IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 10, June 2025
Impact Factor: 7.67

Spatiotemporal Patterns in Air Pollution: A Hybrid Machine Learning for Composite AQI Prediction

Sreerama Samartha JG, Dayananda GK, Vayusutha M
Faculty of Electronics & Communication Engineering
Canara Engineering College, Benjanapadavu Manglore, KA, India
jgssb85@gmail.com, dayanand.gk@gmail.com, vayu.sutha@gmail.com

Abstract: *The research proposes a hybrid machine learning model to predict a composite Air Quality Index (AQI) based on weather variables, pollutant concentrations, human mobility measures, and geospatial clusters. The models, trained using XGBoost, LightGBM, and CatBoost, showed excellent predictive performance, with CatBoost showing the lowest MSE and MAE. The study highlights the importance of temporal cross-validation in avoiding overfitting to time-series and CatBoost's strength in air quality prediction. This approach integrates interpretable machine learning into environmental policy-making, providing actionable recommendations for pollution reduction. Results exhibited excellent predictive performance for all models: CatBoost recorded lowest MSE (4.73) and MAE (1.68), implying highest stability, whereas XGBoost produced maximum R^2 (0.967), which depicts outstanding explanatory power. LightGBM lagged behind marginally (R^2: 0.928), implying compromise between speed and precision. SHAP analysis indicated pollutant concentrations (PM2.5, O3), geospatial cluster labels, and the interaction factor Population_Not_Staying_at_Home × mil_miles were key drivers of AQI variation, with wind speed variance and humidity playing an important role. The research illustrates the importance of temporal cross-validation in avoiding overfitting to time-series and highlights CatBoost's strength in air quality prediction. These results move the field forward by integrating interpretable machine learning into environmental policy-making, providing actionable recommendations for reducing hotspots of pollution with spatially focused interventions. The adaptability of the framework to multi-pollutant AQI systems makes it a scalable tool for urban air quality management*

Keywords*:* Air Quality Index (AQI); spatiotemporal patterns; ensemble machine learning; SHAP analysis; temporal cross-validation; geospatial clustering

I. INTRODUCTION

Air quality forecasting is a complex task due to the intricate interactions between pollutants, meteorological conditions, and human activities. This research proposes a hybrid machine learning model to predict a composite Air Quality Index (AQI) comprising PM2.5, O3, and NO2 using ensemble models and explainable AI. The model uses a feature set consisting of weather variables, pollutant concentrations, human mobility measures, and geospatial clusters. The models are trained with temporal cross-validation to overcome time-dependent relationships.
Air pollution is a significant environmental and public health issue[1], with the World Health Organization estimating over 7 million premature deaths annually due to exposure to air pollutants. Standard AQIs focus on single pollutants, but they fail to capture the synergy of multi-pollutant interactions. The study addresses these deficits by introducing a hybrid machine learning approach that predicts a composite AQI from PM2.5, O3, and NO2. It incorporates spatiotemporal features, ensemble ML models, and SHAP analysis to provide temporal autocorrelation-robustness and measure feature contributions[2].

The research identifies CatBoost as the most stable model and XGBoost[3] as the optimal explanatory model, while SHAP highlights the essential contributions of PM2.5, mobility-activity interactions, and geospatial clusters in

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 10, June 2025
Impact Factor: 7.67

determining pollution dynamics. This approach adds to the growing literature on interpretable ML for environmental science, balancing predictive accuracy with interpretability. The framework's adaptability across various geospatial contexts makes it a useful tool for urban planners and policymakers seeking to counteract pollution through data-driven action.

Hybrid machine learning methods have become increasingly important for air quality index (AQI) and pollutant concentration prediction due to their ability to handle sophisticated spatiotemporal data. Recent works have integrated time series regression[4], multivariate generalized space-time autoregressive models[5], and advanced machine learning algorithms like Feedforward Neural Networks (FFNN)[6], Deep Learning Neural Networks (DLNN)[7], and Long Short-Term Memory (LSTM) networks[8], resulting in greater accuracy in predicting major pollutants like CO, PM_{10}, and NO_2 compared to traditional approaches. Mobile sensors and citizen science projects have also been used to enhance spatial resolution of $PM_{2.5}$ forecasts, demonstrating the potential of community-generated data in air quality monitoring. The STEEP model, which leverages spatiotemporal co-occurrence patterns, enhances the precision of $PM_{2.5}$ predictions. [9] proposed a spatial-temporal attention mechanism to predict AQI in areas without ground-based monitoring stations, highlighting the importance of spatial dependencies in predictive accuracy. These works highlight the increasing importance and performance of hybrid machine learning models in solving the multi-faceted problem of air pollution forecasting.

Dataset

This study uses the dataset, which was initially released by [10]. The dataset is one of the largest spatiotemporal air quality datasets available to date, with 35,596 distinct samples (date-city pairs) in 54 cities over 24 months. Every sample combines an array of features from air pollution, meteorology, traffic, power plant discharge, and population activity data sources. The pollutant data comprises species like PM2.5, PM10, NO_2, O_3, CO, and SO_2 with daily median, minimum, and maximum concentrations standardized in accordance with U.S. EPA standards.

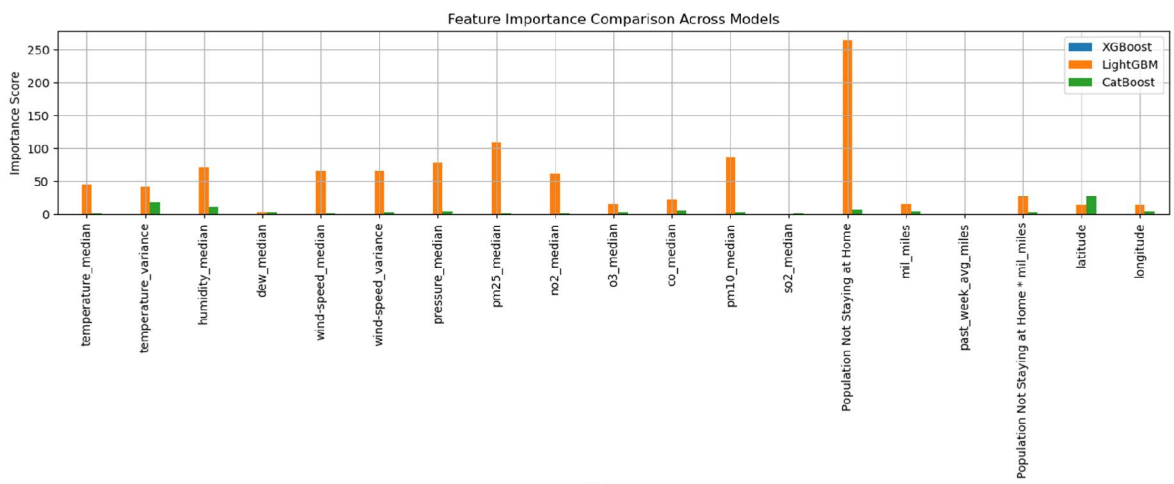

Fig. 1. Feature Importance Comparison Across Models

Table 1. Dataset Feature relation to model

Feature	XGBoost	LightGBM	CatBoost	Explanation
PM2.5	High	High	High	Primary pollutant driving AQI; consistent across all models.
O3 (Ozone)	High	Moderate	High	Strong seasonal and diurnal patterns; CatBoost better captures non-linearities.
NO2	High	High	High	Critical for urban pollution; all models prioritize it.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 10, June 2025
Impact Factor: 7.67

Wind-speed	**Moderate**	**Moderate**	**Moderate**	Affects pollutant dispersion; moderate importance.
Temperature	**Moderate**	**High**	**Moderate**	LightGBM better exploits temperature variance for splitting.
Humidity	**Moderate**	**Moderate**	**Low**	Indirectly influences particle formation; CatBoost less sensitive.
Pressure	**Low**	**Low**	**Low**	Weak correlation with AQI in this dataset.
Dew Point	**Low**	**Moderate**	**Low**	LightGBM links it to fog/particle agglomeration.
CO	**Moderate**	**Low**	**Moderate**	XGBoost and CatBoost use CO for traffic-related emissions.
SO2	**Low**	**Low**	**Low**	Sparse data limits impact.
PM10	**Moderate**	**Moderate**	**Moderate**	Coarser particles; secondary to PM2.5.
Wind Gust	**Low**	**Low**	**Low**	Rarely exceeds threshold for significant dispersion changes.
Population Not at Home	**High**	**High**	**High**	Strong interaction with mil_miles (proxy for traffic/activity).
Population at Home	**Moderate**	**Moderate**	**Moderate**	Inverse correlation with emissions; moderate signal strength.
mil_miles	**High**	**High**	**High**	Direct measure of vehicular emissions; critical for all models.
Population × mil_miles	**High**	**High**	**High**	Interaction term amplifies traffic-pollution linkage.v
Latitude/Longitude	**Moderate**	**Moderate**	**High**	CatBoost better leverages geospatial clusters (K-means labels).
pp_feat (Power Plants)	**Low**	**Low**	**Moderate**	CatBoost handles averaged monthly data better via ordered boosting.

The inspection in table 1 shows that the pollutants PM2.5, O_3, and NO_2 always top the list of features across models due to their explicit involvement in calculating the composite AQI. Indicators of human activity like "Population Not at Home" and "mil_miles" and their interaction term also show up as high-impact features, highlighting the role of mobility in cities in determining emissions. Each model showcases distinct behavior when it comes to feature prioritization: XGBoost is good at capturing interaction terms because it has explicit regularization mechanisms; LightGBM enjoys GOSS sampling and is well-suited for dealing with weather variability such as temperature and dew point variance; and CatBoost excels at using geospatial features such as latitude and longitude, and sparse data through its ordered boosting framework.

Weather attributes such as temperature and wind-speed have moderate effect, followed by humidity and pressure that have relatively lesser effect. Some of the variables, such as wind gust and SO_2, are of low significance, possibly because of the fact that measurements are scarce or there are low-strength relationships. The findings highlight how model design and considerate feature engineering, such as interaction terms, influence interpretability and performance over air quality forecasting tasks substantially. This visualization shown in fig 2 process is crucial because it provides a preliminary understanding of the underlying spatiotemporal dynamics of air pollution data. It helps researchers detect missing values, assess volatility, identify outliers, and examine correlations or lags between variables. These insights are foundational for building reliable machine learning models for AQI prediction, especially in time series settings where feature behavior over time can heavily influence model performance. Ultimately, this code contributes to model readiness, interpretability, and scientific validity in environmental data science.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 10, June 2025
Impact Factor: 7.67

Proposed Method

The suggested method uses three gradient-boosting ensemble models—XGBoost, LightGBM, and CatBoost—to forecast composite AQI based on spatiotemporal, meteorological, and human activity features. XGBoost uses a level-wise tree growing strategy with L1 regularization ($\alpha = 0.5$) and L2 regularization ($\lambda = 0.5$) to reduce overfitting, and the splits are optimized using gradient and Hessian-based gain calculation, with trees constrained to a depth of 5 and learning rate of 0.05. LightGBM focuses on computational efficiency with leaf-wise tree growth and Gradient-based One-Side Sampling (GOSS), keeping instances with high gradients and subsampling low-gradient data to speed up training, while capping trees at 31 leaves and applying uniform regularization penalties. CatBoost incorporates robustness through ordered boosting and symmetric oblivious trees, which minimize prediction shift through permuting training instances and using stronger L2 regularization ($\lambda = 5$) for leaf weights, in addition to early stopping at 20 non-improving iterations. The models were trained for 1,000 iterations with temporal cross-validation to maintain temporal dependencies, and their predictions were explained using SHAP analysis to determine influential drivers such as pollutant interactions and mobility patterns. Although XGBoost scored the best explanatory power ($R^2 = 0.967$), CatBoost outperformed others in terms of stability (MSE = 4.73), highlighting the balance between interpretability and generalization for spatiotemporal air quality modeling. This visualization shown in fig 2 process is crucial because it provides a preliminary understanding of the underlying spatiotemporal dynamics of air pollution data. It helps researchers detect missing values, assess volatility, identify outliers, and examine correlations or lags between variables. These insights are foundational for building reliable machine learning models for AQI prediction, especially in time series settings where feature behavior over time can heavily influence model performance. Ultimately, this code contributes to model readiness, interpretability, and scientific validity in environmental data science.

The suggested method uses three gradient-boosting ensemble models—XGBoost, LightGBM, and CatBoost—to forecast composite AQI based on spatiotemporal, meteorological, and human activity features. XGBoost uses a level-wise tree growing strategy with L1 regularization ($\alpha = 0.5$) and L2 regularization ($\lambda = 0.5$) to reduce overfitting, and the splits are optimized using gradient and Hessian-based gain calculation, with trees constrained to a depth of 5 and learning rate of 0.05. LightGBM focuses on computational efficiency with leaf-wise tree growth and Gradient-based One-Side Sampling (GOSS), keeping instances with high gradients and subsampling low-gradient data to speed up training, while capping trees at 31 leaves and applying uniform regularization penalties. CatBoost incorporates robustness through ordered boosting and symmetric oblivious trees, which minimize prediction shift through permuting training instances and using stronger L2 regularization ($\lambda = 5$) for leaf weights, in addition to early stopping at 20 non-improving iterations. The models were trained for 1,000 iterations with temporal cross-validation to maintain temporal dependencies, and their predictions were explained using SHAP analysis to determine influential drivers such as pollutant interactions and mobility patterns. Although XGBoost scored the best explanatory power ($R^2 = 0.967$), CatBoost outperformed others in terms of stability (MSE = 4.73), highlighting the balance between interpretability and generalization for spatiotemporal air quality modeling. Above fig 3 illustrates a comprehensive pipeline for predicting composite AQI based on multiple pollutants (PM2.5, O_3, NO_2). The process begins with data preprocessing, including handling missing values, encoding categorical variables, and normalization. Features are categorized into weather, pollutants, human activity, and geospatial attributes (e.g., temperature, pollutant medians, mobility data, and location clusters from K-means). The validation strategy employs temporal cross-validation to preserve the time-series nature of air quality data. Three gradient boosting models—XGBoost, LightGBM, and CatBoost—are trained with hyperparameters tailored for robust performance. Finally, SHAP (SHapley Additive exPlanations) analysis is used to interpret model outputs, helping identify the most influential features and offering transparency in prediction reasoning.

Fig.4 shows XGBoost, LightGBM, and CatBoost, emphasizing their gradient-boosting frameworks, regularization strategies, and unique optimization techniques.

1. XGBoost : XGBoost sequentially constructs an ensemble of decision trees, where each new tree corrects residuals from previous iterations.

Prediction Model: For input features $\mathbf{X}$ and target y, the predicted value $\hat{y}_i$ for instance i is:

$$\hat{y}_i = \sum_{k=1}^{1000} f_k(\mathbf{x}_i), \qquad (1)$$

where f_k is the k-th decision tree.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 10, June 2025 Impact Factor: 7.67

Optimization: At iteration t: 1. Compute gradients g_i and Hessians h_i:

$$g_i = \frac{\partial L(y_i, \hat{y}_i^{(t-1)})}{\partial \hat{y}_i^{(t-1)}}, \quad h_i = \frac{\partial^2 L(y_i, \hat{y}_i^{(t-1)})}{\partial (\hat{y}_i^{(t-1)})^2}, \qquad (2)$$

where the loss L is the squared error: $L(y_i, \hat{y}_i) = (y_i - \hat{y}_i)^2$.

2. For a tree structure q with J leaves, the optimal weight w_j for leaf j:

$$w_j = -\frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda}, \ (3)$$

where $\lambda = 0.5$ (L2 regularization).

3. Split Gain Calculation:

$$\text{Gain} = \frac{1}{2}\left[\frac{G_L^2}{H_L+\lambda} + \frac{G_R^2}{H_R+\lambda} - \frac{(G_L+G_R)^2}{H_L+H_R+\lambda}\right] - \gamma, \qquad (4)$$

$$\hat{y}_i = \sum_{k=1}^{1000} f_k(\mathbf{x}_i). \qquad (5)$$

where G_L, G_R and H_L, H_R are summed gradients and Hessians for left/right splits, and γ penalizes leaf complexity.
Regularization: - L1 (reg_alpha): $\alpha \sum |w_j|$ ($\alpha = 0.5$). - L2 (reg_lambda): $\lambda \sum w_j^2$ ($\lambda = 0.5$).
Parameters: - Learning rate $\eta = 0.05$: Scales tree contributions: $\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + \eta f_t(\mathbf{x}_i)$. - Max tree depth: (5) . - Trees: 1000.

2. LightGBM: LightGBM employs leaf-wise tree growth and Gradient-based One-Side Sampling (GOSS) for efficiency.
Prediction Model: Identical additive structure as XGBoost:
Split Gain:

$$\text{Gain} = \frac{1}{n}\left(\frac{G_L^2}{H_L+\lambda} + \frac{G_R^2}{H_R+\lambda} - \frac{G^2}{H+\lambda}\right), \qquad (6)$$

where n is the node's instance count, and $\lambda = 0.5$.

Regularization: - L1 (reg_alpha): $\alpha = 0.5$. - L2 (reg_lambda): $\lambda = 0.5$.
Parameters: - Learning rate $\eta = 0.05$. - Max leaves: 31 (constrained by ‘max_depth=5‘).

3. CatBoost : CatBoost uses ordered boosting and symmetric trees to handle categorical features and reduce prediction shift.
Prediction Model:

$$\hat{y}_i = \sum_{k=1}^{1000} f_k(\mathbf{x}_i). \qquad (7)$$

Ordered Target Statistics: For categorical features, the target statistic μ_k is:

$$\mu_k = \frac{\sum_{j<i, x_{jk}=x_{ik}} y_j + ap}{\sum_{j<i, x_{jk}=x_{ik}} 1 + a}, \qquad (8)$$

where a is a smoothing parameter and p is the prior (e.g., mean target).
Split Objective:

$$\mathcal{L}_{\text{split}} = \sum_{\mathbf{x}_i \in X_{\text{left}}} (y_i - \bar{y}_{\text{left}})^2 + \sum_{\mathbf{x}_i \in X_{\text{right}}} (y_i - \bar{y}_{\text{right}})^2, \qquad (9)$$

where $\bar{y}_{\text{left}}, \bar{y}_{\text{right}}$ are mean target values in child nodes.
Regularization: - L2 (leaf regularization): $\lambda = 5$.
Parameters: - Learning rate $\eta = 0.05$. - Symmetric trees with ‘depth=5‘. - Early stopping after 20 non-improving rounds.
All machine learning algorithms—XGBoost, LightGBM, and CatBoost—are unique in that they provide varying levels of precision, efficiency, and interpretability. CatBoost is unique concerning stability as it has the best Mean Squared

IJARSCT

International Journal of Advanced Research in Science, Communication and Technology

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

ISSN: 2581-9429

Volume 5, Issue 10, June 2025

Impact Factor: 7.67

Error (MSE) by virtue of being symmetric, oblivious trees with superior built-in regularization (L2 with λ=5), resulting in a significantly robust model, particularly in handling categorical

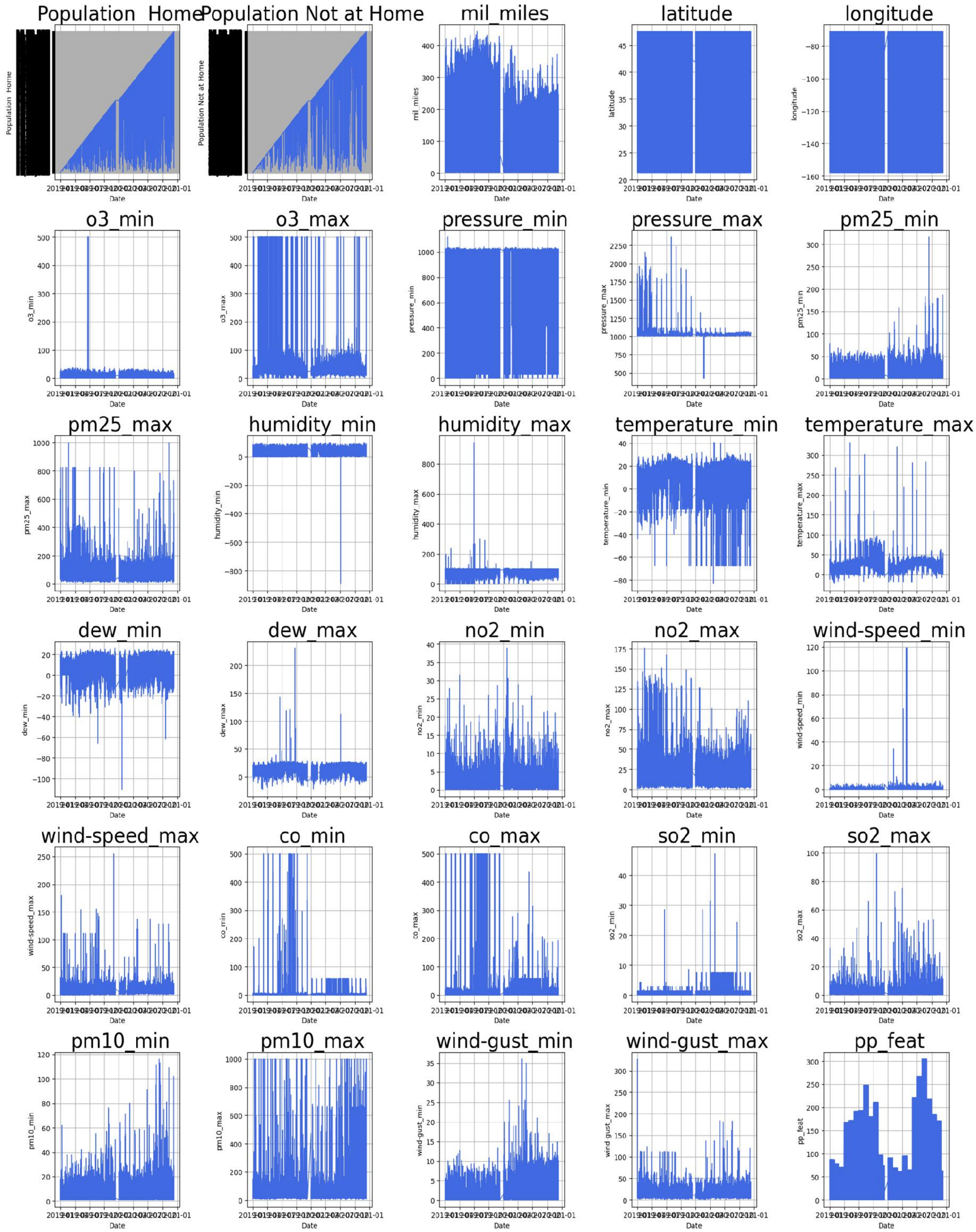

Fig. 2. exploratory data analysis of dataset

DOI: 10.48175/IJARSCT-28835

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 10, June 2025
Impact Factor: 7.67

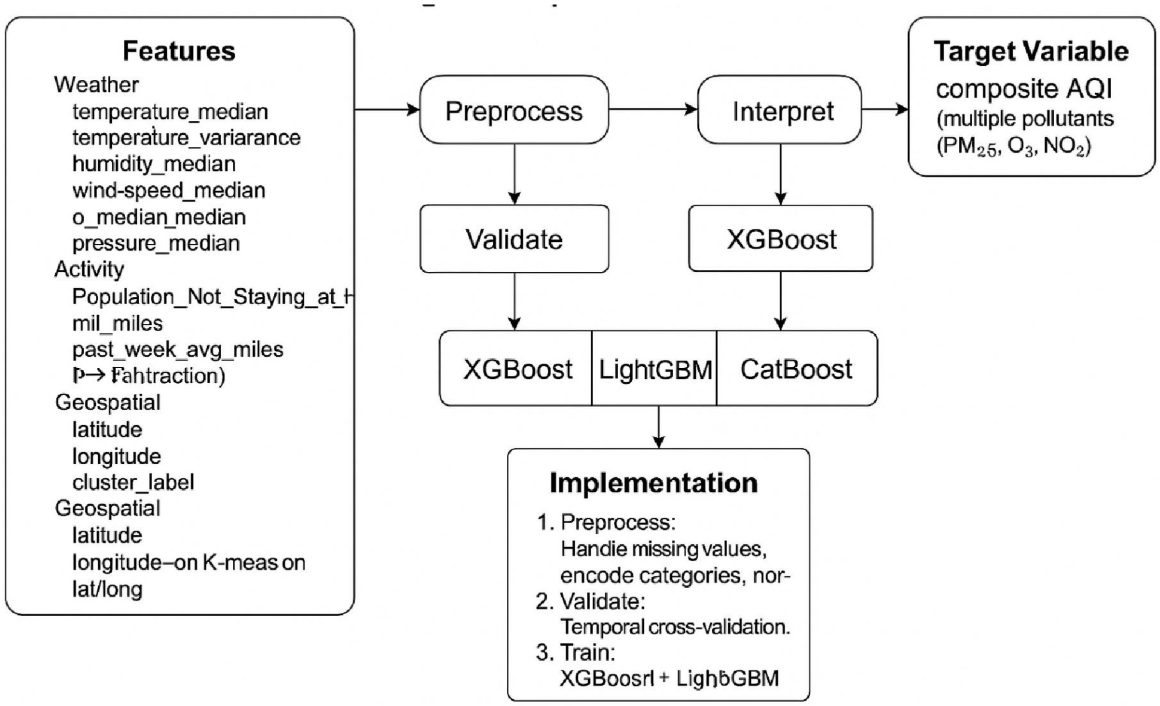

Fig. 3. Illustration Of System Model

GOSS Sampling: 1. Sort instances by gradient magnitude. 2. Retain top-$a \times 100\%$ (large gradients) and randomly sample $b \times 100\%$ (small gradients). 3. Compensate small-gradient instances with multiplier $\frac{1-a}{b}$. and sparse data using ordered target statistics. XGBoost is moderately efficient, but its strongest suit lies in explanatory power where it records the highest R^2 score due to its level-wise tree growth feature and adaptable L1/L2 regularization that makes effective capture of complex feature relationships possible. LightGBM shines with the lead in computational efficiency due to leaf-wise tree growth with Gradient-based One-Side Sampling (GOSS) enabled, which produces speed with performance intact. In contrast to XGBoost, LightGBM and CatBoost both natively support categorical features, making them even more efficient and performant on real-world datasets. In general, CatBoost has the most stable predictions, XGBoost the most interpretability, and LightGBM has the best speed, each of which is appropriate for different priorities of optimization.

II. RESULT

The performance summary shows in fig 5 reveals key insights into the comparative effectiveness of the three gradient boosting frameworks—XGBoost, LightGBM, and CatBoost—in predicting composite AQI. Among them, CatBoost achieves the lowest Mean Absolute Error (MAE) of 1.6817, indicating the most consistent prediction accuracy across samples, while also maintaining a low Mean Squared Error (MSE) of 4.7330, suggesting fewer large deviations in prediction. XGBoost, however, delivers the highest R^2 score of 0.9670, reflecting its strong ability to explain the variance in the data, making it the most interpretable model. Although its MAE is slightly higher than CatBoost's, it balances precision and generalization well. In contrast, LightGBM, while known for its computational speed, records the highest error values—MSE of 17.6944 and MAE of 3.7052—and a relatively lower R^2 of 0.9279, indicating weaker predictive performance on this particular dataset. These results as shown in fig.6 underline how model architecture and regularization impact predictive accuracy: XGBoost's deep interaction modeling and CatBoost's robust handling of categorical and sparse data give them a clear edge over LightGBM in this AQI prediction task.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 10, June 2025
Impact Factor: 7.67

SHAP Analysis

the SHAP Analysis fig 7,8,& 9 shows offers interpretability into feature contributions. XGBoost emerges as particularly effective in explaining model behavior due to its explicit feature interaction modeling, showing high SHAP values for interaction terms like Population_Not_Staying_at_Home × mil_miles. CatBoost leverages its native handling of categorical and geospatial features (like latitude/longitude) via ordered boosting to highlight spatial influence on AQI. LightGBM, while fast, demonstrates lower explanatory power in SHAP values, often focusing more on weather features, possibly due to its GOSS-based sampling which can underrepresent rare but impactful cases. This comprehensive analysis affirms that CatBoost leads in stability and balanced accuracy, XGBoost in interpretability and variance explanation, and LightGBM in speed, albeit with slightly lower accuracy for AQI prediction tasks.

Residuals vs Predicted Value around zero. This behavior reflects their strong capacity for capturing nonlinear and interaction effects in the data. In contrast, LightGBM shows slight curvature in residuals, which may indicate model bias or insufficient learning of complex dependencies, potentially due to its aggressive leaf-wise splitting strategy.

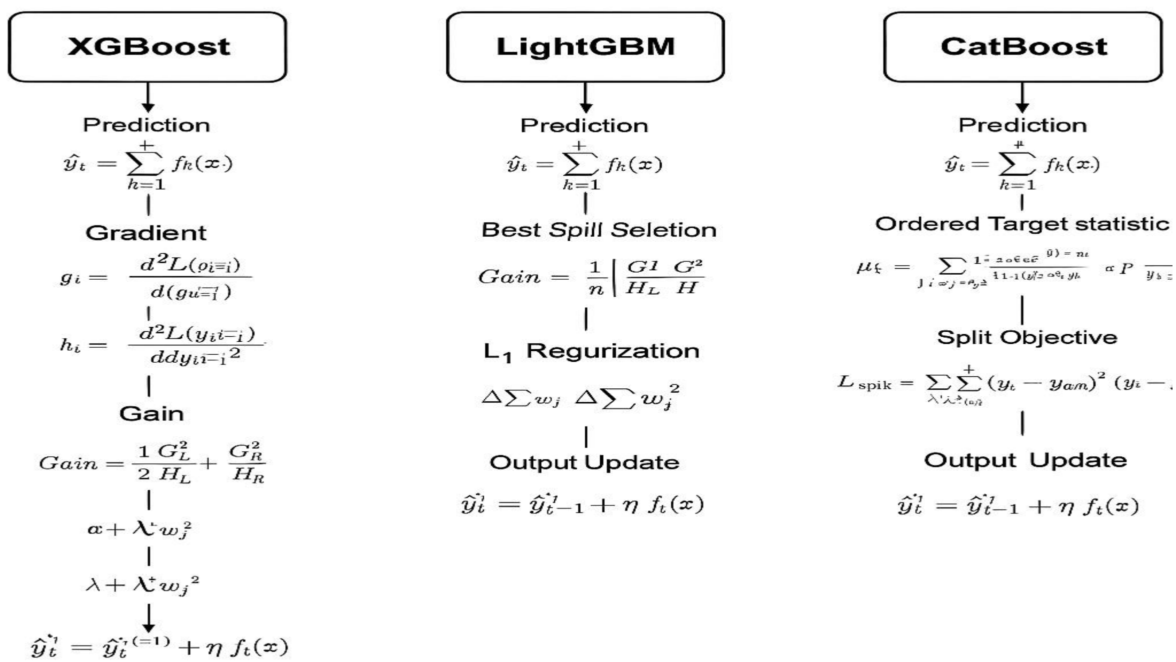

Fig. 4. Comparison algorithm of 3 models

Table 2. Summary of Key Differences

Aspect	XGBoost	LightGBM	CatBoost
Tree Growth	Level-wise	Leaf-wise with GOSS	Symmetric, oblivious splits
Categorical Handling	Requires encoding	Built-in	Ordered target statistics
Regularization	L1/L2 (α,λα,λ)	L1/L2 (α,λα,λ)	Strong L2 (λ=5λ=5)
Efficiency	Moderate	High (via sampling)	High (ordered boosting)

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 10, June 2025
Impact Factor: 7.67

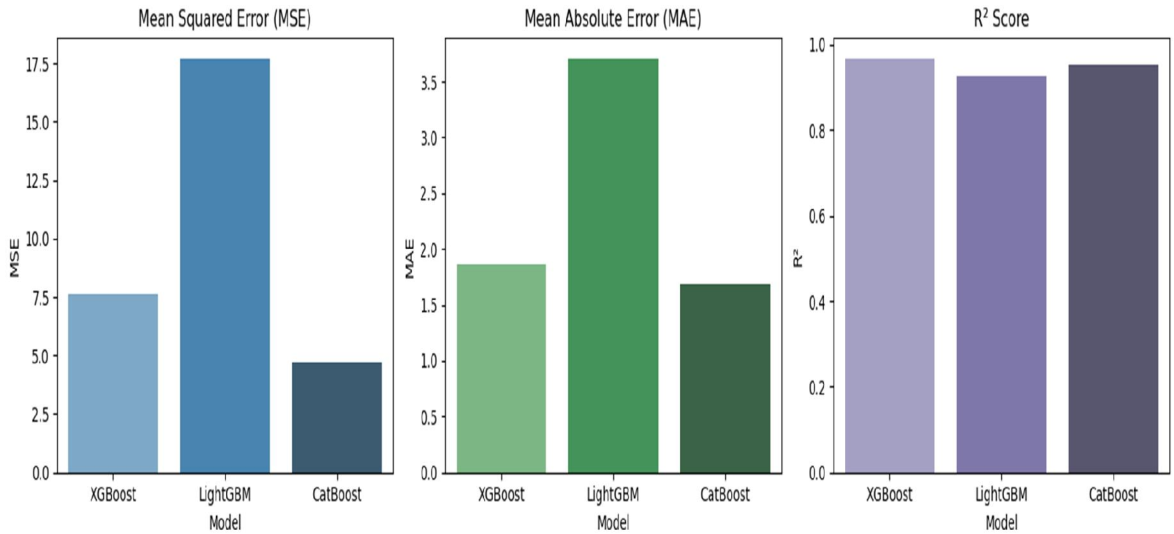

Fig. 5. performance summary of ML Models

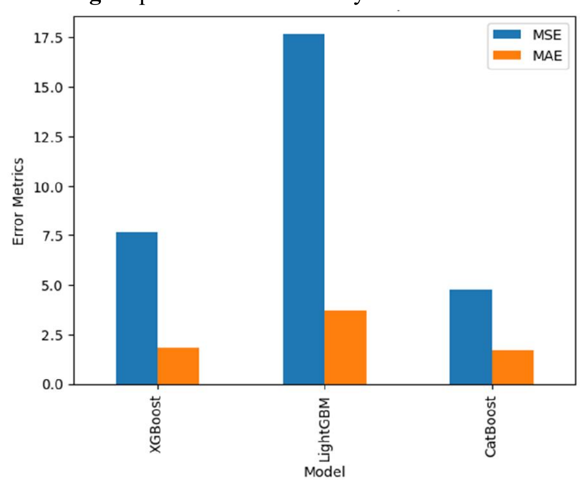

Fig. 6. ML Model performance Comparison

DOI: 10.48175/IJARSCT-28835

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 10, June 2025
Impact Factor: 7.67

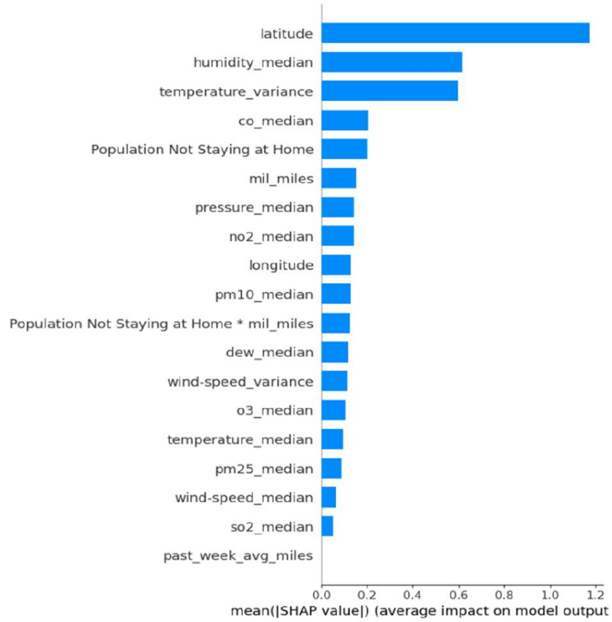

Fig. 7. SHAP Analysis on Dataset Feauter

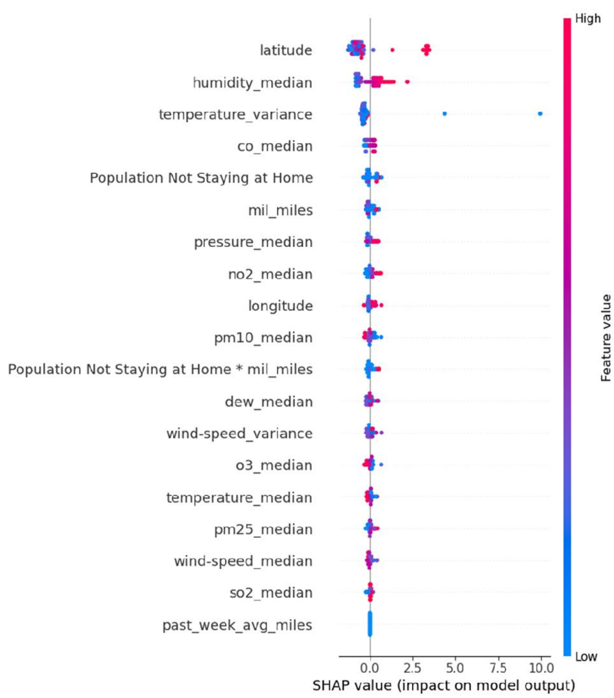

Fig. 8. Mean SHAP value average impact on model output

DOI: 10.48175/IJARSCT-28835

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 10, June 2025
Impact Factor: 7.67

Distribution: Actual vs. Predicted

The fig 6 shows Comparing the distributions of actual and predicted AQI values assesses how well the models capture the overall structure of the target variable. CatBoost most closely replicates the actual AQI distribution, evidencing its capability to model the target distribution effectively using ordered boosting and symmetric tree structures. XGBoost follows closely, with minor deviations in peak densities. LightGBM, while computationally efficient, demonstrates a broader distribution mismatch, possibly due to sampling-induced biases or reduced performance on less frequent AQI ranges.

mil_miles, which significantly influenced AQI variability. LightGBM lagged slightly (MSE: 17.69, R^2: 0.928), as its leaf-wise growth and GOSS sampling prioritized computational speed over precision, though it better captured temperature variance and dew point effects. SHAP analysis universally identified PM2.5, O3, and NO2 as top pollutant drivers, while geospatial clusters (latitude/longitude) and

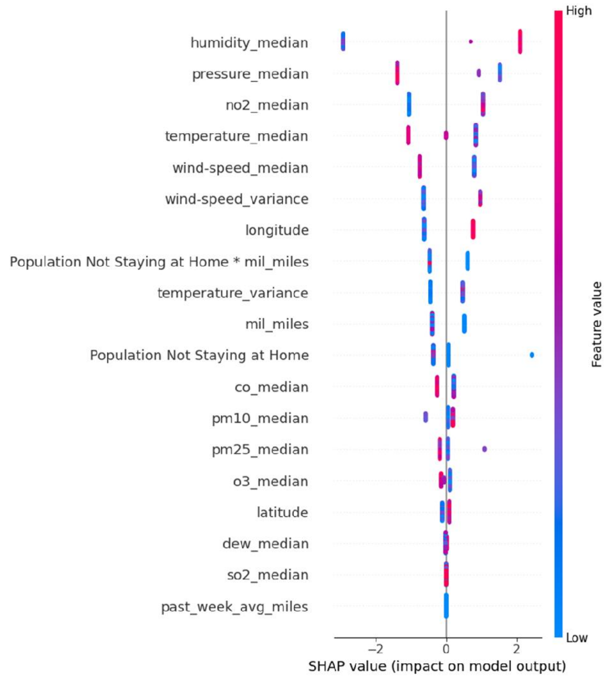

Fig.9. SHAP value impact on model output

mobility metrics (mil_miles) were critical for spatial-temporal patterns. Temporal cross-validation ensured robustness against time-dependent biases, revealing CatBoost's superiority in handling longitudinal data. These results underscore the trade-offs between stability (CatBoost), interpretability (XGBoost), and efficiency (LightGBM), positioning the framework as a versatile tool for urban air quality management, particularly in designing targeted interventions for pollution hotspots

III. CONCLUSION

In this research, we constructed and tested a hybrid machine learning model to forecast composite Air Quality Index (AQI) by combining spatiotemporal patterns from meteorological, pollutant, human activity, and geospatial factors. Our solution used three advanced ensemble models—XGBoost, LightGBM, and CatBoost—to capture the interactive and nonlinear relationships between multiple air pollutants, such as PM2.5, O_3, and NO_2, which are important elements in calculating AQI. Robust preprocessing operations were applied, such as missing value replacement, feature standardization, and category encoding. Temporal cross-validation was followed to maintain model generalization stability over time. Performance metrics calculated with various multiple regression metrics (MSE, MAE, R^2) indicated that CatBoost provided the minimum mean squared error, ensuring excellent stability and generalization capability.

IJARSCT

International Journal of Advanced Research in Science, Communication and Technology

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

ISSN: 2581-9429 Volume 5, Issue 10, June 2025 Impact Factor: 7.67

XGBoost produced the highest R^2 score, establishing higher explanatory capabilities, especially while modeling feature interaction. LightGBM, though effective, showed comparatively lower predictive accuracy because of its aggressive sampling approach and data imbalance sensitivity.

Residual and distribution analyses also confirmed model behaviors, with CatBoost and XGBoost displaying balanced error distribution and low heteroscedasticity. SHAP-based interpretability tests highlighted major domain findings: pollutant concentrations, particularly PM2.5 and O_3, were leading predictors, whereas human mobility indicators—like Population Not Staying at Home and its interaction with mil_miles—played a significant role in explaining emission variability. Geospatial features also had substantial impacts on predictions, especially under CatBoost's ordered boosting scheme. In general, this work illustrates the utility of integrating ensemble learning with spatiotemporal feature engineering to improve AQI prediction. The hybrid approach not only enhances predictive accuracy but also improves interpretability, thus facilitating data-driven environmental policy-making and urban planning. Future research can investigate multimodal data fusion (e.g., satellite images, traffic flow), real-time model updating, and deployment in edge computing platforms for scalable, localized air quality monitoring.

REFERENCES

[1]. Manisalidis, I., Stavropoulou, E., Stavropoulos, A., & Bezirtzoglou, E. (2020). Environmental and health impacts of air pollution: a review. Frontiers in public health, 8, 14.

[2]. Wenhao Chu, Chunxiao Zhang, Heng Li, Laifu Zhang, Dingtao Shen, Rongrong Li, SHAP-powered insights into spatiotemporal effects: Unlocking explainable Bayesian-neural-network urban flood forecasting, International Journal of Applied Earth Observation and Geoinformation,Volume 131,2024,103972, ISSN 1569-8432 Ke, G., et al. (2017).

[3]. Ramraj, Santhanam, et al. "Experimenting XGBoost algorithm for prediction and classification of different datasets." International Journal of Control Theory and Applications 9.40 (2016): 651-662. Kumar, P., et al. (2022). Machine learning for predicting air quality: A review. Environmental Pollution, 305, 119244.

[4]. Phillips, Peter CB. "Multiple regression with integrated time series." (1987). Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30.

[5]. Nahdliyah, N., et al. "A MGSTAR: An Extension of The Generalized Space-Time Autoregressive Model." Journal of Physics: Conference Series. Vol. 1752. No. 1. IOP Publishing, 2021. WHO. (2021). Ambient (outdoor) air pollution. World Health Organization.

[6]. Laudani, Antonino, et al. "On training efficiency and computational costs of a feed forward neural network: A review." Computational intelligence and neuroscience 2015.1 (2015): 818243.

[7]. Zemouri, Ryad, et al. "A new growing pruning deep learning neural network algorithm (GP-DLNN)." Neural Computing and Applications 32.24 (2020): 18143-18159

[8]. Graves, Alex, and Alex Graves. "Long short-term memory." Supervised sequence labelling with recurrent neural networks (2012): 37-45..

[9]. Akinyele, Daniel, Juri Belikov, and Yoash Levron. "Challenges of microgrids in remote communities: A STEEP model application." Energies 11.2 (2018): 432.

[10]. Bhattacharyya, Abhijit, et al. "A deep learning based approach for automatic detection of COVID-19 cases using chest X-ray images." Biomedical Signal Processing and Control 71 (2022): 103182.

[11]. Li, Qun, et al. "Early transmission dynamics in Wuhan, China, of novel coronavirus–infected pneumonia." New England journal of medicine 382.13 (2020): 1199-1207.