

# A Hybrid Machine Learning and Deep Learning Framework for Air Quality Index Prediction Using Advanced Temporal Feature Engineering

Aayushi Patel<sup>\*</sup>, Prof Deepak Agrawal<sup>\*\*</sup>, Prof. Prateek Gupta<sup>\*\*\*</sup>

<sup>\*</sup>Research Scholar Department of CSE, Shriram Institute of Science & Technology, Jabalpur, M.P.

<sup>\*\*</sup>Prof., Department of CSE, Shriram Institute of Science & Technology, Jabalpur, M.P.

<sup>\*\*\*</sup>

**Abstract-** AQI prediction is a challenging task due to the complex, nonlinear, and time-dependent interactions among various pollutants. This study proposes a comprehensive and hybrid framework for AQI prediction by integrating advanced feature engineering, machine learning, deep learning, and ensemble techniques. Initially, extensive data preprocessing was performed, followed by the generation of multiple feature types, including temporal features, lag features, rolling statistics, and interaction features, to capture seasonal patterns and temporal dependencies in air quality data. Several machine learning models, such as Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, K-Nearest Neighbors, and XGBoost, were implemented and evaluated. Among these, XGBoost demonstrated superior performance for AQI classification tasks. To address class imbalance, various resampling techniques were explored; however, a manual class weighting strategy was found to be more effective in improving class-wise prediction balance. In addition, a deep learning model was developed and optimized using techniques such as batch normalization, dropout, and early stopping. Although deep learning models alone did not outperform tree-based methods, they contributed to improved representation learning and balanced predictions. Furthermore, ensemble learning through stacking was explored to combine the strengths of multiple models. The experimental results demonstrate that feature engineering and model optimization play a critical role in improving AQI prediction performance. The proposed framework achieves high

accuracy and balanced classification across AQI categories, making it suitable for real-world applications in environmental monitoring.

**Keywords:** Air Quality Index (AQI), Machine Learning, Deep Learning, XGBoost, Feature Engineering, Time-Series Analysis, Classification, Environmental Monitoring, Air Pollution Prediction.

## I. INTRODUCTION

The resilience of our society's sustainability is under significant threat due to rising pollution levels and diminishing infrastructural resilience. Pollution, driven by both anthropogenic activities and natural processes, poses a severe risk to public health and the environmental balance. Among various forms of pollution, deteriorating air quality has emerged as a critical global concern. Rapid industrialization and urbanization have led to significant emissions from automobiles, industries and coal burning, resulting in severe air pollution that endangers human health and ecological stability. Recognizing these challenges, the Party's 20th National Congress has emphasized the importance of strengthening environmental pollution prevention and control, with the goal of reducing pollution events and promoting the harmonious coexistence between humanity and nature.

Several methods have been proposed to predict environmental phenomena to mitigate these risks. Soomro et al. [1] highlighted the importance of understanding precipitation-vegetation relationships in the Kunhar River basin, while Soomro et al. [2] analyzed flood susceptibility using geomorphometric parameters and hydrological models, providing insights for river basin management and disaster preparedness. Han et al. [3] demonstrated how vegetation changes significantly influence surface air temperature and broader climate patterns, exacerbating

environmental degradation and extreme weather. Sun et al. (2022) showcased the integration of machine learning, such as K-means clustering, with traditional hydrological models to enhance urban flood simulations. Tang et al. [4] reinforced the significance of advanced machine learning methods in predicting environmental phenomena like floods, bolstering responses to emerging environmental threats.

In the realm of air quality prediction, researchers have employed diverse technologies and models. Biancofiore et al. [5] developed a recursive neural network model for analyzing and forecasting PM10 and PM2.5 concentrations. Yang and Wang (2017) [6] proposed a comprehensive air quality prediction system that combines data exploration and machine learning techniques. Wang et al. [7] introduced a CT-LSTM model specifically designed for air quality prediction. Huang et al. [8] developed a large-scale air pollution prediction model using deep convolutional networks, highlighting the potential of deep learning in this domain. Zhang et al. [9] created a new air quality monitoring and early warning system that combines air quality assessment and air pollutant concentration prediction.

Recently, novel approaches have been proposed for advancing air quality prediction. Pawul [10] demonstrated neural networks' efficacy of machine learning for predicting gas pollution. Bekkar et al. [11] proposed an effective deep neural network model for smart cities, while Seng et al. [12] used LSTM for spatiotemporal predictions. Zhang et al., (2021a) used a semi-supervised bidirectional LSTM, incorporating both labelled and unlabelled data. Zeng et al. [13] combined LSTM with extended stationary wavelet transform, capturing temporal and frequency characteristics. Liu and Guo (2022) [14] combined LSTM with the Sparrow Search Algorithm (SSA) to achieve promising results. Ji et al. (2022) [15] proposed a multi-scale evolutionary deep learning model integrating CEEMDAN, an improved whale optimization algorithm, a regularized extreme learning machine, and LSTM for AQI prediction. Li et al. [16] introduced an evolutionary deep learning model based on an improved grey wolf optimization algorithm and DBN-ELM for enhanced AQI prediction. Mengara and Park [17] developed an attention-based distributed deep learning model for more accurate forecasts. Zhao et al. (2023) [18] developed a hybrid deep learning framework considering temporal and spatial

factors, enhancing prediction accuracy during the COVID-19 pandemic.

Additionally, studies have explored the application of machine learning models for air quality prediction. Wang and Song [19] introduced a deep spatial-temporal ensemble model considering spatial and temporal correlations. Zhu et al. [20] enhanced prediction accuracy with model regularization and optimization. Bhalgat et al. [21] found SVM performed best for air quality prediction. Dun et al. [22] Used fractional grey linear regression with SVM for accurate short-term air quality prediction.

As a result of technological advancements and rapid industrialization, air pollution has become a serious global problem. It is caused by a combination of gases and particles, such as PM2.5, nitrogen oxides (NO<sub>x</sub>), and sulphur dioxide (SO<sub>2</sub>), which are released into the atmosphere and lead to major problems. The issue of air pollution is particularly severe in urban areas due to high traffic emissions, industrial activities, and energy consumption. According to a study by the Centers for Disease Control, air quality in rural areas is often better than in urban areas. However, although pollutant levels are low, the study still detected measurable contaminants in rural areas. Studies have shown that exceeding the World Health Organization (WHO) threshold limits for air pollutants can have severe consequences for human health. The WHO recommends maintaining annual mean concentrations 8-hour mean ozone concentration below 60 µg/m<sup>3</sup> (World Health Organization, 2021). According to the WHO, air pollution is the fourth leading cause of death worldwide, following cardiovascular disease, smoking, and poor diet (World Health Organization, 2021). In 2019, air pollution, both ambient and household, was responsible for 6.7 million deaths globally, highlighting the significant health burden posed by poor air quality (World Health Organization, 2024). In addition to its impact on human health, air pollution also negatively affects both the environment and the economy.

## II. LITERATURE SURVEY

**Paper 1:** Accurate hourly AQI prediction using temporal CNN-LSTM-MHA+GRU: A case study of seasonal variations and pollution extremes in Visakhapatnam, India. [23]

Method Used: Hybrid Deep Learning architecture combining CNN for spatial feature extraction, Multi-Head

Attention (MHA) for important temporal feature weighting, LSTM for long-term dependency modeling, and GRU for residual error correction. Feature engineering included lag features (AQI\_lag1–AQI\_lag5) and rolling statistics (7-day mean and std). Feature selection was done using Spearman correlation and Recursive Feature Elimination (RFE).

Dataset used: CPCB (Central Pollution Control Board, India) dataset from Visakhapatnam city. Dataset period Oct 2022 – Oct 2024 with 17,592 hourly records and 21 variables including pollutant concentrations (PM2.5, PM10, NO<sub>2</sub>, SO<sub>2</sub>, CO, O<sub>3</sub>, etc.) and meteorological parameters (temperature, humidity, wind speed, rainfall, solar radiation, etc.).

Results: The proposed hybrid model outperformed baseline models such as CNN, LSTM, CNN-LSTM, Transformer, AddNet, and MGSFormer. Performance improvement was observed using metrics such as R<sup>2</sup>, RMSE, and MAE, demonstrating better AQI prediction accuracy due to integration of attention mechanism and residual learning.

Limitations: The model was tested only on one city dataset, limiting generalization. Computational complexity is high due to hybrid deep learning architecture. Real-time deployment feasibility and cross-regional validation were not explored.

**Paper 2:** Ensemble learning for air quality index prediction: integrating gradient boosting, XGBoost, and stacking with SHAP-based interpretability [24]

Method Used: Multiple ML and DL models including Linear Regression, KNN, SVR, Decision Tree, AdaBoost, Gradient Boosting, XGBoost, LightGBM, CatBoost, LSTM, CNN-LSTM, Transformer with Weighted Ensemble Voting Regressor

Dataset Used: Taiwan Air Quality Dataset (2016–2024) containing pollutant data (PM2.5, PM10, NO<sub>2</sub>, SO<sub>2</sub>, CO, O<sub>3</sub>) and meteorological variables (wind speed, wind direction, temperature, humidity, pressure). Total cleaned records: 4,608,039 (subset of 10,000 samples used for training).

Results: Best performance achieved by Weighted Ensemble model combining GBR, CatBoost, XGBoost and LightGBM. Validation R<sup>2</sup> = 0.9969, Test R<sup>2</sup> = 0.9978, showing strong prediction accuracy. Tree-based ensemble models significantly outperformed classical regression and deep learning models for tabular AQI data.

Limitations: 1. Only 10,000 samples used from a much

larger dataset, reducing full data utilization.

2. Study limited to Taiwan region, so global generalization not fully validated.

3. Focus only on AQI regression prediction, not classification of AQI categories.

4. Real-time deployment not implemented.

5. Temporal forecasting capability limited despite temporal validation.

**Paper 3:** Title: Hybrid deep learning model for air quality prediction and its impact on healthcare [25].

Method Used: Deep Learning models: CNN, LSTM, and Hybrid CNN-LSTM with extensive feature engineering (lag features, moving averages, cyclic encoding, and pollutant ratios)

Dataset Used: Gurugram Air Quality Dataset collected from OpenWeather Air Pollution API (Jan 10 – May 9, 2024), 2898 samples, pollutants: PM10, PM2.5, O<sub>3</sub>, SO<sub>2</sub>, NO<sub>2</sub>, CO.

Results: Hybrid CNN-LSTM achieved best performance. Precision = 0.9101, Recall = 0.9118 with engineered features. Hybrid model outperformed individual CNN and LSTM.

Limitations: Dataset size very small (2898 samples). Only single city (Gurugram) is used. Limited pollutants used. No classical ML models comparison (RF, XGBoost). No real-time deployment system.

**Paper 4:** Machine learning for air quality prediction and data analysis: Review on recent advancements, challenges, and outlooks [26].

Method Used: Machine Learning (ML) algorithms have proven to be powerful tools for enhancing air quality prediction and addressing monitoring challenges. However, a comprehensive review compiling the research space of ML for air quality is seldom available. This review analyzes over 70 recent studies that apply ML techniques to air quality monitoring, categorizing them based on the type of learning approach employed, with a focus on identifying the most effective algorithms in each category. The findings demonstrate that ensemble models such as Random Forest (RF) and Extreme Gradient Boosting (XGBoost) consistently achieve high accuracy in structured datasets, while deep learning (DL) approaches like Long Short-Term Memory (LSTM) and Convolutional Neural Networks (CNN) excel in capturing temporal dependencies and spatial patterns in pollution forecasting. Unsupervised

approaches like clustering and anomaly detection effectively enhance data quality and sensor calibration, whereas reinforcement learning shows promise in adaptive control scenarios, despite challenges related to computational intensity and interpretability. This review is highly significant, offering valuable insights for policymakers and researchers in developing strategies to mitigate air pollution and improve public health using advanced ML techniques.

**Paper 5:** Explainable forecasting of air quality index using a hybrid random forest and ARIMA model [27].

Method Used: Hybrid Random Forest Regressor + ARIMA with SHAP Explainable AI.

Dataset Used: Kaggle AQI dataset (city\_day, city\_hour, station\_day, station\_hour, stations metadata).

Results: Hybrid model improves AQI forecasting by combining nonlinear modeling (RFR) and temporal modeling (ARIMA). Explainability achieved using SHAP.

Limitations: Model evaluated mainly on aggregated India AQI. Limited comparison with deep learning models like Transformer or CNN-LSTM.

**Paper 6:** PM2.5: Air Quality Index Prediction Using Machine Learning: Evidence from Kuwait's Air Quality Monitoring Stations [28].

Method Used: Linear Regression, Random Forest, SVM, Neural Network, Gradient Boosting, AdaBoost, Decision Tree, KNN.

Dataset Used: Air quality data from 12 monitoring stations in Kuwait (2019–2021). Initially 300,000 records, reduced to 37,731 after preprocessing.

Results: ML models successfully predicted PM2.5 AQI levels using environmental and meteorological parameters. Feature selection improved model performance and interpretability.

Limitations: Dataset only from Kuwait region, limited geographic generalization. Focus mainly on regression models, limited deep learning approaches

Title: Air quality prediction by machine learning models: A predictive study on the Indian coastal city of Visakhapatnam [29].

Method Used: LightGBM, Random Forest, CatBoost, AdaBoost, XGBoost.

Dataset Used: CPCB air quality data (Visakhapatnam, India) with 1920 observations (2017–2022).

Results: Ensemble models successfully predicted AQI with

high correlation between PM2.5, PM10 and AQI.

Limitations: Small dataset size, no deep learning models, limited spatial analysis.

### III. PROPOSED WORK

This study proposes a hybrid machine learning and deep learning framework for Air Quality Index (AQI) prediction and classification. The framework is designed to overcome the limitations of existing studies by integrating large-scale datasets, advanced feature engineering, ensemble learning, time-series forecasting, and explainable artificial intelligence (XAI) techniques. The proposed methodology consists of several sequential stages including data acquisition, preprocessing, exploratory analysis, feature engineering, model development, ensemble learning, evaluation, and interpretability analysis.

The overall workflow of the proposed framework is illustrated through the following stages.

#### 1. Data Collection

The first stage involves collecting air pollution and meteorological datasets from multiple reliable sources. The dataset contains both pollutant concentration data and environmental parameters that influence air quality.

The main pollutant attributes considered in this study include:

- PM2.5
- PM10
- NO2
- SO2
- CO
- O3

In addition to pollutant concentrations, meteorological variables are also incorporated to improve prediction accuracy.

These include:

- Temperature
- Humidity
- Wind speed
- Wind direction
- Atmospheric pressure
- Rainfall

The Air Quality Index (AQI) is considered as the target variable.

The use of multi-source and multi-city datasets helps improve the generalization capability of the prediction system.

## 2. Data Preprocessing

Raw environmental datasets often contain missing values, noise, and inconsistent measurements, which must be handled before model training.

The following preprocessing steps are applied.

### 2.1 Missing Value Handling

Missing values in numerical features are handled using mean or median imputation, while categorical attributes are filled using mode imputation.

### 2.2 Outlier Detection and Removal

Outliers in pollutant concentrations can significantly affect model performance. Therefore, the Interquartile Range (IQR) method is used to detect and remove abnormal values.

$$IQR=Q3-Q1$$

Values outside the range

$$[Q1-1.5IQR, Q3+1.5IQR]$$

are treated as outliers.

### 2.3 Feature Scaling

To ensure that features contribute equally to the learning process, feature scaling techniques such as:

- Min–Max Normalization
- Standardization

are applied.

Standardization formula:

$$Z=X-\mu\sigma$$

Where

$X$  = feature value.

$\mu$  = mean.

$\sigma$  = standard deviation

### 2.4 Data Splitting

The dataset is divided into three subsets:

- Training set (70%)
- Validation set (15%)
- Testing set (15%)

Additionally, k-fold cross-validation is applied to improve model reliability.

## 3. Exploratory Data Analysis (EDA)

EDA is performed to understand the underlying patterns and relationships within the dataset.

The following analysis techniques are applied:

### Univariate Analysis

Distribution of individual variables using:

- Histograms
- Box plots
- Density plots

### Bivariate Analysis

Relationships between AQI and pollutants using:

- Scatter plots
- Correlation matrices
- Heatmaps

### Multivariate Analysis

Interaction between multiple variables using:

- Pair plots
- Cluster visualization

EDA helps identify:

- Highly correlated pollutants
- Seasonal pollution patterns
- Data imbalance issues

## 4. Feature Engineering

Feature engineering plays a crucial role in improving model performance.

Several new features are derived from existing variables.

### Temporal Features

- Hour of the day
- Day of the week
- Month
- Seasonal indicators

### Rolling Statistical Features

Moving averages such as:

- 24-hour PM2.5 average
- 8-hour O3 average

### Lag Features

To capture temporal dependencies, lag features are generated:

$PM2.5_{t-1}$

These features help models learn time-dependent pollution behavior.

## 5. Model Development

To address limitations in existing literature, this study evaluates three categories of models.

### 5.1 Classical Machine Learning Models

Traditional machine learning models are used as baseline methods.

These include:

- Linear Regression
- Support Vector Regression (SVR)
- K-Nearest Neighbor (KNN)
- Decision Tree
- Random Forest

These models help provide a benchmark for performance comparison.

### 5.2 Advanced Ensemble Models

Ensemble models are used because they perform exceptionally well on structured environmental data.

The following models are implemented:

- Gradient Boosting Regressor
- XGBoost
- LightGBM
- CatBoost

These algorithms combine multiple weak learners to improve prediction accuracy. The architecture of the proposed model is shown below:

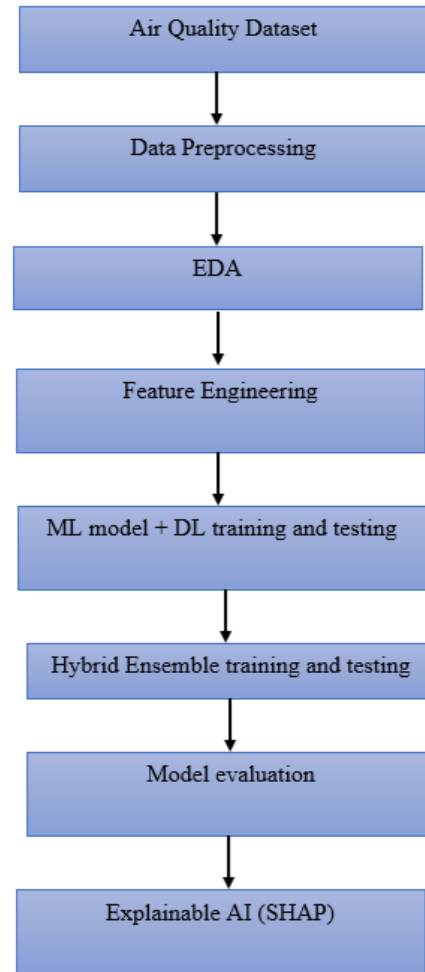


Figure 3.1: Proposed model architecture.

## IV. RESULTS

For the implementation of the work, six machine learning models have been used for classification of the AQI. After that some data balancing techniques are applied and again models are trained and tested for the AQI prediction. All the results are shown below.

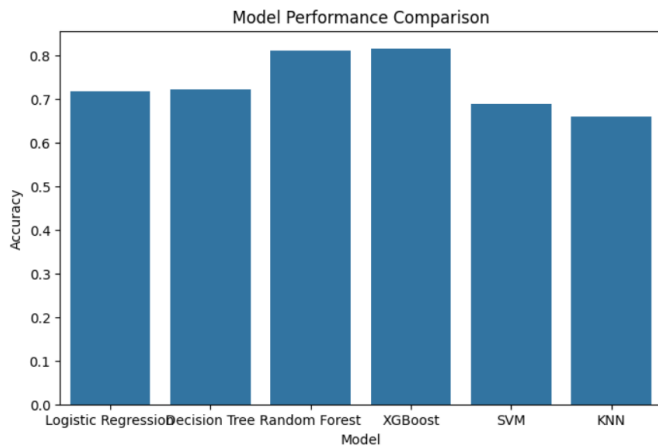


Figure 4.1: Accuracy Comparison Chart of ml models.

Model	Accuracy after SMOTE
0 Logistic Regression	0.658011
1 Decision Tree	0.706723
2 Random Forest	0.805757
3 XGBoost	0.800523
4 SVM	0.646538
5 KNN	0.582528

Figure 4.2: Results after SMOTE.

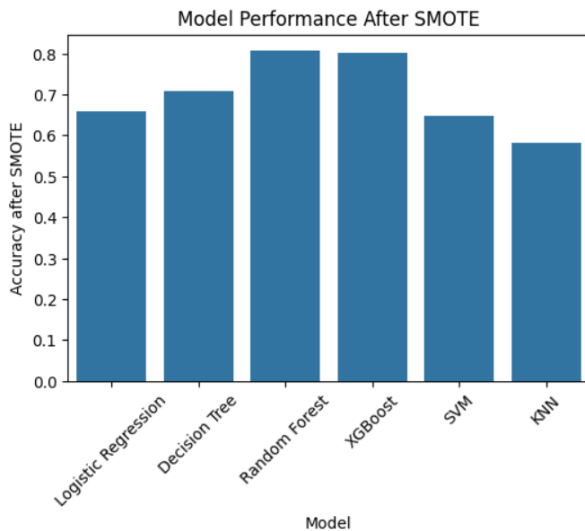


Figure 4.3: Model results after SMOTE.

## V. CONCLUSION

This study presented a comprehensive framework for Air Quality Index (AQI) prediction using a combination of advanced feature engineering, machine learning, deep learning, and ensemble techniques. The objective was to develop an accurate and robust model capable of handling the complex, nonlinear, and time-dependent nature of air pollution data.

Initially, multiple baseline machine learning models were implemented and evaluated. Among them, tree-based models demonstrated superior performance for tabular AQI data. To further enhance prediction accuracy, extensive feature engineering was performed, including temporal features, lag features, rolling statistics, and interaction features. These features successfully captured seasonal patterns, temporal dependencies, and relationships among pollutants, leading to a significant improvement in model performance.

## VI. REFERENCES

- [1] Soomro, S. E. H., Hu, C., & Jian, S. (2021). Precipitation changes and their relationships with vegetation responses during 1982–2015 in Kunhar River basin, Pakistan. *Water Supply*, 21(7), 3657–3671.
- [2] Soomro, S. E. H., Hu, C., & Boota, M. W. (2022). River flood susceptibility and basin maturity analysed using a coupled approach of geo-morphometric parameters and SWAT model. *Water Resources Management*, 36(7), 2131–2160.
- [3] Han, Z., Wu, Q., & Lai, R. (2022). Spatio-temporal variations of vegetation cover and its influence on surface air temperature change over the Yellow River Basin, China. *Journal of Water and Climate Change*, 13(9), 3239–3252.
- [4] Tang, Y., Sun, Y., & Han, Z. (2023). Flood forecasting based on machine learning pattern recognition and dynamic migration of parameters. *Journal of Hydrology: Regional Studies*, 47, 101406.
- [5] Biancofiore, F., Busilacchio, M., & Verdecchia, M. (2017). Recursive neural network model for analysis and forecast

of PM10 and PM2.5. *Atmospheric Pollution Research*, 8(4), 652–659.

[6] Yang, Z., & Wang, J. (2017). A new air quality monitoring and early warning system: Air quality assessment and air pollutant concentration prediction. *Environmental Research*, 158, 105–117.

[7] Wang, J., Li, J., & Wang, X. (2021). Air quality prediction using CT-LSTM. *Neural Computing and Applications*, 33, 4779–4792.

[8] Huang, G., Ge, C., & Xiong, T. (2021). Large scale air pollution prediction with deep convolutional networks. *Science China Information Sciences*, 64, 1–11.

[9] Zhang, Z., Zeng, Y., & Yan, K. (2021b). A hybrid deep learning technology for PM2.5 air quality forecasting. *Environmental Science and Pollution Research*, 28, 39409–39422.

[10] Pawul, M. (2019). Application of neural networks to the prediction of gas pollution of air. *New Trends in Production Engineering*, 2(1), 515–523.

[11] Bekkar, A., Hssina, B., & Douzi, S. (2021). Air-pollution prediction in smart city, deep learning approach. *Journal of Big Data*, 8(1), 1–21.

[12] Seng, D., Zhang, Q., & Zhang, X. (2021). Spatiotemporal prediction of air quality based on LSTM neural network. *Alexandria Engineering Journal*, 60(2), 1517–1526.

[13] Zeng, Y., Chen, J., & Jin, N. (2022). Air quality forecasting with hybrid LSTM and extended stationary wavelet transform. *Building and Environment*, 213, 108822.

[14] Liu, X., & Guo, H. (2022). Air quality indicators and AQI prediction coupling long-short term memory (LSTM) and sparrow search algorithm (SSA): A case study of Shanghai. *Atmospheric Pollution Research*, 13(10), 101551.

[15] Ji, C., Zhang, C., & Hua, L. (2022). A multi-scale evolutionary deep learning model based on CEEMDAN, improved whale optimization algorithm, and regularized extreme learning machine and LSTM for AQI prediction. *Environmental Research*, 215, 114228.

[16] Li, Y., Peng, T., & Hua, L. (2022). Research and application of an evolutionary deep learning model based on improved grey wolf optimization algorithm and DBN-ELM for AQI prediction. *Sustainable Cities and Society*, 87, 104209.

[17] Mengara, A., & Park, K. (2022). Arrangement of ecological civilization-construction in the Report of the 20th CPC National Congress. *Tibet Development Forum*, 192(2), 25–30.

[18] Zhao, Z., Wu, J., & Cai, F. (2023). A hybrid deep learning framework for air quality prediction with spatial autocorrelation during the COVID-19 pandemic. *Scientific Reports*, 13(1), 1015.

[19] Wang, J., & Song, G. (2018). A deep spatial-temporal ensemble model for air quality prediction. *Neurocomputing*, 314, 198–206.

[20] Zhu, D., Cai, C., & Yang, T. (2018). A machine learning approach for air quality prediction: Model regularization and optimization. *Big Data and Cognitive Computing*, 2(1), 5.

[21] Bhargat, 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.

[22] Dun, M., Xu, Z., & Chen, Y. (2020). Short-term air quality prediction based on fractional grey linear regression and support vector machine. *Mathematical Problems in Engineering*, 2020(1), 8914501.

[23] Sreenivasulu T, Mokesh Rayalu G, Accurate hourly AQI prediction using temporal CNN-LSTM-MHA+GRU: A case study of seasonal variations and pollution extremes in

Visakhapatnam, India, Results in Engineering, Volume 27, 2025, 106303, ISSN 2590-1230, <https://doi.org/10.1016/j.rineng.2025.106303>.

139518, ISSN 0045-6535, <https://doi.org/10.1016/j.chemosphere.2023.139518>

[24] Singh, S., Kumar, M., Sengar, V. *et al.* Ensemble learning for air quality index prediction: integrating gradient boosting, XGBoost, and stacking with SHAP-based interpretability. *Sci Rep* (2026). <https://doi.org/10.1038/s41598-026-39232-w>.

[25] Madan, T., Sagar, S., Singh, Y. *et al.* Hybrid deep learning model for air quality prediction and its impact on healthcare. *Sci Rep* **16**, 6036 (2026). <https://doi.org/10.1038/s41598-026-36564-5>.

[26] Manal Karmoude, Brenton Munhungewarwa, Isaiah Chiraira, Ryan Mckenzie, Jude Kong, Bevan Smith, Gelan Ayana, Nkosiphendule Njara, Thuso Mathaha, Mukesh Kumar, Bruce Mellado, "Machine learning for air quality prediction and data analysis: Review on recent advancements, challenges, and outlooks", *Science of The Total Environment*, Volume 1002, 2025, 180593, ISSN 0048-9697, <https://doi.org/10.1016/j.scitotenv.2025.180593>.

[27] Anuradha Yenikar, Ved Prakash Mishra, Manish Bali, Tabassum Ara, "Explainable forecasting of air quality index using a hybrid random forest and ARIMA model", *MethodsX*, Volume 15, 2025, 103517, ISSN 2215-0161, <https://doi.org/10.1016/j.mex.2025.103517>.

[28] Alrashidi, H., Sibai, F. N., Abonamah, A., Alrashidi, M., & Alsaber, A. (2025). PM2.5: Air Quality Index Prediction Using Machine Learning: Evidence from Kuwait's Air Quality Monitoring Stations. *Sustainability*, *17*(20), 9136. <https://doi.org/10.3390/su17209136>.

[29] Gokulan Ravindiran, Gasim Hayder, Karthick Kanagarathinam, Avinash Alagumalai, Christian Sonne, "Air quality prediction by machine learning models: A predictive study on the Indian coastal city of Visakhapatnam", *Chemosphere*, Volume 338, 2023,