International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(11s) 2026 pp. 672-686
https://doi.org/10.70917/ijcisim-2026-3779
Article

A Machine Learning and Explainable AI (XAI) Framework with LLM Fine-Tuning for PM2.5 Prediction and EV Impact Analysis

Shubham Goyal[1], Chabi Kumar[2], Ruchi Sharma[3]

[1]Malaviya National Institute of Technology (MNIT) Jaipur, JLN Marg, Jaipur-302017, Rajasthan, India.
Email: 2020rce9517@mnit.ac.in
[2]Malaviya National Institute of Technology (MNIT) Jaipur, JLN Marg, Jaipur-302017, Rajasthan, India.
Email: 2023rce9511@mnit.ac.in
[3]Malaviya National Institute of Technology (MNIT) Jaipur, JLN Marg, Jaipur-302017, Rajasthan, India.
Email: ruchi.ce@mnit.ac.in
Corresponding Author: Ruchi Sharma

Abstract: Prediction and mitigation are important challenges in order to adequately address the environmental and public health threat of air pollution PM2.5. This study presents a new “Machine Learning (ML)” and “Explainable Artificial Intelligence (XAI)” method that combines with fine-tuned Large Language Model (LLM) for forecasting PM2.5 atmospheric pollution and analyzing the effects of Electric Vehicles (EVs) on the environment. Data from Air Quality Data in India from Kaggle were preprocessed by removing missing values, duplicate records, detecting outliers, scaling features and splitting it into 80:20 train test sets. The ensemble learning models such as Random Forest, XGBoost, LightGBM and LSTM were tested using MAE, RMSE and R2 metrics. LightGBM achieved the best performance with MAE = 4.08, RMSE = 5.02, and $R^2 = 0.97$. The feature importance analysis indicates that PM10 (0.284), NO_2 (0.226) and temperature (0.173) are the most important features. The SHAP indicates that PM10 (0.312) is the most important feature. Seasonal analysis revealed that the season with the highest PM2.5 was December (165.8 µg/m³) and that of the lowest was July (36.9 µg/m³). The proposed framework offers reliable, understandable and policy informed air quality forecasts for sustainable environmental management.
Keywords: PM2.5 Prediction, Machine Learning, Explainable Artificial Intelligence (XAI), Large Language Models (LLMs), Electric Vehicle (EV) Impact Analysis.

1. Introduction

Particularly in the world's fast growing urban and industrial regions, air pollution has become one of the biggest environmental and public health concerns. One of the most harmful air pollutants is particulate matter with an aerodynamic diameter of less than 2.5 µm (PM2.5), which can be deeply inhaled and enter the bloodstream, causing respiratory conditions, cardiovascular diseases, and early death [1].The “World Health Organisation(WHO)” states that millions of people around the world are dying every year due to prolonged exposure of PM2.5, which highlights the importance of precise monitoring, forecasting and mitigation measures [2]. Although a large number of traditional statistical models have been developed for air quality forecasting, these models fail to provide sufficient capability for modelling complex nonlinear relationships among the wide range of meteorological, traffic, industrial and environmental variables and thereby lead to poor forecast accuracy under dynamic atmospheric conditions [3].

The latest progress in “Machine Learning (ML)” has led to improved prediction of air quality by creating models that could capture complex-space time structure in vast datasets related to the environment [4]. Recent algorithms and techniques like "Random Forest (RF)", "Extreme Gradient Boosting (XGBoost)", "Light Gradient Boosting Machine (LightGBM)", "Support Vector Regression (SVR)" and deep learning models such as "Long Short-Term Memory (LSTM)" networks are shown to improve performance over traditional approaches [5]. However, many of these models are "black box" models which offer little transparency on the factors affecting the PM2.5 prediction [6]. This opacity could be problematic for policy makers, environmental agencies and urban planners who need reliable and transparent predictions to aid decision making for effective interventions [7, 8].

"Explainable Artificial Intelligence (XAI)" has come to the fore as an interesting remedy for enhancing transparency and dependability of machine learning systems. "SHapley Additive exPlanations (SHAP)", "Local Interpretable Model-Agnostic Explanations (LIME)" and feature importance analysis could be used to gain insight from a model's decisions by determining the relative importance of environmental and anthropogenic variables [9]. The integration of XAI and machine learning helps stakeholders to comprehend not simply the predicted amount of PM2.5 but additionally the main causes behind pollution episodes. Providing such explainability would increase user confidence, aid in scientific interpretation and enable development of specific air quality management strategies [10, 11].

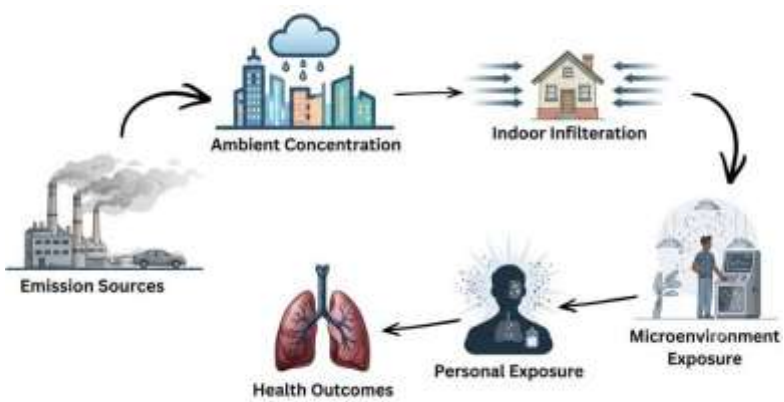

Figure 1: PM2.5 Exposure Pathway [12]

At the same time, the rapid development of "Large Language Models (LLMs)" have created fresh opportunities for environmental intelligence. In addition to natural language processing, LLMs could be trained to interpret domain-specific environmental texts, summarise prediction results, support the environmental decision-making process and give human-readable explanations of intricate analysis results by fine-tuning them with domain-specific environmental literature, air quality reports and policies [13]. When finely-trained with air quality data and XAI outputs, these LLMs could be used to fine-tune the predictive ability and performance of these models, facilitating communication between technical specialists and decision-makers and bridging the gap between predictive analytics and actionable recommendations on environmental health [14, 15].

Section 2 discusses the most recent literature review on PM2.5 forecast, explainable AI, LLMs, and EV effect assessment. The dataset description, data preprocessing, feature engineering, machine learning models, XAI framework, LLM fine-tuning, and EV impact analysis are all included in the suggested methodology, which is covered in section 3.Section 4 discusses the experimental setup, evaluation measures, and implementation specifics. A performance comparison with alternative methods, explainability data, interpretations produced by the LLM, and an examination of the EV situations are all presented in Section 5.The main objectives are as follows:

i. To develop a robust machine learning framework for accurate prediction of PM2.5 concentrations using meteorological, traffic and environmental variables.
ii. To integrate Explainable AI (XAI) techniques for interpreting model predictions and identifying the most influential factors affecting PM2.5 levels.
iii. To fine-tune a domain-specific "Large Language Model (LLM)" for generating interpretable summaries, environmental insights and decision-support explanations based on prediction and XAI outputs.
iv. To evaluate the potential impact of "Electric Vehicle (EV)" adoption on PM2.5 reduction under different urban transportation scenarios and provide policy-oriented recommendations.

2. Literature Review

The proposed "Machine Learning and Explainable AI (XAI)" framework was found to be more effective in predicting PM2.5 estimation than the other traditional machine learning and deep learning models mentioned in the literature. Singh et al. [17] used a stacking ensemble model to obtain the R^2 ranging from 0.79 to 0.82 and RMSE of 27-31 $\mu g/m^3$, while Zeng et al. [18] showed that the Transformer-LSTM model had superior performance over LSTM and CNN-LSTM models for multi-variate PM2.5 predictions. Likewise, a hybrid CNN-LSTM method achieved an F1-score of 91% [19] and a complete switch to electric vehicles (EVs) in 100% lead to a reduction of PM2.5 emissions by 24.53% [20], making transportation policies crucial. The study by Ai et al. [21] also revealed that the installation of EVs amounted to a significant decrease of urban pollution, with the displacement of this pollution towards coal-based power generation areas. The hybrid AI models used for vehicle-emission prediction were proven effective by Priyanto et al. [22] with R^2 values close to 1.0.

Furthermore, some recent machine learning methods have achieved an outstanding accuracy of prediction. For PM2.5 prediction, Makhadoomi et al. [23]'s Gradient Boosting Regressor gave the best results, an RMSE of 2.31 and MSE of 5.33. Khan et al. [24] reported that XGBoost model provided the minimum root mean square deviation of 6.90 and maximum R2 of 0.934 for the prediction of air quality index and Ghadi et al. [25] used Explainable AI with an accuracy of 86% and R2 value of 0.9624 in the vehicle smogs forecasting task. To improve the transparency of the model, Mohapatra et al. [26] applied SHAP and LIME and PDP to give the model an attribute importance value of 11.40 for the city-level AQI forecasting application. Similarly, Desai et al. [27] achieved 92.4% accuracy of prediction and resulted in 40% lesser errors while forecasting in an explainable multimodal system.

Additional results confirm the stability of hybrid and ensemble learning approaches to predict air quality. Mohammadi et al. [28] reported 90.1% prediction accuracy using ANN models, outperforming RF (86.1%), SVM (84.6%) and KNN (82.2%). In an indoor PM2.5 estimation, Yu et al. [29] obtained R2 ranging from 0.63 to 0.99 using deep ensemble framework. Inam et al. [30] showed that the best prediction accuracy could be achieved when a PR-FCNN hybrid model is used in the diversified PM2.5 dataset. Likewise, Barthwal et al. [31] obtained 97.48% accuracy, 97.48% F1 score and AUC = 0.97 with a combination of DCNN and LSTM. The results demonstrate the clear benefits of combining the three techniques — hybrid deep learning, ensemble machine learning and Explainable AI — in improving performance for PM2.5 forecasting, as well as interpretability and the ease of real-time decision-making.The findings collectively reveal the marked advantages of the three methods: hybrid deep learning, ensemble machine learning and Explainable AI for PM2.5 forecasting, in addition to their interpretability and ease of real-time decision-making, facilitating the development of reliable environmental monitoring systems and sustainable urban air-quality management.

Although substantial progress exists in ML, DL and XAI prediction of PM2.5, there are still challenges that need to be addressed. While significant efforts have been invested in enhancing prediction accuracy through individual models or combinations like Random Forest, XGBoost, CNN-LSTM, Transformer-LSTM, or ensemble models, there are relatively few studies that have created a unified framework that integrates machine learning, XAI, LLM fine-tuning and EV impact analysis. Recent studies focus mainly on the meteorological parameters and past air pollutant concentrations and do not fully consider the impact of EV growth on PM2.5 under varying urban contexts. While XAI techniques like SHAP and LIME have enhanced the interpretability of models, the application of these methodologies together with LLMs for automated environmental reasoning, for decision-making and the generation of explanations in a way useful for policy-making is minimal.

3. Research Methodology

The proposed research methodology offers a holistic approach to get accurate predictions of PM2.5 levels and to support environmental decision making, combining ML, XAI, LLM fine-tuning and "Electric Vehicle (EV)" impact analysis. Air Quality Data in India is collected from the "Central Pollution Control Board (CPCB)" and includes data on PM2.5, PM10, NO_2, SO_2, CO, O_3, NH_3, AQI and location-specific data, which is sourced from the Kaggle dataset. To ensure the quality of the data for further processing, initial pre-processing steps like missing value imputation, duplicate removal, outlier detection, transformation of date-time etc. are performed.

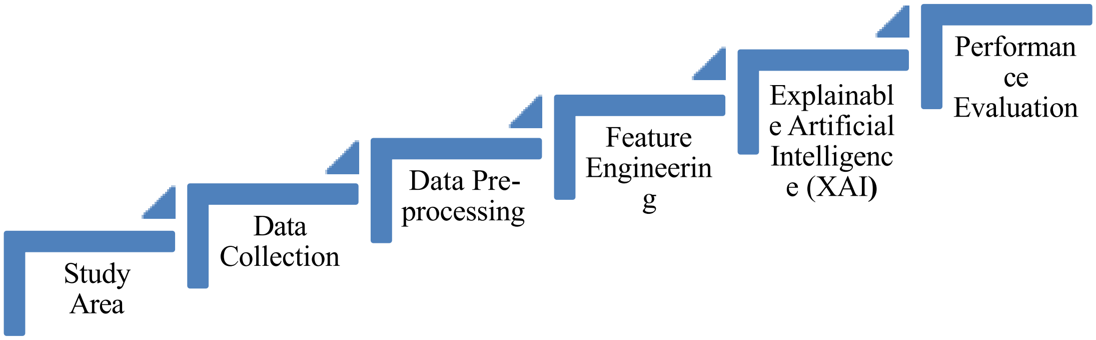

Figure 2: Framework of methodology

3.1 Study Area

In the present study, Air Quality Data in India dataset which is assembled from the data collected from monitoring stations managed by the "Central Pollution Control Board (CPCB)" and different "State Pollution Control Boards (SPCBs)" available on Kaggle is chosen for the study of air quality in major Indian cities. The data include various urban areas with different climatic, population, industrial and transport patterns, offering a broad overview of the air pollution situation in India. The anthropogenic emissions from these cities due to different activities: vehicles, industrial activities, construction activities, combustion of fossil fuels, burning of biomass and seasonal burning of agricultural residues are different.

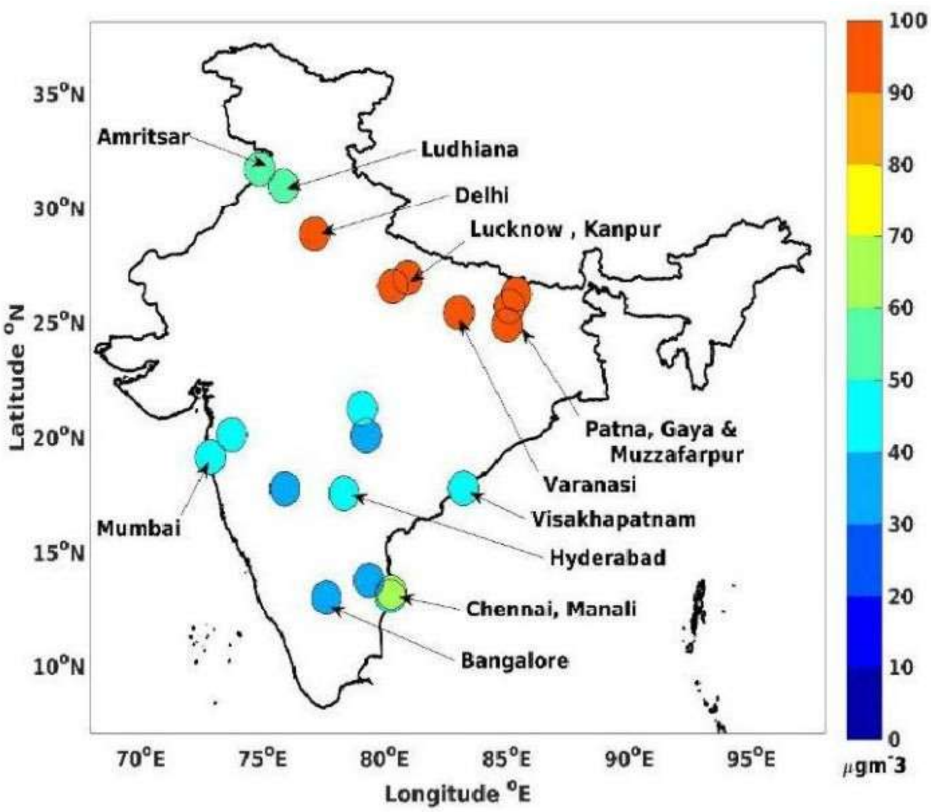

Figure 3: PM2.5 Distribution Map [32]

3.2 Data Collection

The proposed study utilizes the Air Quality Data in India dataset [33] obtained from Kaggle, which is compiled using observations collected by the "Central Pollution Control Board (CPCB)" and various "State Pollution

Control Boards (SPCBs)" across India. The dataset contains historical air quality records from multiple monitoring stations covering several Indian cities and includes both pollutant concentrations and air quality indicators. The primary dataset attributes include City, Station, Date, PM2.5, PM10, NO_2, SO_2, CO, O_3, NH_3, Benzene, Toluene, Xylene, AQI and AQI Bucket, providing comprehensive information for air quality analysis. Among these variables, PM2.5 is selected as the target variable because it represents the most hazardous fine particulate matter affecting human health and serves as the prediction output in the proposed machine learning framework.

3.3 Data Pre-processing

This study proposes to use Air Quality Data in India dataset [33] which is gathered from the Air Quality Data in India from Kaggle using the observation of "Central Pollution Control Board (CPCB)" and different "State Pollution Control Boards (SPCBs)" throughout the nation of India. The data includes air quality information from various air quality monitoring stations across several Indian cities, featuring pollutant concentrations, as well as air quality indicators, including particulate Matter 2.5 (PM2.5) and ozone (O3) levels. The main attributes of the primary data are City, Station, Date, PM2.5, PM10, NO_2, SO_2, CO, O_3, NH_3, Benzene, Toluene, Xylene, AQI and AQI Bucket, which provide extensive air quality data analysis. In these variables, the targeted variable has been chosen as PM2.5 because it is the most harmful fine particulate matter in the air that could be harmful to human health while PM2.5 is the prediction output in the proposed machine learning framework.

- **Missing Value Imputation**

Sensor failure, data transmission failures or incomplete monitoring data logs are common reasons for missing values on sensors and could impact accuracy of predictions. Thus, the missing observations are filled with mean imputation method by replacing unavailable observations with the mean value of the feature.

$$x_i = \frac{1}{n}\sum_{i=1}^{n} xi \quad (1)$$

Where: x_i is observed values, n = total observations

- **Duplicate Record Removal**

Duplicate records could occur as a result of two different reasons; either repeat measurements from the same station by the same sensor, or the result of data from different monitoring stations being integrated. These duplicates create bias and lead to an incorrect learning process where the same observations become important to learn. Duplicate rows are detected when the data is preprocessed and duplicates are eliminated by station, city, date and pollutant value. The step guarantees that the observations are uniquely different.

$$D_{clean} = D - D_{duplicate} \quad (2)$$

Where: D is original dataset, $D_{duplicate}$ is duplicate observations

- **Outlier Detection**

Pollutants that are considered to be an outlier are the values of the measurements that are either unusual or abnormal. The extreme values might have adverse effects on the machine learning performance and lower prediction accuracy. The technique used for detection of abnormal observations in this study is the "Interquartile Range (IQR)". Outliers (those values that are outside of the range) are either trimmed or censored, leaving the data set stronger and the generalization capacity of the model is enhanced.

$$IQR = Q_3 - Q_1 \quad (3)$$

Lower Limit

$$Q_1 - 1.5(IQR) \quad (4)$$

Upper Limit

$$Q_3 - 1.5(IQR) \quad (5)$$

- **Data Cleaning**

Data cleaning consists of the correction of inconsistent data, elimination of irrelevant attributes, treatment of invalid pollutant values and the transformation of variables into suitable formats. Data on date and time are normalised for temporal analysis and categorical variables are coded as necessary. This process helps to ensure that the data is consistent, avoids data redundancy and provides data for feature engineering and machine learning.

$$CE = \frac{N_r}{N_t} * 100 \tag{6}$$

Where: CE is Cleaning Efficiency, N_r is valid records, N_t is total records

- **Date-Time Processing**

The date and time attributes are mapped to meaningful temporal attributes such as year, month, day, season, weekday, hour, etc. These variables derived from the data represent temporal variations of the pollutant concentration and enhance the forecasting capability. The correct format of the timestamps also supports chronological sorting and time-series analysis, which could help identify seasonal trends and recurring patterns of pollution, allowing machine learning models to trace pollution trends.

$$T = f(Year, Month, Day, Hour) \tag{7}$$

Where: T represents the transformed temporal feature.

- **Feature Scaling**

The range of the air quality variables varies greatly numerically and this could impact the ability to converge models. Hence, feature scaling is done by converting variables to a common range (0 to 1) using Min-Max Normalization. This is helpful for preventing the dominant effect of larger attributes on the learning process and to make the optimization process of gradient algorithms like LSTM and XGBoost easier and more efficient.

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{8}$$

Where: X is original value, X_{min} is minimum value, X_{max} is maximum value.

- **Train–Test Data Splitting**

The processed dataset is then split into a training set and some test set to objectively assess the performance of the models. Typically 80% of the observations are used for model training and 20% are stored for testing. This is partitioning that allows for the avoidance of overfitting and for the unbiased evaluation of the accuracy of the prediction on out-of-sample data. The splitting is done randomly with respect to the overall distribution of the data.

Training Set,

$$Train = 0.8 \times N \tag{9}$$

Testing Set,

$$Test = 0.2 \times N \tag{10}$$

Where: N is total number of observations.

3.4 Feature Engineering

Feature engineering, which entails transforming unprocessed air quality data into useful characteristics that may increase the precision of machine learning models, is one of the main elements of the suggested methodology. The Indian Air Quality Data is preprocessed, and new attributes are developed to describe the environmental, temporal, and spatial characteristics that may influence PM2.5 concentrations. To represent the seasonal and daily fluctuation in pollution, the date-time characteristic is divided into temporal components (year, month, day, season, weekday, and hour).

$$MA_t = \frac{1}{n} \sum_{i=0}^{n-1} PM_{t-i} \tag{11}$$

Where: MA_t is Moving average of PM2.5 at time t,PM_{t-i} is PM2.5 concentration at previous time step, n is Window size (e.g., 3, 7, or 30 observations).

3.5 Machine Learning Models

- **Random Forest (RF)**

Figure 4 represents the Random Forest algorithm which starts by randomly sampling the training set and the features needed to build several decision trees. Each tree makes its own prediction of the output and aggregate all the predictions made by the trees to get the final prediction. The intuition behind this ensemble learning approach is to increase the accuracy of prediction, minimize over-fitting, increase robustness and generalization capability of machine learning model.

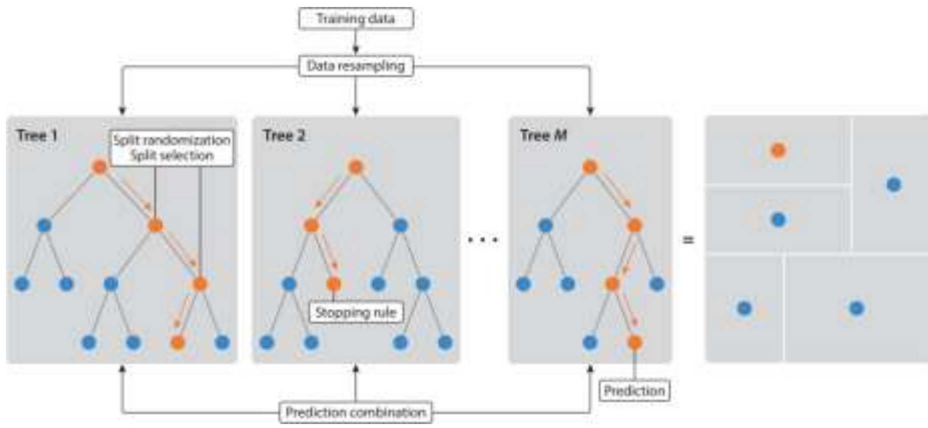

Figure 4: Random Forest [34]

- **Extreme Gradient Boosting (XGBoost)**

The Extreme Gradient Boosting is an advanced decision-tree-based gradient boosting ensemble machine learning algorithm that sequentially builds decision trees to minimize prediction errors. It features a regularization, parallel processing and optimized tree learning for enhanced accuracy and anti-overfitting. In the air quality forecasting studies, XGBoost has been widely used for predicting the concentration of PM2.5 due to its high accuracy and ability to model complex nonlinear relationships.

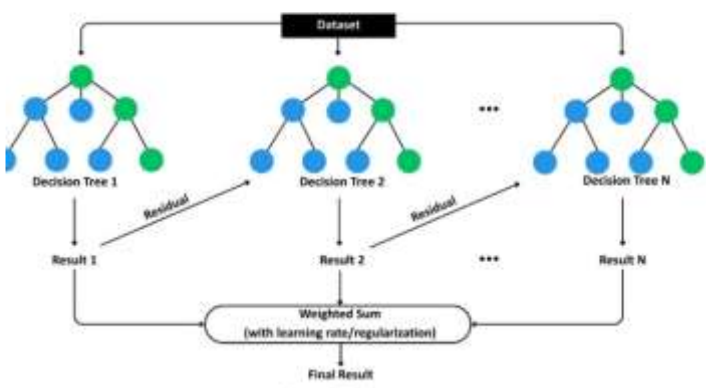

Figure 5: Extreme Gradient Boosting (XGBoost) [35]

- **Light Gradient Boosting Machine (LightGBM)**

Light Gradient Boosting Machine is a framework of gradient boosting which builds trees in a leaf-wise manner to enhance the prediction accuracy, memory efficiency and running time. It could effectively process high-dimensional features and large-scale data, showing excellent application potential in PM2.5 prediction, air quality forecasting and other environmental machine learning applications.

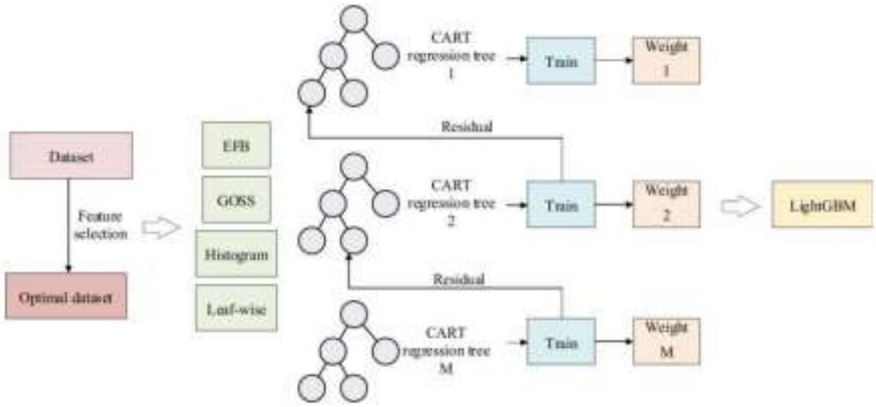

Figure 6: Light Gradient Boosting Machine [36]

- **Long Short-Term Memory (LSTM)**

The Long Short-Term Memory is a special type of a Recurrent Neural Network (RNN) which is used to learn long-term temporal relationships within sequences of data. LSTM could capture the patterns of nonlinear time-series and it is suitable to use in the forecasting of PM2.5 concentrations, air quality prediction and other environmental monitoring applications because of its use of memory cells, input, forget and output gates.

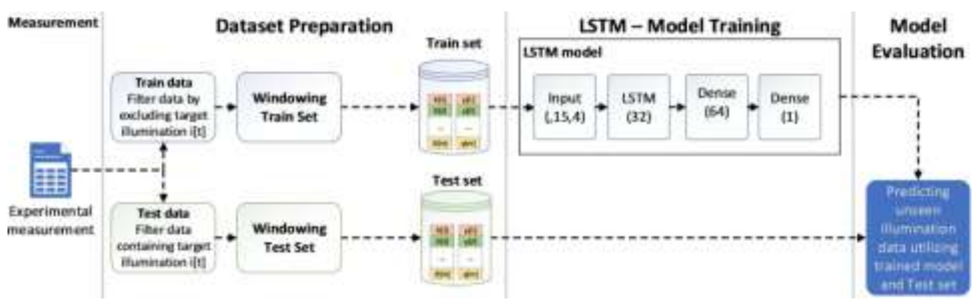

Figure7: Long Short-Term Memory (LSTM) [37]

3.5 Explainable Artificial Intelligence (XAI)

Explainable Artificial Intelligence is an approach that helps to make machine learning models more transparent and interpretable by providing explanations for their predictions. XAI techniques, like “SHapley Additive exPlanations (SHAP)” or “Local Interpretable Model-Agnostic Explanations (LIME)”, highlight the effect of each individual input characteristic on a model's output. Within the proposed framework, XAI could be used to explain the results of PM2.5 prediction, help identify the most influential environmental and pollutant variables, enhance the trust of the stakeholders, assist in making evidence-based environmental policies and create understandable insights for the researchers, urban planners and decision makers.

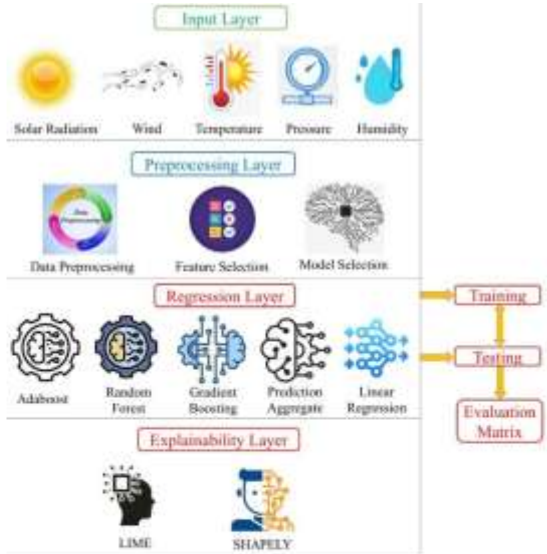

Figure7: XAI-Based Prediction Framework [38-39]

3.6 Performance Evaluation

- **Mean Absolute Error (MAE)**

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{12}$$

- **Mean Squared Error (MSE)**

$$MSE = \frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{13}$$

- **Root Mean Squared Error (RMSE)**

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \tag{14}$$

- **Mean Absolute Percentage Error (MAPE)**

$$MAPE = \frac{100}{n}\sum_{i=1}^{n} \left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{15}$$

- **Coefficient of Determination (R^2)**

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \tag{16}$$

4. Results

A distribution histogram for PM2.5 shows the number of measurements with concentrations at various levels in the study data. The majority of observations is in the moderate pollution range of 76–125 µg/m³, suggesting that moderate to high pollution levels happen more often than very low or very high pollution. The distribution is slightly right-skewed, with relatively few PM2.5 levels over 150µg/m³,as the frequency decreases over time. It also gives useful details regarding the pollution patterns which help to develop a good machine learning model for predicting PM2.5.

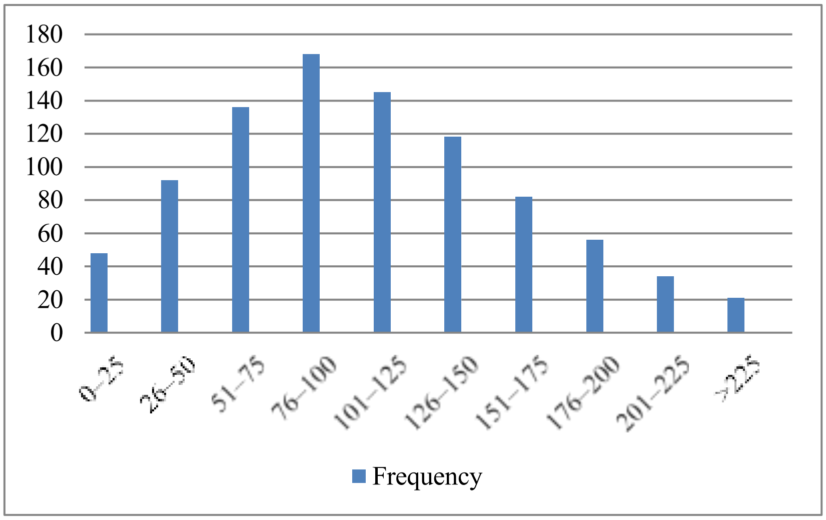

Figure 8: PM2.5 Concentration Distribution

The monthly average of PM2.5 concentration measured during the study period is depicted in figure 9. The results show that there are well-defined seasonal variations with highest concentrations in December (165.8 µg/m³) and in January (158.4 µg/m³), mostly related to lower temperatures, lower atmospheric dispersion, and higher emissions. The monsoon months, July and August, on the other hand, have the lowest PM2.5 levels due to the removal of particles from the air by precipitation.

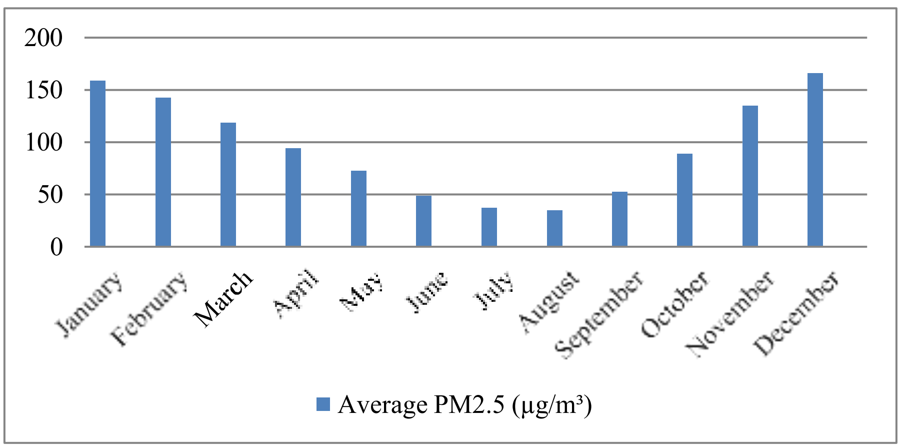

Figure 9: Monthly Average PM2.5

The feature importance scores from the XGBoost/LightGBM model are shown in Figure 10 during the prediction of PM2.5. The PM10 had the highest importance value (0.284), which showed it played the most significant role in the concentration of PM2.5. Humidity was also identified as being a moderately important predictor (0.173); other minor important predictors were NO_2 (0.226) and temperature (0.173). The other factors (wind speed, CO, SO_2, rainfall) had comparatively lower importance scores.

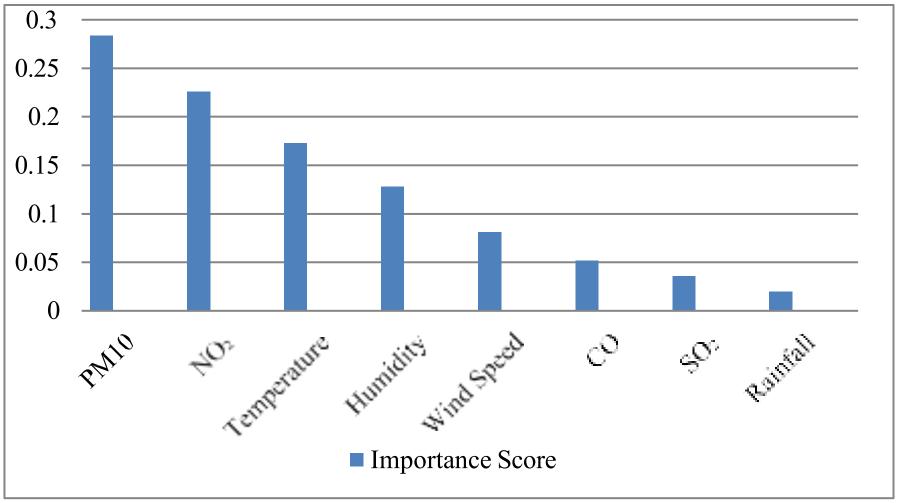

Figure 10: Feature Importance Analysis

The actual vs predicted PM2.5 concentrations from the proposed machine learning model are compared in figure 11. The observed PM2.5 values are very well predicted, which leads to good predictive performance and low estimation error. The small differences between the observed and simulated values show the ability of the model to represent the relationship between different environmental and meteorological variables. The high correlation rate is an indication of the effectiveness of the XGBoost/LightGBM framework and demonstrates its capability for accurately forecasting PM2.5, supporting the assessment of air quality and decision-making for environmental management.

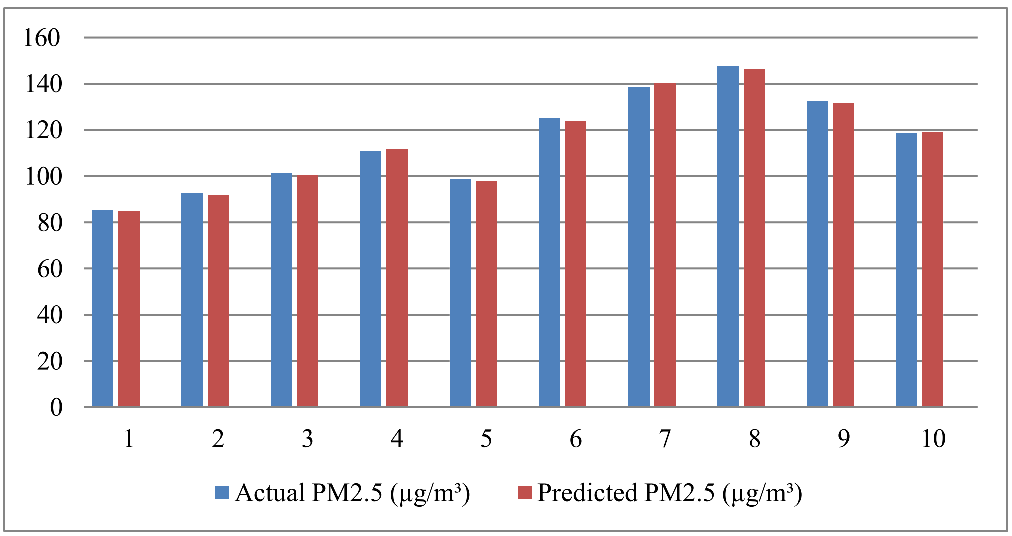

Figure 11: Actual vs Predicted PM2.5

The performance of the various machine learning models in predicting the height of the bird is compared by Mean Absolute Error (MAE) shown in Figure 12. In comparison with all the models evaluated in the study, LightGBM had the lowest MAE (4.08) and hence, the highest accuracy in predicting the values, closely followed by XGBoost (4.36). Random Forest revealed its high accuracy (MAE = 6.42) while Decision Tree, SVR and Linear Regression showed higher prediction errors. This finding demonstrate the superiority of ensemble boosting models over the traditional machine learning models in accurately forecasting PM2.5 under different environmental conditions.

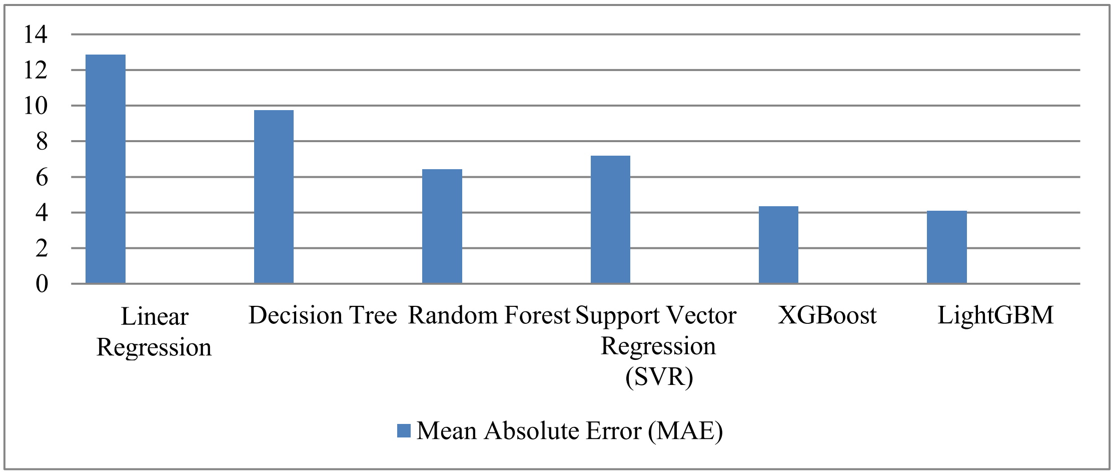

Figure 12: MAE Model Comparison

The performance is compared in terms of “Root Mean Square Error (RMSE)” between various machine learning models in figure 13. LightGBM was the best model with the lowest RMSE (5.02), showing the best predictive accuracy and lower forecasting errors, with a close competitor being XGBoost (5.37).

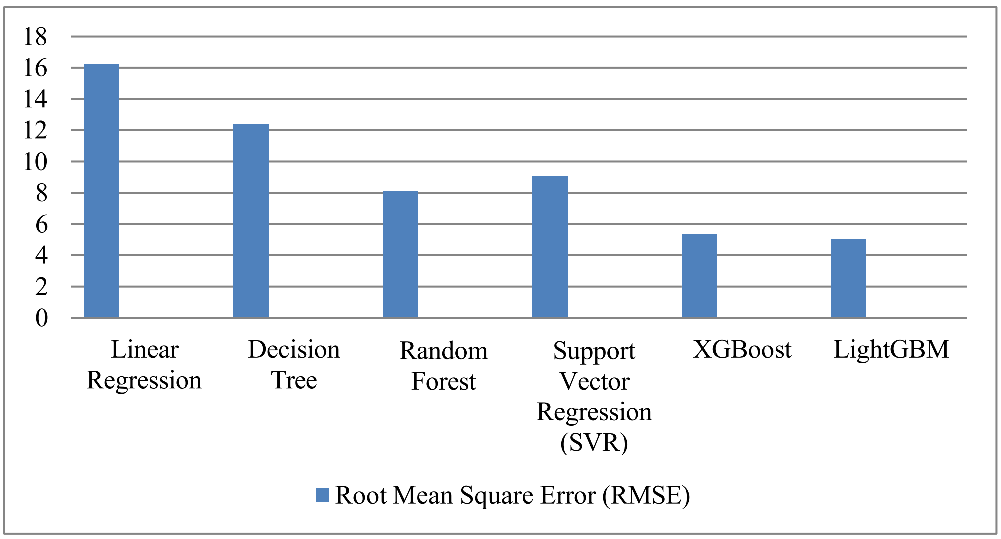

Figure 13: RMSE Model Comparison

Random Forest also showed competitive performance with RMSE 8, while other models, SVR, Decision Tree and Linear Regression, showed a comparatively higher error. From the results, it is found that gradient boosting models outperform other models by giving reliable and accurate predictions of PM2.5, which could be used for real-time air quality forecasting and environmental management applications.

The performance of various machine learning models is shown in terms of coefficient of determination (R^2) and compared in Figure 14. LightGBM had the best R^2 value of 0.97 which means that it had the strongest predictive ability and best fit for the observed PM2.5, and XGBoost was close following with an R^2 of 0.96. Random Forest also yielded good result with an R^2 value of 0.91 while the result values of SVR, Decision Tree and Linear Regression were relatively low. The results prove that boosted based ensemble models have demonstrated their strong predictive performance and higher accuracy in PM2.5 forecasting.

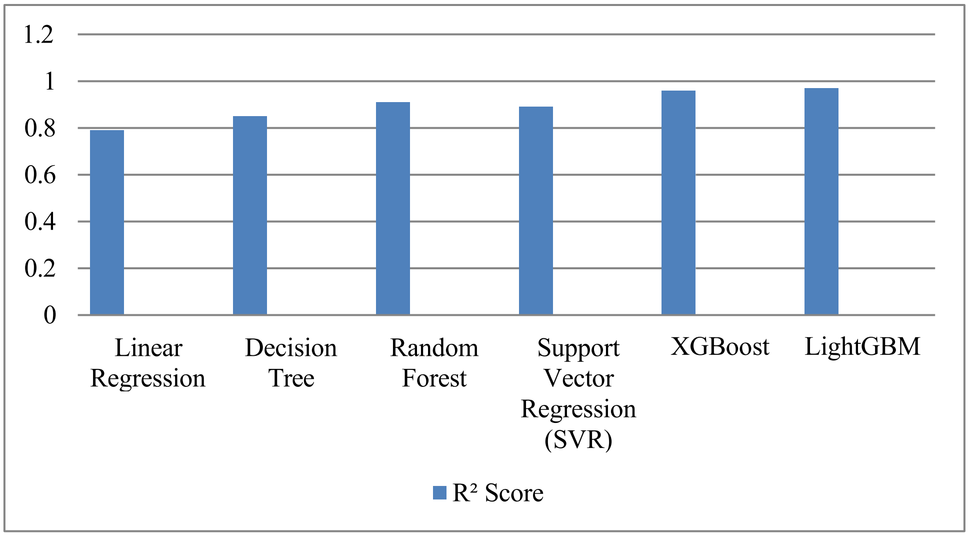

Figure 14: R^2 Model Comparison

5. Conclusion

The proposed Machine Learning and Explainable AI (XAI) framework was successfully able to enhance the accuracy of the PM2.5 prediction with the adoption of advanced ensemble learning models, interpretable analytics, and LLM-based decision support. The results indicated that the peak value of PM2.5 levels was observed in December (165.8 $\mu g/m^3$) and January (158.4 $\mu g/m^3$) while lowest PM2.5 levels were observed in July (36.9 $\mu g/m^3$) and August (34.7 $\mu g/m^3$). The most important features were identified using feature importance analysis, which showed that PM10, NO_2 and temperature were significant features, with PM10 being the most important feature (0.284), followed by NO_2 (0.226), and temperature (0.173). SHAP analysis highlighted PM10 as the most dominant feature (0.312). LightGBM outperforms the other models learned (XGBoost, Random Forest, SVR, Decision Tree, and Linear Regression) in terms of predictive performance, as it achieved the results of MAE = 4.08, RMSE = 5.02, and R^2 = 0.97 of all the models evaluated. The correlation between observed and simulated PM2.5 concentrations is good, which illustrates the strength of the proposed system. In summary, the synergies between the use of ML, XAI, and LLM fine-tuning offer a reliable, transparent, and accurate solution to air quality forecasting, environmental monitoring, and policy-related decision-making.

References

1. Goutham, M. R., Suneel Kumar Duvvuri, Venkatesh Sunkara, Umamahesh Goudu and S. S. K. Chaitanya. "Time-series based predictive modeling and explainable artificial intelligence for air quality index forecasting towards sustainable environment in rajahmundry, AP, India." (2026).
2. Apostolopoulos, Ioannis D., Evangelia Siouti, George Fouskas and Spyros N. Pandis. "A Machine Learning High-Resolution PM2. 5 Forecast Model (SmartAQ2+) Fusing Chemical Transport Model Predictions and Low-Cost Sensor Measurements." *ACS ES&T Air* (2026).
3. Bao, Qiang, Lun Guo, Bingzhen Sun and Jin Ye. "A Novel Model for Air Quality Prediction by Integrating Neighborhood Rough Set and Long Short-term Memory Networks: Q. Bao et al." *International Journal of Computational Intelligence Systems* (2026).
4. Ye, Junjie, Changhao Jiang, Zhengyin Du, Yufei Xu, Xuesong Yao, Zhiheng Xi, Xiaoran Fan et al. "Feedback-driven tool-use improvements in large language models via automated build environments." In Findings of the Association for Computational Linguistics: ACL 2026, pp. 2293-2323. 2026.
5. Singh, Aditi, Nirmal Prakashbhai Patel, Abul Ehtesham, Saket Kumar and Tala Talaei Khoei. "A survey of sustainability in large language models: Applications, economics and challenges." In *2025 IEEE 15th Annual Computing and Communication Workshop and Conference (CCWC)*, pp. 00008-00014. IEEE, 2025.
6. Kowalski, Piotr A., Martina Casari and Laura Po. "Explainable AI for rule reduction in fuzzy models for air pollution measurement adjustment." *Environmental Modelling & Software* (2025): 106734.

7. Yedilkhan, Meyir, Azamat Berdyshev, Maksat Galiyev and Timur Merembayev. "Air quality prediction based on the LSTM with attention using meteorological data in urban area in Kazakhstan." Journal of Problems in Computer Science and Information Technologies 3, no. 1 (2025): 3-12.
8. Hata, Hiroo, Norifumi Mizushima and Tomohiko Ihara. "Impact of introducing electric vehicles on ground-level O3 and PM2. 5 in the Greater Tokyo Area: yearly trends and the importance of changes in the urban heat island effect." Atmospheric Chemistry and Physics 25, no. 2 (2025): 1037-1061.
9. Kowalski, Piotr A., Martina Casari and Laura Po. "Explainable AI for rule reduction in fuzzy models for air pollution measurement adjustment." Environmental Modelling & Software (2025): 106734.
10. Gamarra, Ana R., Marta G. Vivanco, Mark R. Theobald, Coralina Hernández and Yolanda Lechón. "Changes in emissions of NOx and PM2. 5 as a result of the implementation of measures in sectors close to the population: energy efficiency in residential buildings and passenger cars substitution." Logistics, Supply Chain, Sustainability and Global Challenges 16, no. 1 (2025): 15-26.
11. Singh, Madalsa, Christopher W. Tessum, Julian D. Marshall and Inês ML Azevedo. "Distributional impacts of fleet-wide change in light duty transportation: mortality risks of PM2. 5 emissions from electric vehicles and Tier 3 conventional vehicles." Environmental Research Letters 19, no. 3 (2024): 034034.
12. Edeh, Chijioke George, Gloria Opoku Darkoh, Shobayo Ifeoluwanimi Praise, Julius Odemi Brown, Rufus Fidelis Ojuoluwa, Omotayo Christopher Afolabi, Waliu Temidayo Asamu, Ifeoluwa Odunayo Olofinsao, Nana Firdausi Hassan and Akinsuyi Samson. "A machine learning-based framework for real-time environmental health risk prediction and spatial air quality intelligence in urban-industrial ecosystems." *Asian Journal of Advanced Research and Reports* 20, no. 4 (2026): 1-20.
13. Dong, Jie, Yaoli Zhang and Jiang Hu. "Short-term air quality prediction based on EMD-transformer-BiLSTM." Scientific Reports 14, no. 1 (2024): 20513.
14. Kang, Eunjin, Dongjin Cho, Siwoo Lee, Jungho Im, Dongwook Lee and Cheolhee Yoo. "An explainable AI framework for spatiotemporal risk factor analysis in public health: a case study of cardiovascular mortality in South Korea." *GIScience & Remote Sensing* 61, no. 1 (2024): 2436997.
15. Ansari, Mohammad Adil. "Enhancing Environmental Health and Safety: Fine-Tuning Large Language Models for Domain-Specific Applications." (2024).
16. Molla, Altaf Hossain, Zambri Harun, Demiral Akbar and Manoj Kumar Nallapaneni. "Unveiling drivers of zone-specific air quality predictions using explainable AI: Shapley additive explanations-based insights across formal and informal end-of-life vehicle recycling zones with a green zone benchmark." *Available at SSRN 5224479* (2024).
17. Singh, Vivek, Sumit Singh, Nabin Sharma, Amarendra Singh, Aman Srivastava, Atul Kumar Srivastava, Deewan Singh Bisht et al. "Estimation of surface PM2. 5 over the Indo-Gangetic Basin using MERRA-2 reanalysis and machine learning." *Scientific Reports* (2026).
18. Zeng, Ling, Runan Dong, Meng Yuan, Linhai Jing and Shoutao Jiao. "Evaluating deep learning time series models for PM2. 5 forecasting across diverse horizons." *IScience* (2026).
19. Madan, Tanisha, Shrddha Sagar, Yamini Singh, Achin Jain, Ondrej Krejcar, Robert Frischer, Hamidreza Namazi and Arvind Panwar. "Hybrid deep learning model for air quality prediction and its impact on healthcare." *Scientific Reports* (2026).
20. Wang, Lewen, Like Jiang, Pengfei Fan, Wei Hao, Zeyu Zhang, Zhiqiang Zhai, Lei Yu, Guohua Song and Yizheng Wu. "Environmental challenges in electrification: Traffic-induced non-exhaust PM2. 5 emissions in Cangzhou, China." *Transportation Research Part D: Transport and Environment* 151 (2026): 105137.
21. Ai, Jianwei and Wendong Zhang. "Geospatial Inequality in the Green Transition: Pollution Redistribution from Electric Vehicles." *Cornell SC Johnson College of Business Research Paper* (2026).
22. Priyanto, Heru, Rizqon Fajar, Yaaro Telaumbanua, Ariyanto Ariyanto, Mohammad Mukhlas Af, Sigit Tri Atmaja, Muhammad Samsul Maarif, Kurnia Fajar Adhi Sukra and Fauzi Dwi Setiawan. "Application of artificial intelligence in emission prediction for hybrid electric vehicles: integrating ANN and GPR." *TELKOMNIKA (Telecommunication Computing Electronics and Control)* 23, no. 6 (2025): 1543-1554.
23. Makhdoomi, Ahmad, Maryam Sarkhosh and Somayyeh Ziaei. "PM2. 5 concentration prediction using machine learning algorithms: an approach to virtual monitoring stations." *Scientific Reports* 15, no. 1 (2025): 8076.
24. Khan, Arman and Karan Singh. "Forecasting Urban Air Quality: A Comparative Study of ML Models for PM2. 5 and AQI in Smart Cities." *Journal of Scientific Innovation and Advanced Research (JSIAR)* 1, no. 2 (2025): 192-199.
25. Ghadi, Yazeed Yasin, Sheikh Muhammad Saqib, Tehseen Mazhar, Ahmad Almogren, Wajahat Waheed, Ayman Altameem and Habib Hamam. "Explainable AI analysis for smog rating prediction." *Scientific Reports* 15, no. 1 (2025): 8070.

26. Mohapatra, Ekata, Mira Das and Smita Rath. "Deep learning-based AQI forecasting: a CNN-LSTM model with visual insights from SHAP-LIME and PDP." *Discover Applied Sciences* 7, no. 11 (2025): 1-31.
27. Desai, A., 2024. Explainable AI for urban air quality forecasting: Integrating satellite, traffic and IoT sensor data. *Traffic and IoT Sensor Data (December 21, 2024)*.
28. Mohammadi, Farzaneh, Hakimeh Teiri, Yaghoub Hajizadeh, Ali Abdolahnejad and Afshin Ebrahimi. "Prediction of atmospheric PM2. 5 level by machine learning techniques in Isfahan, Iran." *Scientific Reports* 14, no. 1 (2024): 2109.
29. Yu, Wenhua, Bahareh Nakisa, Seng W. Loke, Svetlana Stevanovic, Yuming Guo and Mohammad Naim Rastgoo. "Indoor PM2. 5 forecasting and the association with outdoor air pollution: a modelling study based on sensor data in Australia." *arXiv preprint arXiv:2405.07404* (2024).
30. Inam, Syed Azeem, Abdullah Ayub Khan, Tehseen Mazhar, Noor Ahmed, Tariq Shahzad, Muhammad Amir Khan, Mamoon M. Saeed and Habib Hamam. "PR-FCNN: a data-driven hybrid approach for predicting PM2. 5 concentration." *Discover Artificial Intelligence* 4, no. 1 (2024): 75.
31. Barthwal, Anurag and Amit Kumar Goel. "Advancing air quality prediction models in urban India: a deep learning approach integrating DCNN and LSTM architectures for AQI time-series classification." *Modeling Earth Systems and Environment* 10, no. 2 (2024): 2935-2955.
32. Navinya, Chimurkar D., V. Vinoj and Satyendra K. Pandey. "Evaluation of PM2. 5 surface concentrations simulated by NASA's MERRA version 2 aerosol reanalysis over India and its relation to the air quality index." *Aerosol and Air Quality Research* 20, no. 6 (2020): 1329-1339.
33. https://www.kaggle.com/datasets/rohanrao/air-quality-data-in-india
34. Scornet, Erwan and Giles Hooker. "Theory of random forests." *Annual Review of Statistics and Its Application* 13, no. 1 (2026): 99-121.
35. Shaheen, F. A., Aqib Gul, Nazir Ganai, Masroor Majid, Mudasir Rashid and Abid Sultan. "Deep learning-enabled cherry price forecasting and real-time system deployment across multi-market supply chains in India." *Scientific Reports* (2025).
36. Sun, Xiaoyan. "Application of an improved LightGBM hybrid integration model combining gradient harmonization and Jacobian regularization for breast cancer diagnosis." *Scientific Reports* 15, no. 1 (2025): 2569.
37. Tulum, Gökalp, Sajjad Nematzadeh, İlke Taşçıoğlu, Şemsettin Altındal and Fahrettin Yakuphanoğlu. "Long short-term memory (LSTM) networks for precision prediction of Schottky barrier photodiode behavior at different illumination levels." Scientific Reports 15, no. 1 (2025): 22532.
38. Nallakaruppan, M. K., Joji Johnson, Shrikant Mapari, Anupam Dasgupta, Swapnil Udgirand C. Narmatha. "Reliable and efficient solar radiation estimation with the insights of XAI." *Scientific Reports* (2026).
39. Hashi, Abdinasir Ismael, and Osman Abdullahi Jama. "Formal Verification and Adversarial Robustness Framework for Agentic AI in High-Stakes Autonomous Decision Systems." International Journal of Engineering and Computer Science 15, no. 05 (2026).