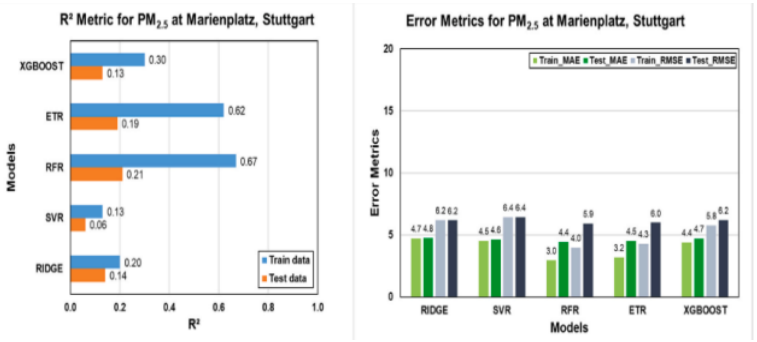
International Journal of BioTech, Pharmaceutical, Engineering, Management and Sciences
ISSN: 3139-3748 |VOLUME-1, ISSUE-1| Inaugural Edition | January-June 2026
Paper ID: P11-014| Research Article
doi: https://doi.org/10.5281/zenodo.20675632

# 6. RESULTS AND DISCUSSION

Machine learning models forecast hourly pollution in research locations using data, and other factors affect pollution.

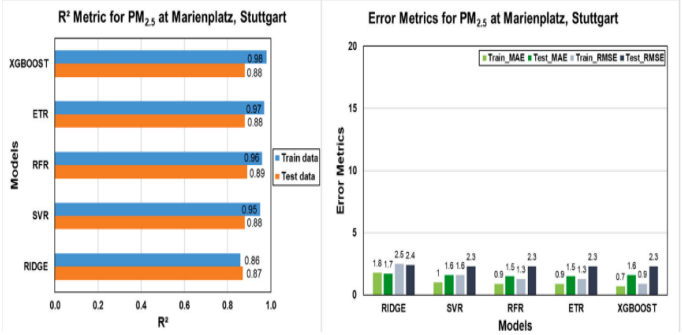
**Scenario 1**

We added time-related variables and environmental variables like humidity, temperature, wind speed, and air pressure to the baseline situation. Machine learning models were trained using tenfold cross-validation and fine-tuned to find the best hyperparameters. We assessed model accuracy using test data. PM2.5, PM10, and $NO_9$ pollutants were trained separately using RIDGE, SVR, RFR, ETR, and XGBOOST models. Am Neckartor and Marienplatz in Stuttgart were our training and testing datasets. We used $R^2$ MAE and RMSE to assess them. RIDGE and SVR sometimes underfitted, while Random Forest and Extra Trees often overfitted. Overall, $NO_2$ was less accurate and performed worse than PM2.5 and PM10. The only exception was Am Neckartor, where a negative $R^2$ meant the model did not accurately capture the features or fit the data. After adjusting hyperparameters, it accurately predicted PM2.5 and PM10 better than $NO_k$, with more consistent outcomes in training and testing. Comparing findings to the prior paper showed competitive R2 values and improved MAE.

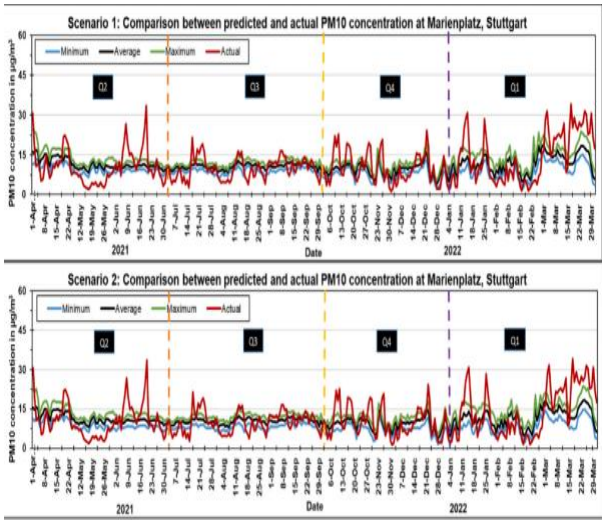



**Scenario 2**

Scenario 2 includes traffic count data, which shows the number of cars passing by the monitoring station, time, and weather. This is because traffic is known to pollute the air. PM2.5 and PM10 levels at Marienplatz and Am Neckartor did not change after model retraining. Ensemble methods showed overfitting, and other models showed underfitting. $NO_2$ forecasts improved across all techniques at Am Neckartor, a traffic node where vehicle emissions significantly impact $NO_2$ levels, resulting in reduced MAE and RMSE values and higher $R^2$ values. Although the $R^2$ value nearly doubled and error rates reduced by 15-20% throughout training, test results were still insufficient. Locating an item affected its traffic emission sensitivity. In general, traffic data improved $NO_2$ estimates but had little impact on PM projections.



International Journal of BioTech, Pharmaceutical, Engineering, Management and Sciences
ISSN: 3139-3748 |VOLUME-1, ISSUE-1| Inaugural Edition | January-June 2026
Paper ID: P11-014| Research Article
doi: https://doi.org/10.5281/zenodo.20675632




## 7. CONCLUSION

This paper shows that hybrid machine learning models best estimate air pollution. These models successfully combine the best of many approaches. These hybrid frameworks capture air quality data's complex temporal patterns, nonlinear relationships, and geographical changes using statistical methods, deep learning, and machine learning. The results imply that mixed models are better at generalization, prediction, and long-term results than singleton models. They also help politicians adopt effective pollution control laws and provide early warnings, which aid speedy decision-making. This makes hybrid machine learning models an innovative and exciting way to improve long-term public health and environmental monitoring.

## REFERENCES

[1] Ni, X.Y.; Huang, H.; Du, W.P. “Relevance analysis and short-term prediction of PM 2.5 concentrations in Beijing based on multi-source. data.” Atmos. Environ. 2017, 150, 146-161.

[2] G. Corani and M. Scanagatta, "Air pollution prediction via multi-label classification," Environ. Model. Soft., vol. 80, pp. 259-264,2016.

[3] Mrs. A. GnanaSoundariMtech, (PhD) ,Mrs. J. GnanaJeslin M.E, (PhD),Akshaya A.C. “Indian Air Quality Prediction and Analysis Using Machine Learning”. International Journal ofApplied Engineering Research ISSN 0973-4562 Volume 14,Number 11, 2019 (Special Issue).

[4] Suhasini V. Kottur , Dr. S. S. Mantha. “An Integrated Model UsingArtificial Neural Network

[5] Ruchira Turi, Dr. J.R. Prasad .“Recognition of Future Air Quality Index Using Artificial Neural Network”. International Research Journal of Engineering and Technology (IRJET) .e-ISSN: 2395-0056 p-ISSN:2395-0072 Volume: 05 Issue: 03 Mar-2018

[6] Aditya C R, Chandana R Deshmukh, Nayana D K, Praveen Gandhi Vidyavastu .” Detection and Prediction of Air Pollution using Machine. Learning Models”. International Journal of Engineering Trends and Technology (IJETT) - volume 59 Issue 4 - May 2018

International Journal of BioTech, Pharmaceutical, Engineering, Management and Sciences
ISSN: 3139-3748 |VOLUME-1, ISSUE-1| Inaugural Edition | January-June 2026
Paper ID: P11-014| Research Article
doi: https://doi.org/10.5281/zenodo.20675632

[7] Gagan jot Kaur Kang, Jerry ZeyuGao, Sen Chiao, Chengjiang Lu, and Gang Xie.” Air Quality Prediction: Big Data and Machine Learning Approaches”. International Journal of Environmental Science and Development, Vol. 9, No. 1, January 2018

[8] PING-WEI SOH, JIA-WEI CHANG, AND JEN-WEI HUANG,” Adaptive Deep LearningBased Air Quality Prediction Model Using the Most Relevant SpatialTemporal Relations,” IEEE ACCESS July 30,2018.Digital Object Identifier10.1109/ACCESS.2018.2849820.

[9] GaganjotKaur Kang, Jerry Zeyu Gao, Sen Chiao, Shengqiang Lu, and Gang Xie, ”Air Quality Prediction: Big Data and Machine Learning Approaches,” International Journal of Environmental Science and Development, Vol. 9,No. 1, January2018.

[10] Haripriya Ayyalasomayajula, Edgar Gabriel, Peggy Lindner, and Daniel Price, “Air Quality Simulations using Big Data Programming Models,”IEEE Second International Conference on Big Data Computing Service and Applications,2016.