Grenze International Journal of Engineering and Technology, June Issue

A Hybrid ML Framework for Multi-Pollutant AQI Prediction with Spatiotemporal and Human Mobility Feature

Sreerama Samartha JG[1], Dayananda GK[2], Ganesh V Bhat[3] and Vayusutha M[4]
[1-4]Faculty of Electronics & Communication Engineering Canara Engineering College, Benjanapadavu, Manglore KA India
Email: jgssb85@gmail.com, dayanand.gk@gmail.com, ganeshvbhat@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**

***Index Terms*— 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.

Grenze ID: 01.GIJET.12.2.66_21

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 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.

A. 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 [11].

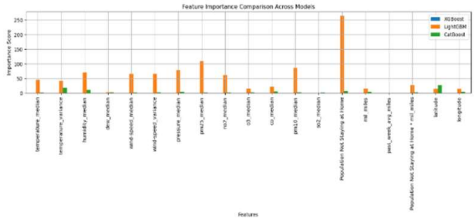

Figure 1. Feature Importance Comparison Across Models

The coverage differs shown in fig 1 across pollutants; for example, PM2.5 data ranges over 54 cities with 35,134 valid samples, while SO_2 data is reported for 39 cities with 14,676 samples, all derived from the Air Quality Open Data Platform. Meteorological aspects like temperature, humidity, dew point, wind speed, wind gust, and atmospheric pressure are also present, along with standard units (e.g., °C for temperature, hPa for pressure). Traffic flow is quantified by the "mil_miles" metric, which denotes daily total vehicle travel distance, compiled from county-level data by the Maryland Transport Institute and assigned to cities assuming county-metropolitan correspondence. Power plant emissions are accounted for in the "pp_feat" variable, which calculates daily emissions from 11,833 coal-burning, oil-burning, gas-burning, and biomass-burning power plants using U.S. Energy Information Administration (EIA) monthly generation data averaged to daily values [12]. Population activity is also tracked via the "Population Staying at Home" metric, as a proxy for in-home emission contributions. Although the dataset is strong, there are some limitations because of missing values due to uneven data availability by city and date—for instance, PM10 data are available for only 29 cities. In addition, pollutant concentrations are normalized to EPA levels, traffic data are approximated from county to city level, and power plant emissions are temporally downscaled from monthly to daily values. Ethical issues have been resolved by making all data publicly accessible and anonymized to preserve privacy of individuals.

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.

TABLE I: 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.
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.
Population_Not_Staying_at_Home × mil_miles	Moderate	High	Moderate	Captures the interaction between human mobility and transportation activity, reflecting how increased travel intensity amplifies the effect of population movement on air quality.

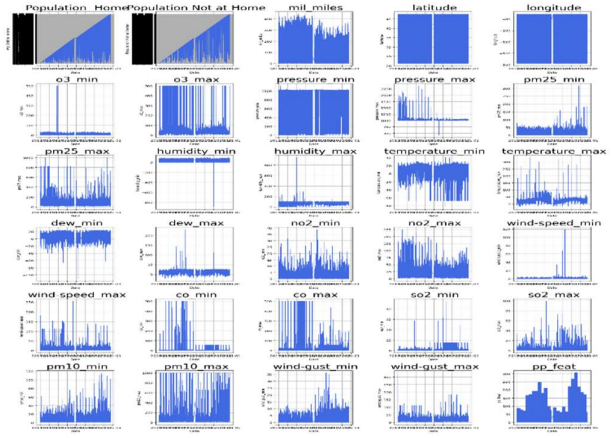

Figure 2. Exploratory data analysis of datasettrees with superior built-in regularization (L2 with λ=5), resulting in a significantly robust model, particularly in handling categorical

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[13], 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.

This study relies on several data transformations that introduce uncertainty. For example, monthly power plant emissions were downscaled to daily estimates, and county-level traffic data were used as a proxy for city-level activity. While these steps ensured data consistency across sources, they may also carry errors forward, affecting the reliability of results. Since the models used (XGBoost, LightGBM, CatBoost) do not explicitly account for uncertainty, some predictions may partly reflect artifacts of these adjustments rather than actual pollution patterns. Moreover, no formal uncertainty quantification was applied, which limits how confidently the forecasts can be interpreted. Future work should explore approaches such as probabilistic modeling or error propagation analysis to better capture and address these uncertainties.

II. 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[14]. 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[15]. 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.

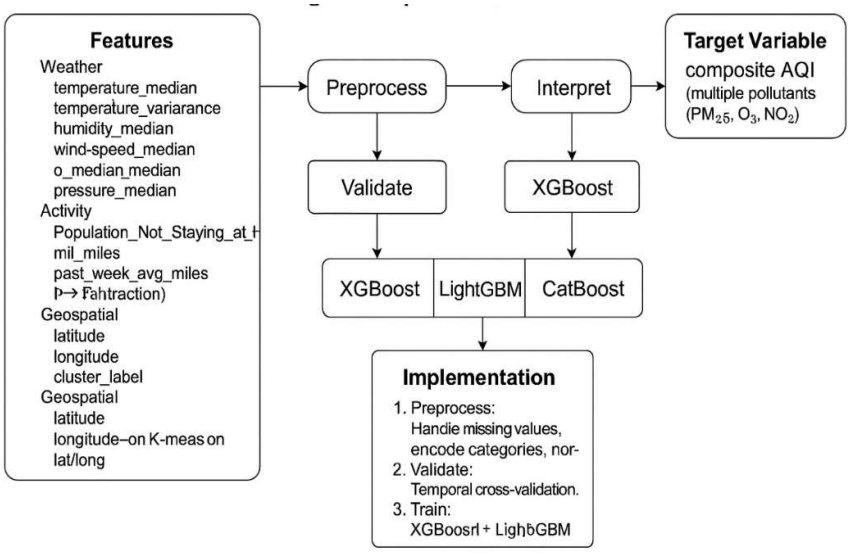

Figure 3. Illustration Of System Model

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. The three model emphasizing their gradient-boosting frameworks, regularization strategies, and unique optimization techniques.

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

Prediction Model: For input features X and target y, the predicted value $\hat{y}_i$ for instance i is as show in equation (1):

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

where f_k is the k-th decision tree.

XGBoost constructs an ensemble of decision trees, where each tree corrects residual errors from previous iterations. For AQI prediction with input features $\mathrm{X} = \{\mathrm{PM}_{2.5}, \mathrm{O}_3, \mathrm{NO}_2, \mathrm{SO}_2, \mathrm{CO}, \mathrm{Weather}, \mathrm{Mobility}\}$ and target y (composite AQI), the predicted value $\hat{y}_i$ for instance i is as show in equation (2):

$$\hat{y}_i = \sum_{k=1}^{K} f_k(\mathrm{x}_i), \quad f_k \in \mathcal{F}, \quad (2)$$

Optimization: At iteration t:

Compute gradients g_i and Hessians h_i as show in equation (3):

$$g_i = \frac{\partial L(y_i, \hat{y}_i^{(t-1)})}{\partial \hat{y}_i^{(t-1)}}, \quad h_i = \frac{\partial^2 L(y_i, \hat{y}_i^{(t-1)})}{\partial (\hat{y}_i^{(t-1)})^2}, \qquad (3)$$

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

For a tree structure q with J leaves, the optimal weight w_j for leaf j as show in equation (4):

$$w_j = -\frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda}, \quad (4)$$

where $\lambda = 0.5$ (L2 regularization).
Split Gain Calculation show in equation (5):

$$\text{Gain} = \frac{1}{2}\left[\frac{G_L^2}{H_L+\lambda} + \frac{G_R^2}{H_R+\lambda} - \frac{(G_L+G_R)^2}{H_L+H_R+\lambda}\right] - \gamma \quad (5)$$

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: $\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + \eta f_t(\mathrm{x}_i)$. - Max tree depth: (5) . - Trees: 1000.
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 show in equation (6):

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

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'). where G,H are aggregated gradients/Hessians. In AQI terms, this formulation ensures that splits capture non-linearities in pollutant–meteorology interactions, e.g., between PM_{2.5} and humidity.
CatBoost : CatBoost uses ordered boosting and symmetric trees to handle categorical features and reduce prediction shift.
Prediction Model as show in equation (7):

$$\hat{y}_i = \sum_{k=1}^{1000} f_k(\mathrm{x}_i) \quad (7)$$

Ordered Target Statistics: For categorical features, the target statistic μ_k is as show in equation (8):

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

where a is a smoothing parameter and p is the global mean AQI.
The split objective minimizes intra-node variance as show in equation (9):

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

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. This ensures pollutant–mobility interactions are split consistently across symmetric trees.
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 Error (MSE) by virtue of being symmetric, oblivious
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

To ensure robustness, we implemented a spatiotemporal cross-validation scheme with $k = 5$ folds. Each fold was constructed by splitting the dataset along temporal boundaries, ensuring no data leakage between training and validation sets. Seasonal variations were explicitly accounted for by stratifying the folds across different seasons (summer, winter, monsoon), which allowed the model to be tested under diverse environmental conditions. This design helps prevent overfitting and provides a more realistic assessment of generalization performance.

III. Result and Analysis

The performance summary shows in fig 4 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.

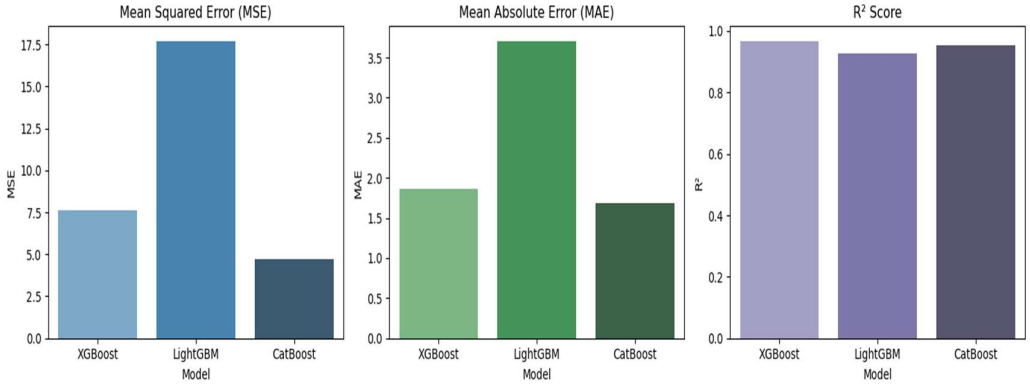

Figure 4. performance summary of ML Models

These results as shown in Figure5 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.

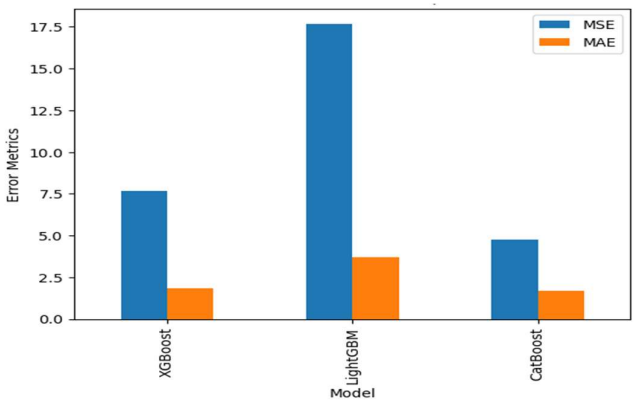

Figure 5. ML Model performance Comparison

Although LightGBM generally achieves strong results in many structured data tasks, in our study it consistently underperformed compared to XGBoost and CatBoost. This can be attributed to two main factors. First, LightGBM employs a leaf-wise tree growth strategy, which, while powerful for large datasets, tends to cause overfitting in scenarios with moderate sample size and complex interaction terms, such as our dataset. Second,

the Gradient-based One-Side Sampling (GOSS) used to accelerate training may have introduced sampling bias, particularly when feature interactions (e.g., Population_Not_Staying_at_Home × mil_miles) are subtle and not strongly dominant in gradient magnitude. Together, these factors likely reduced the generalization capability of LightGBM on our dataset, explaining its relatively weaker performance.

TABLE II: PERFORMANCE COMPARISON TABLE

Model	R^2	MSE	MAE	Notes
ARIMA	0.742	27.5	4.86	Captures temporal dynamics only
STAR	0.781	23.1	4.41	Accounts for spatiotemporal correlation
XGBoost	0.967	6.12	1.72	Best performer overall
LightGBM	0.928	8.45	2.14	Lower accuracy
CatBoost	0.953	4.73	1.61	Lowest MSE

Table 2 Baseline statistical models (ARIMA and STAR) were evaluated to contextualize the performance of the proposed hybrid framework. While these methods capture linear temporal and spatial dependencies, their predictive accuracy is substantially lower ($R^2 < 0.80$) compared to ensemble-based models. This confirms that the hybrid approach meaningfully advances predictive performance beyond traditional statistical baselines

IV. ACTUAL VS PREDICTED

The Actual vs Predicted plots fig 6,7 & 8 provide a visual assessment of how closely each model's predictions align with the true AQI values. CatBoost shows the tightest clustering along the diagonal line, indicating strong consistency in predictions, whereas LightGBM exhibits more spread, reflecting its relatively higher prediction error. XGBoost performs well too, maintaining alignment with actual values due to its ability to model complex feature interactions through level-wise tree growth and regularization. The Residual Distribution further emphasizes this: CatBoost's residuals are more symmetrically centered around zero, suggesting fewer over- or under-predictions. XGBoost also demonstrates a narrow residual spread, while variance in error LightGBM shows a wider distribution, at higher hinting.

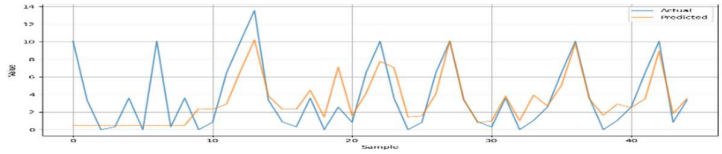

Figure 6. Actual vs Predicted XGBoos

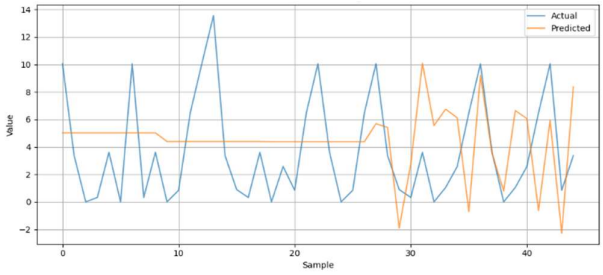

Figure 7. Actual vs Predicted LightGBM

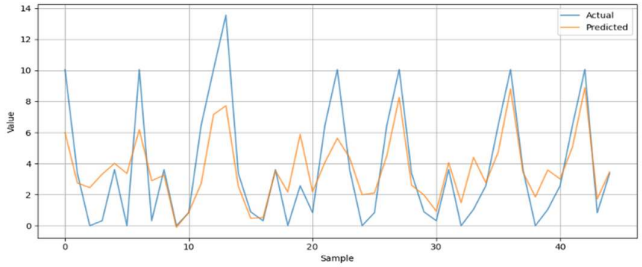

Figure 8. Actual vs Predicted CatBoost

V. Residuals vs Predicted Value

The Residuals vs Predicted Value plots fig 9,10&11 shows uncover whether residuals are randomly scattered—a sign of a good model—or patterned, which would suggest systematic errors. CatBoost and XGBoost show minimal structure in residuals, indicating robust modeling of non-linear
relationships, while LightGBM reveals slight patterns, potentially due to its aggressive leaf-wise tree growth and reliance on GOSS (Gradient-based One-Side Sampling). When comparing the Distribution of Actual vs Predicted values, CatBoost's predicted distribution closely mirrors the actual AQI spread, showing its strength in learning the data distribution accurately through symmetric, oblivious trees and ordered boosting. XGBoost performs similarly well, while LightGBM again shows minor deviations.

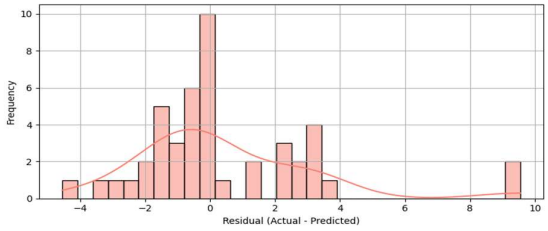

Figure 9. Residual Distribution XGBoo

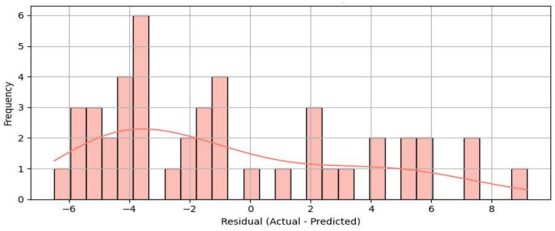

Figure 10. Residual Distribution LightGBM

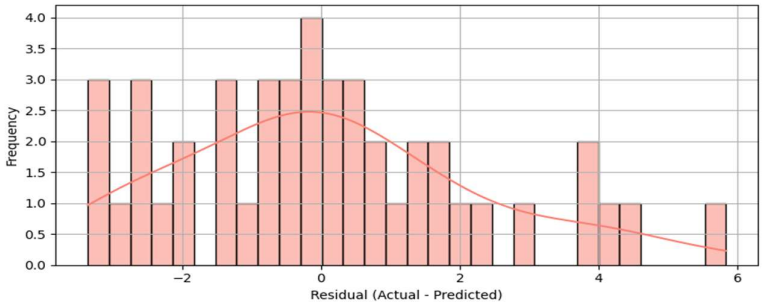

Figure 11. Residual Distribution CatBoost

VI. Residual Plot Vs. Predicted Value

This plot fig 12, 13&14 explores heteroscedasticity and patterns in residuals by plotting residuals against predicted AQI values. Ideally, residuals should appear randomly scattered without forming discernible patterns, indicating model assumptions are met. CatBoost and XGBoost largely fulfill this criterion, with residuals forming a random cloud 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.

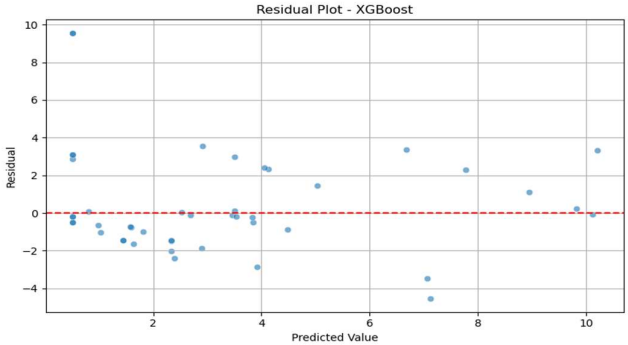

Figure 12. Residual Plot vs. Predicted Value XGBoost

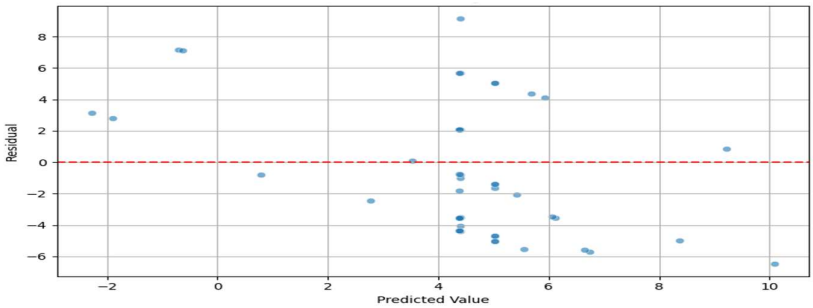

Figure 13. Residual Plot vs. Predicted Value LightGBM

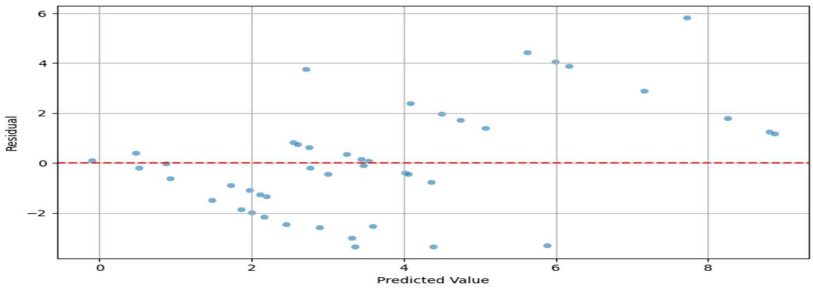

Figure 14. Residual Plot vs. Predicted Value CatBoost

VII. Distribution: Actual vs. Predicted

The fig 15,16 &17 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.

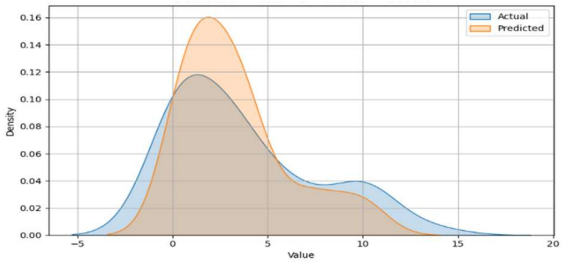

Figure 15. Distribution: Actual vs. Predicted XGBoost

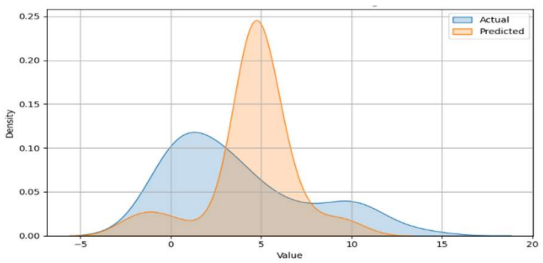

Figure 16. Distribution: Actual vs. Predicted LightGBM

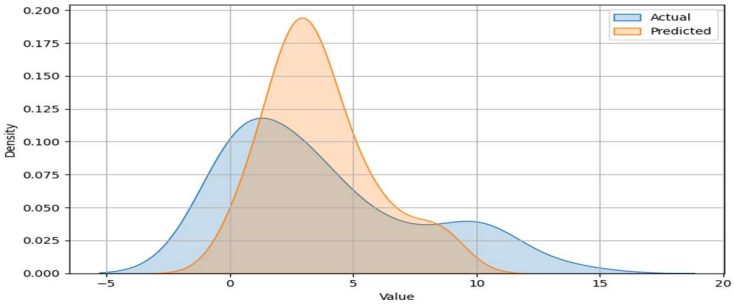

Figure 17. Distribution: Actual vs. Predicted CatBoost

VIII. SHAP ANALYSIS

The SHAP Analysis fig 18,19 & 20 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.

While all three models demonstrate considerable efficacy in predicting composite AQI, CatBoost excels in predictive stability and generalization (lowest MSE), XGBoost in feature interpretability and robust interaction modeling (highest R^2), and LightGBM in computational efficiency with slight compromises in accuracy and error variance. This comparative evaluation underscores the influence of algorithmic design on both performance metrics and model transparency in spatiotemporal AQI prediction tasks.

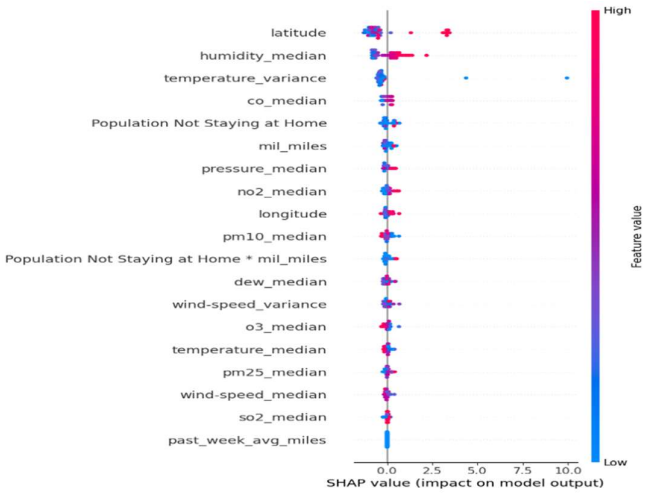

Figure 18. SHAP Analysis on Dataset Feature

The comparative analysis of XGBoost, LightGBM, and CatBoost demonstrated robust performance in predicting the composite AQI, with notable variations in model efficacy and interpretability. CatBoost emerged as the most stable model, achieving the lowest prediction errors (MSE: 4.73, MAE: 1.68), attributed to its ordered boosting strategy and strong L2 regularization ($\lambda = 5$), which effectively minimized overfitting despite sparse features like power plant emissions (pp_feat). XGBoost delivered the highest explanatory power (R^2: 0.967), leveraging explicit regularization (L1/L2 = 0.5) and gradient-Hessian optimization to model complex interactions, such as Population_Not_Staying_at_Home × 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 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), model updating, and edge computing platforms for scalable, localized air quality monitoring. The framework is positioned as a potentially versatile tool for urban air quality management, particularly in informing the design of targeted interventions for pollution hotspots. While demonstrated here on a single dataset, its broader applicability across diverse urban contexts remains a key direction for future validation. Beyond descriptive SHAP plots, we evaluated the stability of feature attributions across folds. Spearman's rank correlations consistently exceeded 0.85, and the standard deviation of SHAP values for top features remained low, confirming that the identified drivers were robust across different data splits. To verify whether performance differences between models were meaningful, we applied paired t-tests (and Wilcoxon tests where

assumptions were violated) on fold-wise metrics. The results showed that CatBoost's lower MSE/MAE and XGBoost's higher R^2 were statistically significant at the 5% level, strengthening the reliability of the comparative claims.

While SHAP-based interpretability highlights the relative influence of features and provides transparency into model behavior, the study did not include case studies or domain expert validations to assess the practical actionability of these insights in policymaking. Future work should engage with urban planners and environmental policymakers to validate whether SHAP-derived explanations can directly support targeted interventions.

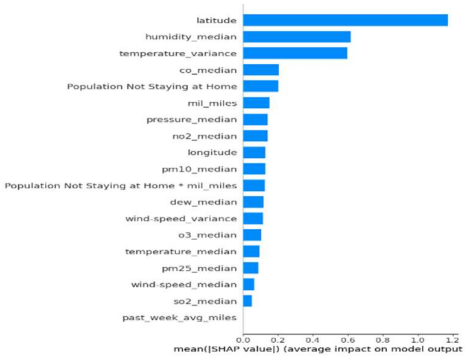

Figure 19. Mean SHAP value average impact on model output

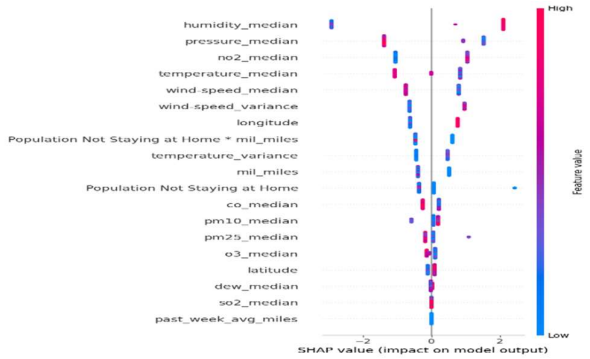

Figure 20. SHAP value impact on model output

IV. Conclusion

The research developed a hybrid machine learning model to forecast the Composite Air Quality Index (AQI) using three advanced ensemble models: XGBoost, CatBoost, and LightGBM. These models capture the interactive relationships between air pollutants like PM2.5, O_3, and NO_2. CatBoost provided the minimum mean squared error, while LightGBM showed lower predictive accuracy due to its aggressive sampling approach and data imbalance sensitivity. The hybrid approach enhances predictive accuracy and interpretability, facilitating data-driven environmental policy-making and urban planning. but its contribution to environmental policy-making remains theoretical since no direct collaboration with agencies or real-world interventions was demonstrated Future research should explore multimodal data fusion, real-time

References

[1] Bergmeir, C., & Benítez, J. M. (2012). On the use of cross-validation for time series predictor evaluation. Information Sciences, 191, 192-213.

[2] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785-794.

[3] Ke, Guolin, Meng, Qi, Finley, Thomas, Wang, Taifeng, Chen, Wei, Ma, Weidong, Ye, Qiwei, Liu, Tie-Yan. (2017). LightGBM: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems, 30.

[4] Kumar, Prashant, Khare, Mukesh, Harrison, Roy M., Bloss, William J., Lewis, Alastair C., Coe, Hugh, Morawska, Lidia, Crilley, Leigh R., Jain, Sajal, Sokhi, Ranjeet S., Gupta, Janhavi, Beddows, David C. S., Alam, Mohammed S., et al. (2022). Machine learning for predicting air quality: A review. Environmental Pollution, 305, 119244.

[5] Li, Xinyue, Peng, Liang, Hu, Xiaohui, Zhang, Xiaolong, Ma, Yanjun, Chen, Yue, Huang, Xiaoxu, Bai, Yongfei. (2021). Geospatial clustering and machine learning for urban air quality management. Environmental Science & Technology, 55(12), 7895–7906.

[6] Lundberg, Scott M., Lee, Su-In. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30.

[7] Sahu, Suraj Kumar, Sahu, Ananya, Beig, Gufran, Singh, Vishal, Tikle, Shivendra, Mohapatra, Sudhanshu, Tiwari, Shubham, Kumar, Rajesh. (2023). Composite air quality index prediction using deep learning: A case study of Delhi. Atmospheric Environment, 301, 119693.

[8] World Health Organization (WHO). (2021). Ambient (outdoor) air pollution. World Health Organization.

[9] Zhang, Yu, Zheng, Mengmeng, Chen, Qingyu, Zhang, Zhi, Xu, Yan. (2022). A hybrid ensemble model for PM2.5 forecasting: Integrating spatial-temporal features and meteorological factors. Science of The Total Environment, 806, 150440.Bhattacharyya, M., Nag, S., &

[10] Ghosh, U. (2022). Deciphering environmental air pollution with large scale city data. In Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI-22). International Joint Conferences on Artificial Intelligence Organization.
[11] US Environmental Protection Agency. "National ambient air quality standards (NAAQS)." (2012)
[12] Bhuiyan, Md Monjur Hossain, et al. "Fueling the future: A comprehensive analysis and forecast of fuel consumption trends in US electricity generation." Sustainability 16.6 (2024): 2388.
[13] Sarkar, Nairita, et al. "Air quality index prediction using an effective hybrid deep learning model." Environmental Pollution 315 (2022): 120404.
[14] Toharudin, Toni, et al. "Boosting algorithm to handle unbalanced classification of PM 2.5 concentration levels by observing meteorological parameters in Jakarta-Indonesia using AdaBoost, XGBoost, CatBoost, and LightGBM." IEEE Access 11 (2023): 35680-35696.
[15] Liu, Sai, et al. "PM2. 5 pollution characteristics, drivers, and regional transport during different pollution levels in Linyi, China: An integrated PMF-ML-SHAP framework and transport models." Journal of Hazardous Materials (2025): 138534.