A HYBRID APPROACH FOR PREDICTION OF AIR QUALITY INDEX USING MACHINE LEARNING

Sarvesh Garg[1], Dr. Kiran Jyoti[2], Dr. Hardeep Singh Kang[3]

[1]Student & 1765, Phase-1, Urban Estate, Dugri. Ludhiana
[2] Professor, Dept. of Computer science Engineering, GNDEC, Punjab, India
[3] Associate Professor, Dept. of Computer Science Engineering, GNDEC, Punjab, India

---

Abstract - *Air pollution has become one of the most critical environmental and public health concerns worldwide due to its severe impact on human health, ecosystems, and climate conditions. Rapid urbanization, industrialization, and increasing vehicular emissions have significantly contributed to the deterioration of air quality, making accurate forecasting of air pollution levels essential for effective environmental management and policy planning. Predicting the Air Quality Index (AQI) with high precision is challenging because air pollution data exhibit both nonlinear patterns and temporal dependencies influenced by meteorological conditions, pollutant interactions, and seasonal variations. Traditional statistical models often fail to capture nonlinear relationships effectively, whereas standalone machine learning models may not fully address time-series dependencies and residual errors. To overcome these limitations, this research proposes a novel hybrid machine learning framework for air quality prediction by integrating Random Forest (RF), Autoregressive Integrated Moving Average (ARIMA), and Gated Recurrent Unit (GRU) models. The proposed methodology combines the strengths of these three models to improve prediction accuracy and robustness. Initially, the Random Forest algorithm is employed as the primary predictive model due to its strong capability in handling nonlinear relationships, feature interactions, and high-dimensional environmental datasets.*

Key Words: Deep Learning, Gated Recurrent Unit, Elastic Net, Machine Learning, Air Quality.

1. INTRODUCTION

Air quality has become a critical environmental and public health concern due to rapid urbanization, industrialization, and increased vehicular emissions. Poor air quality, characterized by high concentrations of pollutants such as $PM_{2.5}$, PM_{10}, NO_2, SO_2, CO, and O_3, has been directly linked to respiratory diseases, cardiovascular disorders, and premature mortality. To effectively mitigate these impacts, timely and accurate air quality prediction is essential. Traditional statistical and deterministic models, including linear regression and chemical transport models, have been widely used for air quality forecasting. However, these approaches often struggle to capture the highly nonlinear, dynamic, and complex interactions among meteorological variables, emission sources, and atmospheric processes. With the availability of large-scale air quality monitoring data and advances in computational power, data-driven methods have gained significant attention. Machine learning and deep learning models, such as artificial neural networks, convolutional neural networks, and recurrent neural networks, have demonstrated superior predictive capabilities compared to classical models. Nevertheless, single deep learning models may still face limitations related to feature extraction, temporal dependency modeling, spatial correlation handling, and generalization. These challenges have motivated the exploration of hybrid deep learning methods for air quality prediction.

2. LITERATURE REVIEW

S. et. al, 2025 said that the research employed an LSTM model to predict the Air Quality Index (AQI) from a three-year dataset of eight pollutants. The data were smoothed and normalized prior to training with look-back days as one of the important parameters. Experimental results had best performance with MAE 0.032 for training and 0.043 for test at 10 look-back days [1].

Khokhlov et. al, 2025 proposed that this research used ML and DL models to forecast AQI in Tashkent, Uzbekistan, one of the dirtiest cities in the world. Using KNN, RF, DT, SVM, and ANN using optimizers such as Adam, Ada Grad, RMSprop, and SGD with momentum, the problem was set as classification. Model performance was measured using Accuracy, Precision, Recall, and F1-Score, identifying strengths, weaknesses, and appropriate methods to predict AQI [2].

Au et. al, 2025 said that the research used Automated Machine Learning (Auto ML) on time- series IoT sensor data in air quality monitoring. Experiments revealed that Auto ML models had similar or better accuracy compared to customary and neural network algorithms such as Random Forests, Extra Trees, Elastic Net, and LSTM. The findings exhibited the usability and cross-industry applicability of ML-based air quality analysis without the need for specialized technical knowledge [3].

Kocak et, al, 2025 said that this research compared five machine learning models to predict hourly PM2.5 and PM10 concentration based on actual pollutant and meteorological data. Varied performance was observed, with the best performance in PM2.5 prediction by SVR (R^2=0.83), followed by PM10 prediction by Ridge Regression (R^2=0.91), and RF, Extra Trees, and XG Boost with strong overall results. Based on the Air Q+ model, long-term PM2.5 exposure was associated with an average attributable mortality proportion of 10.2%, highlighting the necessity for focused interventions [4].

S. Thanya Charoen, 2025 said that in this research, ML and DL models were compared for air quality classification with 5,000 samples from Bangladesh's Dhaka city. Random Forest was the most accurate ML model (96%), followed by VGG19 as the best DL model with 95%, accurately representing intricate patterns. Trade-offs between performance and efficiency were highlighted by the findings, necessitating scalable, diverse datasets and hybrid methods for effective prediction [5].

Bhaskar et. al, 2025 suggested that this review looked at how intense urbanization and a 30% increase in cars in the last decade amplified air pollution and health hazards in Indian cities. It reviewed the shortcomings of current air quality prediction techniques and looked at how ML, DL, and hybrid models may improve forecasting. The paper also touched on challenges such as faulty sensors, excessive computational time, and interpretability, while proposing AI-based predictions and policy reforms for improved air quality management [6].

Abdelmalek et. al, 2025 said that five ML models (GPR, ER, SVM, RT, and KAR) were used in this study to forecast AQI with prominent pollutants (PM2.5, PM10, and CO) from an actual IoT real-time data set. The results indicated that GPR, ER, SVM, and RT performed above 96%, whereas KAR ranked lower at 82.36%. GPR exceeded all models with minimum RMSE, showing the prospect of ML to accurately control real-time air quality [7].

Mohanraj et. al, 2025 said that the four models, XG Boost, LSTM, Prophet, and Vertex AI, were tested by this research for AQI forecasting and monitoring. Findings indicated that sophisticated tools increased accuracy and convenience to the point where ML applications were accessible even to laymen. The research emphasized the viability of hybrid approaches that combine local tools with cloud platforms for better AQI forecasting [8].

Fang et. al,2025 said that the research contrasted linear regression (LR) and random forest (RF) models for calibrating the Plan tower PMS 3003 air quality sensor in different environmental settings. LR exhibited stable performance for low to moderate $PM_{2.5}$ concentrations with low computational cost, whereas RF found calibration accuracy improved in difficult, humid conditions. [9].

Magesh et. al, 2025 proposed that the research examined the relationship between poverty and air pollution in the contiguous United States through machine learning models such as linear regression, decision trees, and neural networks. Findings indicated that socioeconomic indicators correctly predicted rates of poverty, yet poverty itself did not relate to pollution levels to any notable degree.

Yasin et. al, 2025 said that geospatial modeling coupled with machine learning was applied in the PM_{10} prediction, where the random forest (RF) algorithm was the most accurate in hotspot identification. Spatial analysis showed hotspots of PM_{10} 8–16.5 km east and northwest of Addis Ababa city center. [11].

Nataliya et. al, 2025 suggested that the research built a surrogate modeling methodology to forecast indoor air temperature (IAT) based on an augmented LSTM combined with Rolling Window Cross-Validation (RWCV) to identify temporal dependencies. The model exhibited robust generalization, showing little difference in loss between training and testing, and enhanced adaptability to varying conditions. [12].

Jayaraman et. al., 2025 said that the research proposed a Time-Based-Spatial (TBS) forecasting model combining CNNs and ARIMA models for forecasting the Air Quality Index (AQI) in multiple cities. CNNs extracted spatial relationships based on geographic coordinates, and ARIMA estimated temporal pollutant trends. The combined model produced accurate six-hour AQI forecasts with improved insights into spatial and temporal aspects affecting air quality [13].

Rad et. al, 2025 said that the research projected concentrations of CO, O_3, NO_2, SO_2, PM_{10}, in Tehran (2013–2023) with deep learning (DL) models and contrasted them with conventional ML techniques. GRU and FCNN models attained the greatest accuracy, with R^2 values up to 0.93 and minimal MSEs, surpassing conventional methods. Temperature and humidity were determined as drivers of pollutant variability, validating DL models as effective tools for air quality management [14].

Ajala et. al, 2025 said that the research assessed 14 models, which were statistical, machine learning (ML), and deep learning (DL) models, for forecasting CO_2 emissions on a daily basis for major polluting areas from the year 2022 to 2023. ML and DL models compared to statistical models had greater R^2 (0.714–0.932) and smaller RMSE (0.247–0.480) values. Ensemble ML algorithms such as bagging and voting enhanced accuracy by 9.6% and represented the most efficient option for the forecasting of CO_2 emissions on a daily basis [15].

Zupek et. al, 2025 said that the research examined 9,357 hourly air pollutant records from the Kaggle Air Quality Dataset through ten machine learning regression models. Model accuracy was enhanced through Bayesian optimization, cross-validation, and stacking, with the best performance obtained by optimized SVR (R^2 = 99.94%, MAE = 0.0120, MSE = 0.0005). The results validated the suitability of ensemble and optimization methods in modeling air pollution processes with a high degree of accuracy [16].

Kirti et. al, 2025 said that the research estimated spatial evapotranspiration (ET) variability in northeastern Madagascar based on ECOSTRESS data and machine learning algorithms. Random forest was highly accurate (R^2 = 0.7–0.9), where the most critical variables were meteorological and vegetation structure. SHAP analysis revealed that ET patterns were strongly influenced by vegetation structure under different rainfall and wind conditions [17].

Nguyen et. al, 2025 said that the research created a hybrid deep learning model that integrated ACNN, ARIMA, QPSO-optimized LSTM, and XG Boost to forecast the Air Quality Index (AQI) from Seoul's 2021–2022 air quality data. ARIMA utilized linear trends, while the hybrid structure dealt with nonlinear trends via convolution, optimization, and tuning. The model had a 31.13% decrease in MSE and a 19.03% decrease in MAE compared to traditional methods and exhibited greater accuracy and stability [18].

Trenk et. al, 2025 said that in this research, daily Air Quality Index (AQI) forecasting in eastern Türkiye was evaluated between 2016 and 2024 based on machine learning models trained with pollutant and meteorological data. Out of XG Boost, Light GBM, and SVM, the highest performance was attained by XG Boost (R^2 = 0.999, RMSE = 0.234, MAE = 0.158). The results indicated that ensemble-based models managed to capture AQI variations efficiently, providing insights of great importance for air quality management and policy-making [19].

Yeboah et. al, 2025 proposed that this research compared satellite-derived AOD with data from low-cost sensors to estimate PM2.5 concentrations in Accra, Ghana, using machine learning algorithms. Of the models assessed, Boosted Decision Tree performed optimally when temperature and humidity were used. Findings revealed AOD accounted for merely a fraction of ground-level PM2.5 variability, highlighting the requirement of locally derived correction algorithms to enhance the accuracy of low-cost sensors [20].

3. METHODOLOGY

1. Train Random Forest Regressor

In this step, the model is initialized with `RandomForestRegressor(random_state=42). Here, `random_state` is set to ensure reproducibility in the random selection process while creating multiple decision trees. Following that, the model learns from the training data through the `fit (X_train, y_train)` method, where `X_train` represents the predictors and `y_train` denotes the dependent variable. Once training is completed, the model makes predictions based on both training data (`X_train`) and testing data (`X_test`) using the `predict()` method. This yields two sets of predictions, namely `AQI_pred_rfr_train` and `AQI_pred_rfr_test`, which can be further analyzed for evaluating model performance. The completion of the training process is notified using the print function.

2. Computing Residuals

This step computes the errors (residuals) of the Random Forest Regression model on the training dataset. The errors are the differences between the actual value of the target variable (y_train) and the predicted value of the AQI (AQI_pred_rfr_train). It is calculated by using the equation: `residual_train = y_train - AQI_pred_rfr_train`. Each element of the residuals represents the amount by which the model prediction varies from the actual AQI level; any number greater than zero means under-prediction, while any negative number indicates over-prediction. Once the residuals have been computed, they are transformed into the form of a pandas series using `pd.Series()` method. This is essential because in this way we preserve the same index as for the initial training set (y_train.index); this helps us in aligning individual observations with their respective errors. In the end, there is the print command, confirming the successful computation and assignment of the residuals to the variable `residual_train`. This type of code is usually applied in cases when a hybrid modeling approach is utilized, in which residuals obtained with one model are then further analyzed and modeled with another one (ARIMA or deep learning techniques, for instance).

3. Train ARIMA Model on Residuals

This step applies an ARIMA (AutoRegressive Integrated Moving Average) model to the residuals obtained from the Random Forest model, forming a hybrid modeling approach. First, required libraries are imported, and warnings are suppressed for cleaner output. The ARIMA model is then initialized using ARIMA(residual_train, order=(5,1,0)), where (5,1,0) specifies the model parameters: 5 autoregressive terms (p), first-order differencing (d=1) to make the series stationary, and no moving average terms (q=0). The model is trained on the residual data using the .fit() method, allowing it to capture temporal patterns left unexplained by the Random Forest model. Next, the code forecasts future residual values corresponding to the size of the test dataset. The start_index_forecast is set to the length of the training residuals, and the end_index_forecast extends this range to match the number of test samples. The .predict() method generates these future residual estimates. Since the output is initially an array, it is converted into a pandas Series and aligned with the y_test index to maintain consistency with the test dataset's timeline. Finally, a print statement confirms that the ARIMA model has been successfully trained on residuals. This approach helps improve prediction accuracy by modeling patterns missed by the initial machine learning model.

4. Define and Train attention gru model

This step builds and trains an Attention-based GRU (Gated Recurrent Unit) deep learning model for time-series prediction using synthetic data. First, a function `create_attention_gru_model` is defined, where an input layer accepts sequences of shape `(timesteps, features)`. These inputs are passed through a GRU layer with `return_sequences=True`, meaning it outputs hidden states for every time step instead of just the final one. An Additive Attention layer is then applied, allowing the model to focus on the most important time steps by assigning weights to different parts of the sequence. The output of the attention layer is passed through a Dense layer with one unit to produce predictions. Next, synthetic time-series data is generated using sine waves and linear trends with noise, simulating realistic temporal patterns. The dataset is split into training and testing sets, and both input features and target values are scaled using MinMaxScaler for better neural network performance. The model is compiled using the Adam optimizer and Mean Squared Error loss, and trained for 35 epochs with validation. After training, predictions are made on the test set, and the results are inverse-transformed back to the original scale. Finally, the predicted values and actual test values are prepared for evaluation, completing the workflow of an attention-enhanced GRU forecasting model.

5. ARIMA Residuals computed and processed

This step continues the hybrid modeling pipeline by preparing the final residual component (after Random Forest and ARIMA) for training an Attention-based GRU model. First, in-sample predictions from the ARIMA model are generated using `predict()` over the training residual range, and their index is aligned with `residual_train` to ensure proper subtraction. The ARIMA forecasted residuals (`residual_forecast`) are then subtracted from these test residuals to obtain `residual_test_arima_for_gru`, capturing remaining unexplained patterns. Both training and testing residuals are scaled using `MinMaxScaler`, which is important for stabilizing neural network training.

6. Train Attention GRU model on ARIMA Residuals

This step constructs, compiles, and trains the attention enhanced GRU model on the ARIMA residuals to conclude the third phase of the hybrid forecasting approach. The model's architecture is illustrated through `.summary()` method to ensure that the GRU layer, attention mechanism, and the dense output layer are present in the model's structure. Next, the model is compiled by using the Adam optimizer and MSE loss function and including RMSE as a performance metric, meaning that the task is essentially regression-based. The training of the model relies on the residual time series provided as `X_train_gru_residuals` and the targets as `y_train_gru_residuals`. The objective in this case is to predict the future residual value based on the patterns learned from past residuals. Therefore, the model will be trained for 35 epochs while utilizing a batch size of 32 with a validation split of 20%.

7. Inverse Scaled model

As explained above, since the GRU model is trained with the `return_sequences = True`, it gives the predicted residual value for every time step. However, for forecasting, only the prediction of the last time step of each input data is necessary; hence `[:,-1,:]` is used to obtain the desired value stored in the `scaled_gru_forecast_arima_residuals_single`.

8. Final Hybrid Prediction Made

This step generates the final AQI predictions by using hybrid approach using three different models' output i.e. Random Forest, ARIMA, and Attention-based GRU. Further, the predictions done by Random Forest Model for test dataset `(AQI_pred_rfr_test)` were converted into pandas Series with same index values as that of `y_test`. Due to the look-back window concept followed in the GRU Model `(LOOK_BACK_GRU_RESIDUALS)`, first few observations need to be dropped before using the Random Forest predictions as well as the ARIMA's residual forecast results. The same holds true for `GRU_resid_pred_series`. All three series (Random Forest Predictions, ARIMA Residuals Forecast and GRU residual prediction) were trimmed to have a common minimum length in order to avoid dimensional mismatch issues.

4. RESULTS

Model	R2 Score	MAE	MSE
SVR	0.9546	0.2155	0.0890
KNeighborsRegressor	0.9639	0.0699	0.0708
ElasticNet	0.4666	0.8360	1.0449
BayesianRidge	0.4666	0.8360	1.0449
Hybrid ARIMA-RF-Attention GRU	0.9986	0.0057	0.0026

Fig -1: **Comparison table with hybrid Model**

The picture gives the ultimate performance comparison of the five models discussed in the previous chapters, showing the remarkable efficiency of the developed hybrid architecture in the previous architectural iterations. In particular, the proposed model known as the Hybrid ARIMA-RF-Attention GRU becomes the most advanced solution, providing almost a perfect R2 Score of 0.9986 while lowering the amount of error to the minimum values by minimizing the MAE at the level of 0.0057 and MSE down to 0.0026. It is apparent from the results provided that the Attention GRU architecture, used to analyze the residuals of the hybrid ARIMA-RF process, allows capturing nearly all the remaining information in the air quality dataset that cannot be identified by any other algorithm. In comparison, KNeighborsRegressor (R2: 0.9639) and SVR (R2: 0.9546), which proved themselves as the best single models before, are inferior to the deep learning pipeline. Linear regressors, such as ElasticNet and

BayesianRidge, stay the worst algorithms with the same R2 Score of 0.4666.

5. CONCLUSIONS

This research establishes that a hybrid machine learning approach offers a robust and effective solution for predicting the Air Quality Index (AQI) in complex, dynamic, and data-intensive environments. Air quality forecasting is inherently challenging due to the combined influence of multiple pollutant concentrations such as PM2.5, PM10, NO_2, SO_2, CO, and O_3, along with meteorological factors including temperature, humidity, wind speed, and seasonal variations. Additionally, anthropogenic activities such as traffic flow, industrial emissions, and urbanization further increase data variability and uncertainty. Conventional single-model techniques often struggle to accurately represent these diverse and nonlinear interactions, leading to reduced prediction accuracy. By adopting a hybrid modeling strategy, this study addresses these challenges and demonstrates that combining multiple learning paradigms enables more comprehensive representation of the underlying air pollution dynamics. The proposed hybrid framework integrates complementary machine learning models in a unified architecture, allowing the system to capture both linear trends and complex nonlinear dependencies present in historical AQI data. Systematic data preprocessing plays a crucial role in achieving reliable performance, as missing value treatment reduces data inconsistency, normalization ensures uniform feature scaling, and feature selection helps retain the most influential variables while reducing noise and redundancy. These steps collectively improve model convergence, stability, and generalization capability. The hybrid learning process effectively minimizes prediction errors by leveraging the strengths of individual models while compensating for their limitations. As a result, the framework demonstrates strong resilience against overfitting and performs consistently well across different data conditions. Extensive experimental evaluation using standard performance metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and the coefficient of determination (R^2) clearly shows that the hybrid model outperforms individual machine learning models. The improved accuracy and robustness of predictions underline the suitability of the proposed approach for real-world applications.

REFERENCES

1. B. S, S. S. S, S. Lakshmi DM, N. B. AS, and A. NS, "Air Quality Prediction using LSTM Model with Look Back Days Analysis," in 1st International Conference on Radio Frequency Communication and Networks (RFCON), Institute of Electrical and Electronics Engineers (IEEE), pp. 1–4. Doi: 10.1109/rfcon62306.2025.11085269 (2025).
2. S. Khokhlov and I. Yalgoshev, "A Machine Learning Framework for Forecasting Air Quality Index," in Proceedings - International Russian Smart Industry Conference, Smart Industry Con, Institute of Electrical and Electronics Engineers Inc., pp. 598–602. Doi: 10.1109/SmartIndustryCon65166.2025.10985962 (2025).
3. M. Au and I.-H. Kao, "Air Quality Prediction for Multi-Scale Time Period Based on Auto ML," in IEEE Conference on Technologies for Sustainability (Sus Tech), Institute of Electrical and Electronics Engineers (IEEE), pp. 1–8. Doi: 10.1109/sustech63138.2025.11025765 (2025).
4. E. Koçak, "Comprehensive evaluation of machine learning models for real-world air quality prediction and health risk assessment by Air Q+," Earth Sci Inform, vol. 18, no. 3, Doi: 10.1007/s12145-025-01941-7 (2025).
5. S. Thanya Charoen, "Comparative Analysis of Deep Learning and Machine Learning Approaches for Air Quality Prediction," in Proceedings:13th International Electrical Engineering Congress: Carbon Neutrality: Challenges and Solutions Based on Sustainable Power of Nature, Institute of Electrical and Electronics Engineers Inc. Doi: 10.1109/iEECON64081.2025.10987702 (2025).
6. H. N. V. Bhaskar and V. V. Mayil, "Forecasting Traffic Air Pollution Increases from Urban Transportation Using Machine Learning in Metropolitan Cities-A Review," in 4th International Conference on Sentiment Analysis and Deep Learning, ICSADL- Proceedings, Institute of Electrical and Electronics Engineers Inc., pp. 1280–1285.Doi: 10.1109/ICSADL65848.2025.10933458 (2025).
7. M. M. Abdelmalek, H. Mahmoud, and H. Shokry, "Prognosis of air quality index and air pollution using machine learning techniques," Sci Rep, vol. 15, no. 1, Doi: 10.1038/s41598-025-11260-y (2025).
8. R. Mohanraj, , and R. K. Kodali, "Air Quality Prediction Using Machine Learning," in 17th International Conference on Knowledge and Smart Technology, KST 2025, Institute of Electrical and Electronics Engineers Inc, pp. 169–174.Doi: 10.1109/KST65016.2025.11003310 (2025).
9. R. Fang, S. Collingwood, Y. Zhang, J. B. Stanford and D. Sleeth, "Optimizing Air Quality Monitoring: Comparative Analysis of Linear Regression and Machine Learning in Low-Cost Sensor Calibration," Aerosol Air Qual Res, vol. 25, no. 1, Doi: 10.1007/s44408-025-00009-x (2025).
10. S. Magesh and K. Geng, "A machine learning interpretation of the correlation between poverty and air pollution in the contiguous United States," Sci Rep, vol. 15, no. 1, Doi: 10.1038/s41598-025-87150-0 (2025).

11. K. H. Yasin, M. I. Yasin, A. D. Iguala, D. Tulu, and E. Kebede, "Predictive machine learning and geospatial modeling reveal PM10 hotspots and guide targeted air pollution interventions in Addis Ababa, Ethiopia," Discover Applied Sciences, vol. 7, no. 4, Doi: 10.1007/s42452-025-06723-w (2025).

12. N. Stakhovsky, L. Mourad, R. Caro, and S. Argyrias, "Innovative machine learning approaches for indoor air temperature forecasting in smart infrastructure," Sci Rep, vol. 15, no. 1, Doi: 10.1038/s41598-024-85026-3 (2025).

13. S. Jayaraman, N. T, A. S, and S. G, "Enhancing urban air quality prediction using time- based-spatial forecasting framework," Sci Rep, vol. 15, no. 1, Doi: 10.1038/s41598-024-83248-z (2025).

14. A. K. Rad, M. J. Nematollahi, A. Pak, and M. Mahmoudi, "Predictive modeling of air quality in the Tehran megacity via deep learning techniques," Sci Rep, vol. 15, no. 1, Doi: 10.1038/s41598-024-84550-6 (2025).

15. A. A. Ajala, O. L. Adeoye, O. M. Salami, and A. Y. Jimoh, "An examination of daily CO2 emissions prediction through a comparative analysis of machine learning, deep learning, and statistical models," Environmental Science and Pollution Research, vol. 32, no. 5, pp. 2510–2535, Doi: 10.1007/s11356-024-35764-8 (2025).

16. Y. Zupek, F. Alpaslan, and E. Aslan, "Air Quality Forecasting Using Machine Learning: Comparative Analysis and Ensemble Strategies for Enhanced Prediction," Water Air Soil Pollution, vol. 236, no. 7, Doi: 10.1007/s11270-025-08122-8 (2025).

17. R. Kirti, A. Valdés-Uribe, and D. Hölscher, "Spatial prediction of evapotranspiration in a tropical mosaic landscape using remote sensing and explainable machine learning,", vol. 40, no. 8, Doi: 10.1007/s10980-025-02195-w (2025).

18. A. T. Nguyen, D. H. Pham, B. L. Oo, Y. Ahn, and B. T. H. Lim, "Predicting air quality index using attention hybrid deep learning and quantum-inspired particle swarm optimization," J Big Data, vol. 11, no. 1, Doi: 10.1186/s40537-024-00926-5 (2025).

19. S. Trenk, "Machine learning-based forecasting of air quality index under long-term environmental patterns: A comparative approach with XG Boost, Light GBM, and SVM," P Lo S One, vol. 20, Doi: 10.1371/journal.pone.0334252 (2025).

20. P. Attey-Yeboah et al., "Utility of low-cost sensor measurement for predicting ambient PM2.5 concentrations: evidence from a monitoring network in Accra, Ghana," Environmental Science: Atmospheres, vol. 5, no. 4, pp. 517–529, Doi: 10.1039/d4ea00140k (2025).