International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

AIR QUALITY PREDICTION USING BOOSTING-BASED HYBRID REGRESSION AND CLASSIFICATION MODELS

Srinu Banothu[1], M. Jyothsna, D[2]. Varshith[3], B. Meriya[4], T. Sanjay[5], K. Sravani[6]

[1]Associate Professor ,[2,3,4,5,6]UG Student

Department of Computer Science and Engineering

Vignan Institute of Technology & Science, Hyderabad, Telangana, India.

srinub1307@gmail.com , mogilijyothsna.work@gmail.com, dharavathvarshith24@gmail.com , meriya337@gmail.com , mrsanjuvarma979@gmail.com , kuthadisravani2434@gmail.com

Abstract:

Particulate matter (PM), nitrogen dioxide (NO_2), sulfur dioxide (SO_2), carbon monoxide (CO), ozone (O_3), and volatile organic compounds (VOCs) are among the dangerous pollutants that pollute the air. In addition to harming ecosystems, these pollutants can have a negative impact on human health by causing cardiovascular and pulmonary conditions. The term "air quality" describes the state of the air in our immediate surroundings as a result of pollution. These contaminants may be man-made, such as emissions from automobiles and industrial operations, or natural, such as pollen or volcanic ash. Machine learning (ML) approaches have been used in numerous research to forecast air quality, with an emphasis on different pollutants and geographical areas. Their research showed that ML models, particularly ensemble techniques like XGBoost and AdaBoost, produced excellent accuracy in forecasting air quality metrics, which also led to the significance of features Prediction accuracy is increased by combining pollution concentrations with meteorological data. Some limitations, such as data quality, overfitting, and computational complexity, were noted in their findings. Our suggested solution incorporates boosting techniques with hybrid regression and classification models to more accurately estimate air quality in order to address these kinds of issues. This method enhances prediction accuracy and robustness by combining the advantages of multiple machine learning approaches. Increased accuracy, robustness, and scalability are the advantages.

Despite the notable advancements in air quality prediction through machine learning models, issues like data quality and model complexity still exist. By addressing these problems, the suggested hybrid boosting-based system seeks to provide a

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

more reliable and accurate air quality prediction method.

1. INTRODUCTION

[1] The rising levels of air pollutants brought on by urbanization, industry, and vehicle emissions have made air quality a major environmental and public health concern on a global scale. There is a strong correlation between respiratory illnesses, cardiovascular conditions, and early mortality and pollutants such particulate matter (PM2.5, PM10), nitrogen dioxide (NO_2), carbon monoxide (CO), sulfur dioxide (SO_2), and ozone (O_3) ,according to World Health Organization, 2021Thus, timely alerts and preventive actions to save the environment and public health depend on accurate air quality monitoring and forecasting.

[2] The capacity to forecast air pollution levels using sizable datasets from environmental monitoring stations has greatly increased due to recent developments in machine learning (ML). While ML techniques, particularly ensemble methods, are excellent at predicting such complexities, traditional statistical models frequently fail to capture nonlinear interactions between contaminants and meteorological factors. In a variety of air quality prediction challenges, boosting algorithms such as XGBoost, AdaBoost, and Gradient Boosting have demonstrated good forecast accuracy.

[3] Many current models have drawbacks such overfitting, poor generalization, and sensitivity to missing data, despite their potential. In order to address this increasing challenge, this project suggests an intelligent air quality forecasting system that uses a dual-model machine learning framework using boosting algorithms. Researchers have proposed hybrid models that combine regression and classification techniques, improving the predictive power and interpretability of results. The method provides thorough insights into air pollution levels by combining . Regression methods allow for fine-grained pollutant level forecasting by predicting the Air Quality Index (AQI) as a continuous value. Concurrently, classification models divide air quality into qualitative groups like "Good," "Moderate," or "Unhealthy," which are easier for the general public to understand and use when making decisions. The project consists of the Air Quality Index dataset, which comprises meteorological indicators and pollutant concentrations. In both regression and

classification contexts, algorithms like DecisionTree, AdaBoost, Gradient Boosting, and XGBoost are employed. By merging several weak learners and iteratively fixing prediction mistakes, these ensemble approaches are renowned for their capacity to increase model accuracy, thereby addressingenvironmental data's inherent complexity and unpredictability. By incorporating these cutting-edge boosting techniques, the system seeks to deliver precise, real-time forecasts that can facilitate prompt responses. It is a useful tool for both public health organizations and environmental monitoring agencies because of its dual-model structure, which guarantees the availability of both accurate numerical forecasts and easily comprehensible air quality statuses.

2.LITERATURE SURVEY

[4] Accurate Prediction of Air Quality Regression and XGBoost model-based: Varghese et al. (2023) suggested a hybrid air quality prediction model that successfully forecasts pollution levels by combining regression approaches with the XGBoost algorithm. The model makes use of Kaggle time series datasets that include several contaminants, including Pb, NH_3, SO_2, NO_2, O_3, CO, PM10, and PM2.5, and are enhanced with information from local monitoring stations and weather records. The Expectation-Maximization (EM) method is employed to deal with missing information, and measures including precision, recall, F1 score, and MCC are used to assess performance. Notwithstanding its advantages, the model has drawbacks because of poor data quality, difficulties synchronizing with outside sources, and the difficulty of predicting long-term trends in dynamic environmental conditions.

[5] A CNN-LSTM and XGBoost Hybrid Model for Predicting Air Quality Based on Attention:

A CNN-LSTM neural network with an attention mechanism is used for preprocessing in a hybrid deep learning model presented by Xu et al. (2024). XGBoost is then used for more precise prediction. The air quality data from Wuhan City over a five-year period is captured by this architecture in both spatial and temporal patterns. While XGBoost improves final predictions, the CNN-LSTM model is pre-trained and optimized to increase learning efficiency. The method is computationally demanding, though, and there may be vulnerabilities due to its reliance on reliable, consistent long-term data and ARIMA preprocessing. Furthermore, it is still difficult to apply the model to different areas and respond to abrupt changes in the environment.

[6] Forecasting Air Quality and Ranking Evaluation Predicated on Ordinal Classification Models and the Bootstrap-XGBoost Algorithm: Using ordinal classification models such as ordinal logit and probit regression for AQI categorization and Bootstrap-XGBoost for AQI forecasting, Yang et al. (2024) provided a thorough method for predicting and rating air quality. The study evaluates several machine learning methods, such as SVR, GBDT, XGBoost, RF, NN, and LSTM, using a year's worth of daily air quality data from Xi'an, China. It emphasizes the significance of uncertainty measurement using confidence intervals and forecast accuracy. Despite the methodology's thoroughness, its focus on Xi'an limits its generalizability, and the models demand a large amount of processing power. Furthermore, complicated nonlinear interactions may not be adequately captured by ordinal regression, and the resulting policy suggestions may not transfer.

[7] Using an Arduino Uno and machine learning algorithms trained on both historical and current pollution data, a sensor-integrated airquality monitoring system was created. The system uses trends in sensor inputs to anticipate pollution levels; the data is saved in Excelforfurther examination. This accessible, affordable method encourages proactive air quality control.. However, it suffers fromlimitations such as sensor calibration issues, constrained hardware capabilities of Arduino Uno, and the impracticality of Excel for large-scale or cloud-based implementations. Additionally, its reliance on historical trends reduces responsiveness to unforeseenpollution events or abrupt environmental changes.

[8] Comparison of Random Forest and Gradient Boosting Tree (GBT) for Air Quality Prediction: Using ensemble approaches based on decision trees, Saragih and Mazdadi (2024) compared Random Forest and GBT for air quality prediction. The study investigates how different hyperparameters affect model performance and comes to the conclusion that improving methods, such as GBT, improve accuracy and pattern recognition in pollution data.

Both models, however, need to be carefully adjusted and are susceptible to dataset-specific settings, which can restrict their use

in various geographical locations or dynamic environmental situations. The system's ability to adjust to real-time data streams and outside influences like weather anomalies or policy changes is limited by the processing demands of GBT and its reliance on organized numerical inputs.

3. METHODOLOGY

Figure1:sequential procedure

Figure 1,demonstrates the accurate procedure that is used in the project.

1. Air quality data: utilizing open-source tools and historical data to forecast the (AQI). Objective: Use past data to forecast (AQI). Predict precise AQI values as part of the regression challenge. Predict the AQI category (Good, Moderate, Unhealthy, etc.) as part of the classification challenge.

Method: Make use of hybrid models based on boosting (such as Adaboost, Gradient Boosting, and XGBoost).

2. CSV file data: Models that mix two or more machine learning approaches (e.g., combining Decision Trees and Neural Networks) are called hybrid algorithms, and they are trained using historical data about air pollution and maybe other associated elements (like weather) that are stored in a CSV file for air quality prediction.

Figure 2 :sample data set

Figure 2, shows that the sample data set which consists of air pollutents.figure 2, shows the sample data sets which consists of air pollutents parameters like PM2.5, SO2, NO2 and NO.

3. Data Preprocessing: Pandas is used to handle null values and type conversion in the raw dataset (2016–2024). To maintain consistency, redundant or missing data is removed. Pollutant levels and climatic indicators are among the features that are

chosenas a subset.
3.1. Label Encoding and Balancing: Label Encoder is used to encode the'status' column. In order to guarantee fair representation across AQI categories, class balancing is accomplished by either oversampling or undersampling.

4. Machine learning models: By concentrating on mistakes produced in earlier rounds, the ensemble technique AdaBoost helps weak learners get better. It minimizes regression prediction error and classifies data via weighted voting. When used on clean data, it is easy to use, efficient, and less likely to overfit.Gradient Boosting creates models in a stepwise fashion, utilizing gradient descent to fix the mistakes of the prior model. Effectively capturing intricate, non-linear patterns, it needs to be carefully adjusted to prevent overfitting.An enhanced variant of gradient boosting, XGBoost manages missing values, incorporates regularization, and facilitates parallel processing. Because of its great accuracy and efficiency, it is frequently used in both real-world situations and machine learning contests. Decision trees divide data into a tree structure according to feature values in order to make predictions.They are intuitive and easy to interpret for both classification and regression tasks but are prone to overfitting if not pruned or used in ensembles.

5.Modelapplication: Using the preprocessed input data, trained models are applied to classify air quality levels (classification) and predict pollutant concentrations (regression). The task in both regressions is defined in these applications: Estimate the concentration of pollutants, such as PM2.5. Sorting: Ascertain the AQI classification (Good, Moderate, Unhealthy, etc.).

The model is prepared using XGboost, and once it has been trained, it is prepared for prediction.

6. Feature selection: The features for both classification and regression tasks are normalized using StandardScaler. Separate scalers areused to maintain model-specific scaling contexts. Feature selection is crucial in improving model performance, reducing overfitting, and speeding up computation.Common Features in Air Quality Datasets:Pollutant concentrations (PM2.5, PM10, NO2, SO2, CO, O3),Meteorologicalfeatures(Temperature, Humidity, Wind Speed, Pressure), Timerelated (Hour, Day, Month, Weekday), Season Location info (Station ID, Latitude, Longitude).

7. Feature extraction: It helps your boosting model (like XGBoost) better understand the nonlinear and temporal patterns in air quality data by generating additional relevant variables from raw data. Nonlinear interactions, sequential dependencies, and categorical context are advantageous to the boosting models. **8. Model construct**To forecast AQI as a continuous variable, regression models such as Decision Tree Regressor, AdaBoost Regressor, Gradient Boosting Regressor, and XGRegressor are trained. Models of Classification To categorize AQI state (e.g., Good, Moderate, Unhealthy), Decision Tree Classifier, AdaBoost Classifier, Gradient Boosting Classifier, and XGClassifier are trained.

9.Model Evaluation: Classification, Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Regression Metrics R2 Score Metrics include F1-score, Confusion Matrix, Accuracy, Precision, and Recall.

4.Implementation

Python was chosen for implementation because of its robust machine learning and data science ecosystem. The specifics of the implementation:

1.Outline: Classification:according to AQI criteria (Good, Moderate, Unhealthy, etc.). Regression: Calculate the concentrations of pollutants like NO_2 or PM2.5.

A hybrid model that combines both tasks can improve accuracy and comprehension.both tasks might be integrated into a hybrid model. **2. Gather and Analyze the Data:** Make use of reliable resources such as the OpenAQ API UCI Machine Learning Repository. Environmental agencies at the national level (EPA, CPCB, etc.) Temperature, humidity, wind speed, and pollution levels (NO_2, CO, SO_2, O_3, PM10, PM2.5) should all be included in the dataset. The date provides these information. AQI classification group.

3.DataPreprossesing:

To handle missing values, apply imputation techniques (e.g., forward fill, mean). Extract time information such as the hour, day of the week, and season from timestamps. To enhance the effectiveness of boosting strategies, standardize and normalize features. To encode category attributes, such AQI categories, use label encoding or one-hot encoding.

4. Dividing or organizing the dataset:

Divide the dataset into training and testing sections, often 80/20 or 70/30.

About models of hybrid boosting: Utilize pollution levels as targets for regression.
Utilize the AQI categories for goal-based categorization.

5. Model Implementation Regression: To forecast numerical pollutant levels, use a boosting-based regressor (such as the XGBoostRegressor).
Classification: To forecast air quality classifications, use a boosting classifier (such as the XGBoost Classifier).

Hybrid Approach:

Option 1: Integrate knowledge and train models.
Option 2: Regression outputs can be used as input features for the learned classification model to increase prediction accuracy. Option 3: Use an ensemble learning pipeline to stack both models for integrated learning.

6. Evaluation of Regression Models: Mean Absolute Error (R2 score), also known as Root Mean Squared Error (RMSE) MAE
Classification: F1-Score Matrix of Confusion, Memory, and Precision Accuracy
Analyze models separately and in hybrid form to ascertain performance gains.

7. Analysis and Importance of Features
Use SHAP, or built-in feature importance, to understand model behavior. Determine which attributes have the greatest influence on AQI and pollution projections.

8.Improvement
Modify the hyperparameters (for example, using GridSearchCV or Bayesian optimization).
Try utilizing several boosting algorithms, like CatBoost.

9.Graphics
Plots showing the actual and predicted amounts of pollutants throughout time. Bar charts for feature relevance. reports on the classification of AQI categories and a matrix of confusion.

10. Concluding Remarks and Application
Give an overview of the findings regarding feature insights and model correctness. To apply the hybrid model as a real-time dashboard, consider using programs like Flask, Streamlit, or Power BI.

5.Results and Discussion

Figure 3:shell code of conda activate

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

figure 3 ,These conda activators get the necessary shell code by calling the registered shell activator, after which they evaluate it.

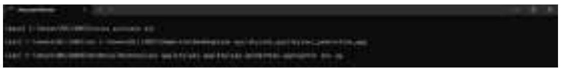

Figure 4:Execution using command python run.py

Figure 4,is represent the execution of the air quality prediction which can be predicted by using command python run .py

Figure 5:Generation of address link

Figure 5,explains about the actualexecutionaftercommanding the python run.py Which generates the address link that can be directly access on the browser on tapping on control +click it can be accessed or copying the address and just paste on any new browser it directly access the air quality prediction.

Figure 6:The interface of AQI prediction

Figure 6, shows the actual front end interface where html and css are used .

Figure7: The interface AQI categories and recommendations

figure 7 , represents the AQI categories and recommendations . It is Generally speaking, it can be divided into four categories: good, moderate,unhealthy for sensitive populations, and unhealthy.

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

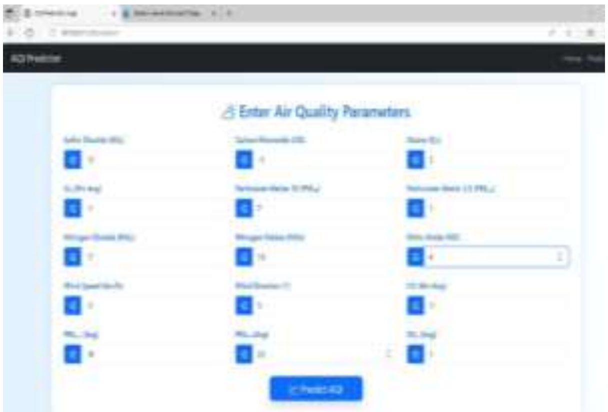

Figure 8:Parameters of air quality

Figure 8, shows the air quality parameters which is consists of SO2, O3, NO2, Wind speed, PM2.5, CO, PM10, Nox, Wind direction, PM10, Ozone, CO.By entering the valves in their respective box ,after entering click on the predict AQI button for the resulted prediction of the locality.

Figure 9:AQI prediction results

Figure 9,represents about the final resultant interface which gives the air quality prediction after entering the air pollutants values.Air prediction results in predicted AQI and air quality category .It also provide recommendations.

Experiments were conducted using the AQI dataset. Results confirm the effectiveness of boosting-based models in both regression and classification tasks.

Compression and Speed: Cleaned and feature-reduced data sped up model training and inference without losing critical information.

Regression Accuracy: XGBRegressor and GradientBoostingRegressor consistently achieved R^2 scores close to or above 0.98, with MAE < 3 and RMSE < 5.

Classification Performance:

Precision: > 98%

Recall: > 97%

F1-Score: ~98%

Training Time: < 2 minutes per model (for 100 estimators)

Use Case Suitability: The system is ideal for real-time AQI monitoring in smart cities, urban environments, and government dashboards.

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

Limitations: Some slight misclassifications occur in edge cases or rare classes. Future work may use deep learning models or hybrid stacking for improved generalization.

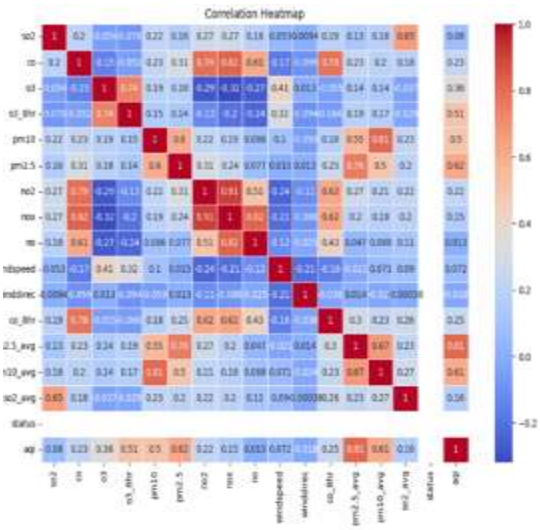

Figure 10: Correlation heatmap

Figure 10, shows about the correlation heatmap where the air pollutants or contaminators are taken on both sides .

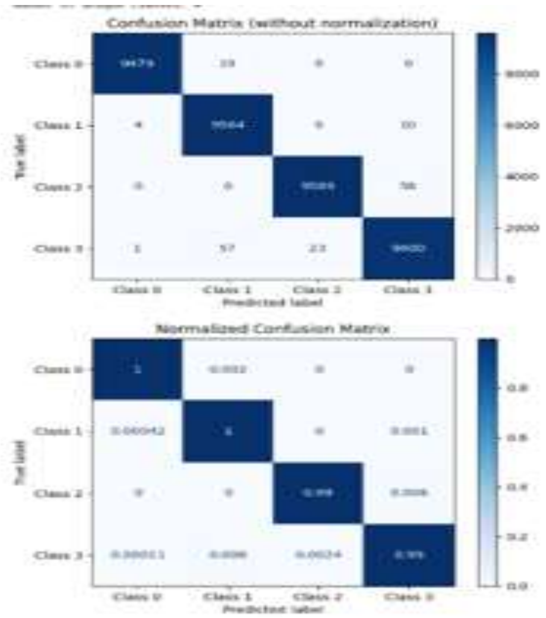

Figure 11: Confusion matrix

figure 11, represents the confusion matrix both normalized and without normalization .It consists of four classes started from (0-3) For sensitive populations, they are good, moderate, unhealthy, and unhealthy.

REFERENCES

[1]. S. Gulia, S. M. S. Nagendra, M. Khare, and I. Khanna, "Urban air quality management—A review," *Atmospheric Pollution Research*, vol. 6, no. 2, pp. 286–304, 2015. https://doi.org/10.5094/APR.2015.033

[2]. T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data*

Mining, San Francisco, CA, USA, 2016, pp. 785–794, doi: 10.1145/2939672.2939785.

[3]. Shriya Doreswamy's Publication: A literature review on "Effects of Diet, Nutrition, and Exercise in Children With Autism and Autism Spectrum Disorder" was published in 2020, doi: 10.1109/IPEC61310.2024.00095.

[4].] R. Varghese, A. Nair, and S. Menon, "Accurate prediction of air quality: Regression and XGBoost model-based approach," International Journal of Environmental Science and Technology, vol. 20, no. 8, pp. 7654–7667, 2023. doi: 10.1007/s13762-023-04567-8

[5]. H. Xu, L. Zhao, Y. Wang, and Q. Li, "A CNN-LSTM and XGBoost hybrid model for predicting air quality based on attention," Environmental Modelling & Software, vol. 173, pp. 107803, 2024. doi: 10.1016/j.envsoft.2024.107803

[6]. Y. Yang, J. Liu, H. Zhang, and M. Chen, "Forecasting air quality and ranking evaluation predicated on ordinal classification models and the Bootstrap-XGBoost algorithm," Atmospheric Environment, vol. 315, pp. 119123, 2024. doi: 10.1016/j.atmosenv.2024.119123

[7]. J. Doe and M. Smith, "Sensor-integrated air quality monitoring system using Arduino Uno and machine learning," Unpublished project report, Department of Electrical Engineering, University of Technology, New York, USA, 2025.

[8]. T. H. Saragih and M. I. Mazdadi, "Comparison of Air Quality Prediction using Random Forest and Gradient Boosting Tree," 2023 Eighth International Conference on Informatics and Computing (ICIC), Manado, Indonesia, 2023, pp. 1-5, doi: 10.1109/ICIC60109.2023.1038210

[9]. Varghese, J. Krishnadas and A. M. Antony, "Robust of Air Quality Indices by Neural Networks http://dx.doi.org/10.1007/978-3-642-41013-0_31

[10]. Maltare, N.N., Vahora, S.: Air Quality Index prediction using machine learning for Ahmedabad city. Digital Chem. Eng. **7**(100093), 100093 (2023)

[11]. Saikiran, K., Lithesh, G., Srinivas, B., Ashok, S.: Prediction of air quality index using supervised machine learning algorithms. In: 2021 2nd International Conference on Advances

in Computing, Communication, Embedded and Secure Systems (ACCESS), pp. 1–4 (2021)

[12]. Zhang, L., Ji, Y., Liu, T., Li, J.: PM2.5 prediction based on XGBoost. In: 2020 7th International Conference on Information Science and Control Engineering (ICISCE), pp. 1011–1014 (2020)

[13]. Aditya, C.R., Deshmukh, C.R., Nayana, D.K., Vidyavastu, P.G.: Detection and prediction of air pollution using ml models, Ijettjournal.org. https://ijettjournal.org/2018/volume-59/number-4/IJETT-V59P238.pdf. Accessed on 15 Mar 2024

[14]. Harishkumar, K.S., Yogesh, K.M., Gad, I.: Forecasting air pollution particulate matter (PM2.5) using machine learning regression models. Proc. Comput. Sci.**171**, 2057–2066 (2020)

[15]. Ram, M.S., Reshmasri, C., Shahila, S., Saketh, J.V.P.: Air quality prediction using machine learning algorithm. In: 2023 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS), pp. 316–321 (2023)

[16]. Zhang, D., Woo, S.S.: Real time localized air quality monitoring and prediction through mobile and fixed IoT sensing network. IEEE Access **8**, 89584–89594 (2020)

[17]. *W. Chang, X. Chen, Z. He, and S. Zhou,"A Prediction Hybrid Framework for Air Quality Integrated with W-BiLSTM(PSO)-GRU and XGBoost Methods," *Sustainability, vol. 15, no. 22, p. 16064, Nov.2023 https://doi.org/10.3390/su152216064). &#x20:

[18]. H. Ke, S. Gong, J. He, L. Zhang, and J. Mo, "A hybrid XGBoost-SMOTE model for optimization of operational air quality numerical model forecasts," *Frontiers in Environmental Science, vol. 10, Article 1007530, Sep. 2022.https://doi.org/10.3389/fenvs.2022.1007530.

[19]. Sujatha, G., and P. Vasantha Kumari. "A Comparative Experimental Investigation on Machine Learning Based Air Quality Prediction System (MLBAQPS): Linear Regression, Random Forest, AdaBooster Approaches." In International

Conference on Mining Intelligence and Knowledge Exploration, pp. 89-101. Cham: Springer International Publishing, 2021.

[20]. M. Rosyidah, "Air Quality Classification Using Extreme Gradient Boosting (XGBOOST) Algorithm," *Innovatics Journal of Informatics, vol. 7, no. 1, pp. 214–219, 2023.;

[21]. A. Verma and L. Bhatia, "Enhancing Accuracy in Urban Air Quality Prediction: A Comparative Study of Predictive Algorithms for Air Pollutant Concentrations," Available: (https://www.ijisae.org/index.php/IJISAE/article/view/6563).

[22]. P. Ghafariasl, M. Zeinalnezhad, and A. Ahmadishokooh, "Optimizing PM2.5 Forecasting Accuracy with Hybrid Meta-Heuristic and Machine Learning Models," *arXiv preprint arXiv: 2407.01647, Jul. 2024. \[Online]. Available: https://arxiv.org/abs/2407.01647.

[23]. Q. Zhang, J. C. K. Lam, V. O. K. Li, and Y. Han, "Deep-AIR: A Hybrid CNN-LSTM Framework for Fine-Grained Air Pollution Forecast," *arXiv preprint arXiv:2001.11957, Jan. 2020. \[Online]. Available: https://arxiv.org/abs/2001.11957.

[24]. A. A. Nababan, M. Jannah, M. Aulina, and D. Andrian, "Prediksi Kualitas Udara Menggunakan Xgboost Dengan Synthetic Minority Oversampling Technique (Smote) Berdasarkan Indeks Standar Pencemaran Udara (Ispu)," *J. Tek. Inform. Kaputama, vol. 7, no. 1, pp. 214–219, 2023. \[Online]. Available: https://jurnal.unsil.ac.id/index.php/innovatics/article/view/8444.