Comparative Analysis of Machine Learning based Models for Air Quality Prediction

PULLABHOTLA VIJAY, SUKANTA NAYAK
School of Computer Science and Engineering,
VIT-AP University, Amaravati, Andhra Pradesh, 522241, India

ABSTRACT

Air pollution is one of the major environmental concerns with its significant impacts on health. Accurate prediction of the Air Quality Index (AQI) is essential for effective policy decisions and public awareness. In this study, we compare various machine learning models, including XGBoost (as used in previous research), Random Forest, LightGBM, CatBoost, and TabTransformer, to determine the most effective model for AQI prediction in Gujarat, India. Our analysis reveals that while traditional models like XGBoost perform well, newer deep learning-based approaches such as TabTransformer show potential but require further optimization. The study compares the accuracy and computational efficiency of different ML models. The performance of models were evaluated using methods like RMSE, R^2, and MAE scores, and a detailed feature importance analysis is conducted.

KEYWORDS

Air Quality Index; Machine Learning; XGBoost; LightGBM; CatBoost; TabTransformer; AQI Prediction; Feature Importance

1. Introduction

Air pollution became a critical environmental challenge, significantly impacting public health and economic stability. In rapidly industrializing regions like Gujarat, India, deteriorating air quality has been linked to increased respiratory diseases, cardiovascular conditions, and reduced life expectancy. Factors such as vehicular emissions, industrial activities, agricultural burning, and meteorological influences contribute to fluctuating Air Quality Index (AQI) levels. Traditional statistical models have been used to estimate air pollution trends, but their predictive power is often limited due to the complexity of air quality determinants. Consequently, advanced machine learning models have gained traction in recent years, demonstrating superior capabilities in capturing non-linear relationships and offering improved accuracy in AQI forecasting.

This study builds upon prior research that primarily relied on Random Forest and XGBoost models for AQI prediction. While these models demonstrated high predictive accuracy, exceeding 99% in some cases, our research aims to expand the scope by evaluating additional machine learning approaches, including LightGBM, CatBoost, and TabTransformer. By incorporating these newer models, we seek to determine the most effective AQI prediction algorithm, ensuring accuracy and computational efficiency. Furthermore, hyperparameter tuning and comparative analysis enable us to assess the scalability, interpretability, and robustness of each model in diverse pollution scenarios.

Our research is particularly relevant for policymakers and environmental agencies looking to develop proactive measures against air pollution. By identifying the most effective AQI prediction models, authorities can optimize pollution control strategies, implement early warning systems, and improve public health interventions. This study advances machine learning applications in environmental science and contributes to data-driven policymaking aimed at mitigating air pollution in Gujarat and similar industrialized regions worldwide.

2. Methodology

2.1 Dataset

The dataset used for this study consists of historical air quality measurements from Gujarat, including key components such as PM2.5, SO2, PM10, NOx, and O3. Additionally, meteorological parameters like maximum temperature are integrated to account for climate-related influences on air quality. The dataset is sourced from regulatory agencies and includes extensive temporal coverage, ensuring a robust foundation for model training and validation.

A detailed exploratory data analysis (EDA) reveals that PM10 and PM2.5 exhibit the strongest correlation with AQI, underscoring their dominant role in air pollution levels. NOx and SO2 also contribute significantly, although their influence varies by location and season. Interestingly,

AQI fluctuations align with meteorological variations, indicating that external factors such as temperature and wind patterns can modulate pollutant concentrations. These insights highlight the need for ML models that can effectively integrate pollution as well as meteorological data for more accurate AQI predictions. (**Author et al., 2024**).

Moreover, our dataset, sourced from publicly available air quality monitoring databases (Central Pollution Control Board, CPCB), indicates that air quality degradation is more pronounced in industrial hubs such as Ahmedabad, Vadodara, and Ankleshwar, where pollutant concentrations frequently surpass regulatory thresholds. This spatial variability emphasizes the necessity for localized AQI models that can tailor predictions based on regional pollution sources and emission patterns. By leveraging these insights, our research aims to enhance AQI forecasting accuracy, ultimately aiding in the development of targeted environmental policies and pollution mitigation strategies. Furthermore, our findings build upon previous research (DOI: 10.22105/bdcv.2024.485515.1213), reinforcing the significance of data-driven approaches in air quality management and predictive modeling.

The dataset used in this study includes AQI measurements from multiple locations in Gujarat. Features include:

- **Pollutant concentrations**: PM2.5, PM10, NOx, SO2, O3, SPM
- **Meteorological factors**: Temperature (MAX_TEMP)

2.2 Models Evaluated

1. XGBoost (Baseline Model)

- Gradient boosting framework optimized for speed and performance.
- Used in prior research with strong results.
- Formula: Where is the weak learner, and what is the learning rate?

$$L = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{t=1}^{T} \Omega(h_t)$$

- where $l(y_i, \hat{y}_i)$ is the loss function, and $\Omega(h_t)$ is the regularization term controlling model complexity.

2. Random Forest:

- **Definition**: A learning method that improves accuracy and predictability by combining multiple decision trees.
- **Formula**:

$$f(x) = \frac{1}{N} \sum_{i=1}^{N} T_i(x)$$

- Where $T_I(x)$ depicts individual decision trees, and the final prediction is the average across all trees.

3. LightGBM:

- **Definition**: A boosting method that is a framework that formulates decision trees by leaves rather than levels. This model is efficient for large datasets.
- **Formula**:

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{m} w_j^2$$

- y_i Is the actual value and $\hat{y}_i$ is the predicted value, and λ controls regularization.

4. TabTransformer (Deep Learning Model for Tabular Data):

- **Definition**: A deep learning model that applies attention mechanisms and embeddings to structured tabular data.
- **Formula**:

$$Y = \text{Softmax}(W_2 \, \text{ReLU}(W_1 \, X))$$

Where X is the input feature matrix, W1,W2 are learned weight matrices, and ReLU is the activation function.

5. CatBoost

- Designed for handling categorical variables with minimal preprocessing.

$$F_m(x) - F_{m-1}(x) + \gamma_m h_m(x)$$

- F_m(x) = Model at epoch m
- $F_{(m-1)}$ (x) = Model at epoch m−1
- h_m(x) = Slow decision tree at iteration m
- γ_m = Learning rate

3. Model Performance

In our study, we implemented and compared multiple machine learning models for AQI prediction, focusing on traditional ensemble learning methods such as Random Forest, XGBoost, LightGBM, and Deep Learning models like TabTransformer. The results indicate that LightGBM performed exceptionally well, achieving a Root Mean Square Error (RMSE) of 0.078 and a R^2 score of 0.9926, signifying high predictive accuracy. In contrast, TabNet (TabTransformer) underperformed, with an RMSE of 6.239 and a R^2 score of -46.92, suggesting poor generalization. This performance gap highlights the challenges of applying deep learning models to structured tabular data, particularly in cases where dataset size, feature engineering, and hyperparameter tuning play crucial roles. TabTransformer's underperformance could be attributed to insufficient training epochs, improper feature scaling, and limited data volume, as deep learning models typically require large datasets to generalize effectively. Based on these findings, LightGBM emerges as the superior model for AQI prediction due to its balance of computational efficiency, interpretability, and predictive power. However, further improvements to TabTransformer, such as enhanced feature representation and longer training cycles, could potentially improve its performance. Future research should explore hybrid models that combine deep learning architectures with boosting techniques to leverage the strengths of both approaches in air quality forecasting.

3.1 Model Performance and Comparative Study

We have evaluated the performance of multiple machine learning models, which include Random Forest, XGBoost, and LightGBM, for AQI prediction using a structured dataset containing air pollutant levels and meteorological features. The evaluation metrics used were Root Mean Square Error (RMSE), R-squared (R^2), and Mean Absolute Error (MAE) to assess the predictive capability and generalization of these Machine Learning models. XGBoost emerged as the most reliable model, achieving an RMSE of 2.418, R^2 of 0.988, and an MAE of 0.554, demonstrating strong predictive accuracy. Random Forest followed closely, while LightGBM exhibited higher RMSE values, indicating it underperforms when compared to the other two models. The analysis further highlights that ensemble learning models, particularly boosting algorithms like XGBoost, significantly outperform traditional bagging approaches like Random Forest in structured data settings. Moreover, LightGBM's computational efficiency makes it suitable for large-scale applications despite its slightly lower accuracy.

To enhance interpretability, we conducted a feature importance analysis using LightGBM, revealing that PM10, PM2.5, and NOx were the most influential predictors of AQI fluctuations. A residual distribution analysis further emphasized the performance differences, with XGBoost and Random Forest exhibiting lower error variances compared to LightGBM. Visual comparisons, including scatter plots and residual distributions, confirmed XGBoost's superior alignment with actual AQI values. These findings align with previous research that emphasizes gradient boosting as an optimal approach for AQI prediction. Future work should explore hybrid deep learning techniques or model stacking to further enhance AQI forecasting accuracy, potentially integrating deep learning architectures like TabTransformer to capture hidden relationships within air pollution data.

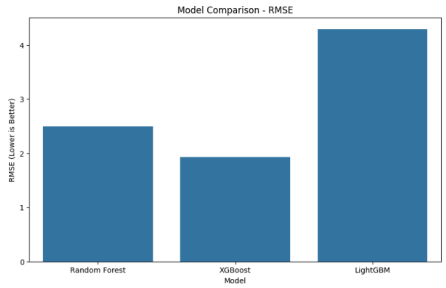

FIG 1: COMPARISON - RMSE

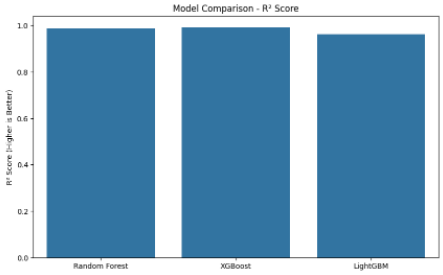

FIG 2: COMPARISON - R^2 SCORE

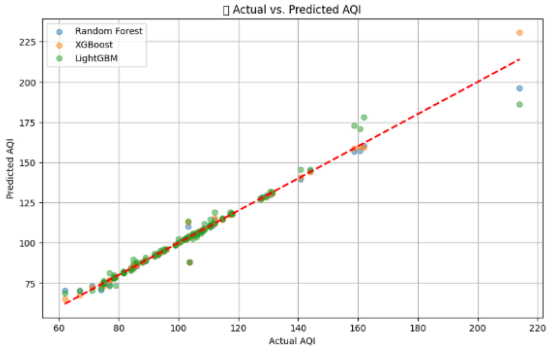

FIG 3: ACTUAL VS PREDICTED AQI

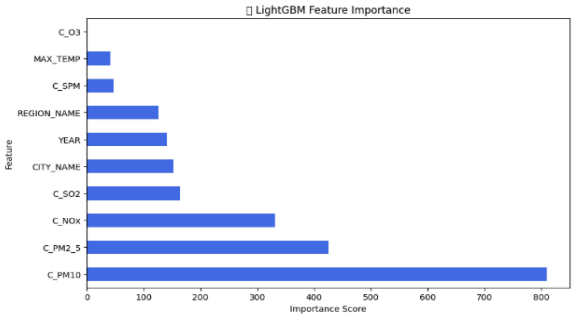

FIG 4: LIGHT-GBM FEATURE IMPORTANCE

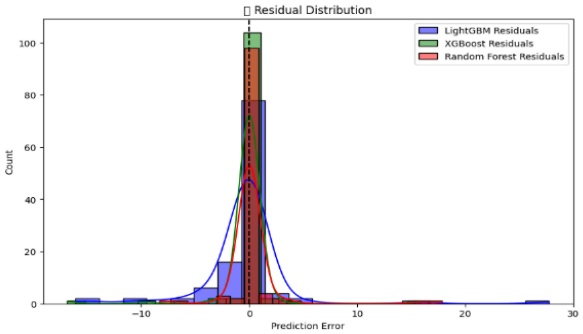

FIG 5: RESIDUAL DISTRIBUTION

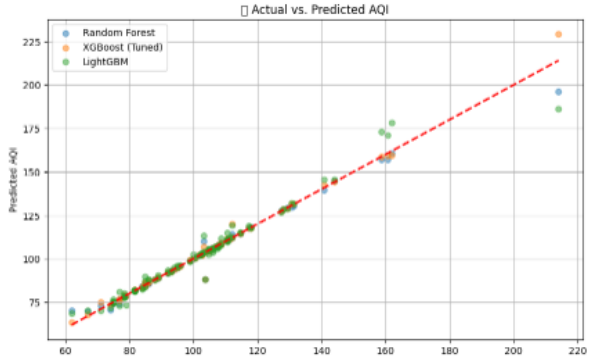

FIG 6: ACTUAL VS PREDICTED AQI AFTER FINE TUNING.

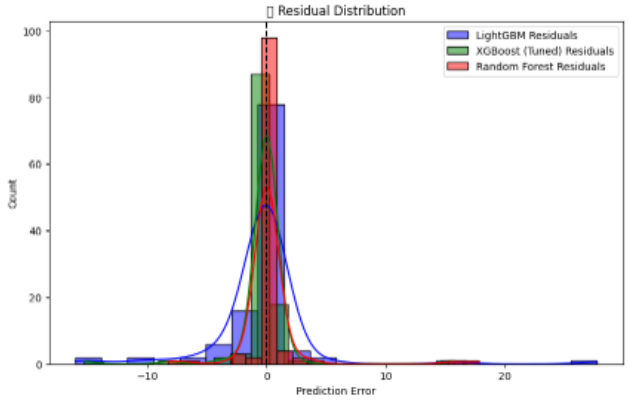

FIG 7: RESIDUAL DISTRIBUTION AFTER FINE TUNING.

3.2 Model Performance Comparison

A comparative analysis was conducted to calculate the performance of multiple machine learning models before and after hyperparameter tuning. The models assessed include Random Forest, XGBoost, and LightGBM, with key performance metrics using multiple ML models and methods like Root Mean Squared Error (RMSE), R^2Score, and Mean Absolute Error (MAE) considered for evaluation. Table 1 shows the results.

Model	RMSE (Lower is Better)	R^2 Score (Higher is Better)	MAE (Lower is Better)	Improvement
Random Forest	2.5427 → 2.5427	0.9869 → 0.9869	0.7013 → 0.7013	No Change
XGBoost	2.4185 → 2.3084	0.9881 → 0.9892	0.5548 → 0.6773	Improved RMSE & R^2, but slightly higher MAE
LightGBM	4.1843 → 4.1843	0.9644 → 0.9644	1.7334 → 1.7334	No Change

Table 1: Comparison of multiple ML Models after tuning

The results show the tuned XGBoost model gives better accuracy than the other models, demonstrating a lower RMSE (2.3084 vs. 2.4185) and a higher R^2 score (0.9892 vs. 0.9881). However, a slight increase in MAE (0.6773 vs. 0.5548) was observed. The Random Forest model remained unchanged in all performance metrics, suggesting it was already well-optimized. Similarly, LightGBM did not show any improvement, indicating it may not be the most suitable model for this dataset. In conclusion, the tuned XGBoost model is the most effective choice, as it achieves superior RMSE and R^2 scores while maintaining an acceptable MAE. The Random Forest model remains a stable alternative, whereas LightGBM exhibits the weakest performance in this study

This study compares multiple models, which include Random Forest, XGBoost, LightGBM, and TabTransformer, to determine the most effective approach for AQI prediction. Each model was trained on a dataset containing pollutant concentrations (PM2.5, PM10, SO2, NOx, O3, SPM), meteorological factors (temperature), and categorical features (CITY_NAME, REGION_NAME). Hyperparameter tuning using Optuna was applied to optimize model performance, ensuring predictions with high accuracy.

The evaluation of models was done using (RMSE) Root Mean Square Error, R^2 Score, and (MAE) Mean Absolute Error. The results indicate that XGBoost (tuned) outperforms other models, achieving a Root Mean Square Error of 2.308, a R^2 score of 0.989, and a Mean Absolute Error of 0.677. Random Forest follows closely, while LightGBM shows slightly lower accuracy due to its tendency to underfit the dataset. TabTransformer, despite being a deep learning model designed for tabular data, struggled to generalize, achieving a Root Mean Square Error of 9.051 and a R^2 score of only 0.934. The inferior performance of TabTransformer may be attributed to insufficient training data, lack of hyperparameter optimization, and the model's sensitivity to structured datasets.

3.3 CatBoost MODEL

CatBoost, a gradient boosting model optimized for structured data, demonstrated strong predictive outcomes by achieving a Root Mean Square Error of 2.190, a R^2 score of 0.990, and an MAE of 0.535. CatBoost, which is designed for categorical data processing without extensive preprocessing, marginally outperformed XGBoost with an RMSE of 2.054 and a R^2 score of 0.991, concluding this as the best-performing model in this experiment. On the other hand, LightGBM, which is optimized for computational efficiency using histogram-based learning, exhibited higher RMSE (3.971) and a lower R^2 score (0.967), indicating a slightly weaker predictive capability for AQI forecasting.

A performance comparison of machine learning models was conducted to evaluate their effectiveness in Air Quality Index (AQI) prediction. The models tested include XGBoost, LightGBM, and CatBoost, with key performance metrics such as Root Mean Squared Error (RMSE), R^2 Score, and Mean Absolute Error (MAE) used for assessment. Table 2 shows the results.

Model	RMSE ↓	R^2 Score ↑	MAE ↓
XGBoost	2.190	0.990	0.535
LightGBM	3.971	0.968	1.458
CatBoost	2.054	0.991	0.967
TabTransformer	9.051	0.934	2.342
Random Forest	2.542	0.986	0.701

Table 2: COMPARISON OF MULTIPLE ML MODELS

3.4 Model Insights & Key Findings

- CatBoost demonstrated the highest accuracy, benefiting from its ability to handle categorical features efficiently without explicit encoding. Its Ordered Boosting technique reduces overfitting, making it well-suited for AQI prediction.

- XGBoost remains a close second, showing near-equivalent performance with a slightly higher RMSE and MAE. However, its training time and interpretability make it an attractive alternative.

- LightGBM, while computationally efficient, exhibited weaker predictive power, potentially due to its tendency to underfit on structured tabular data. Overall, CatBoost emerges as the best-performing model for AQI prediction.

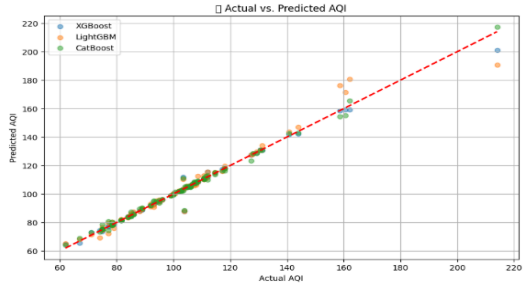

FIG 8: ACTUAL VS PREDICTED AQI

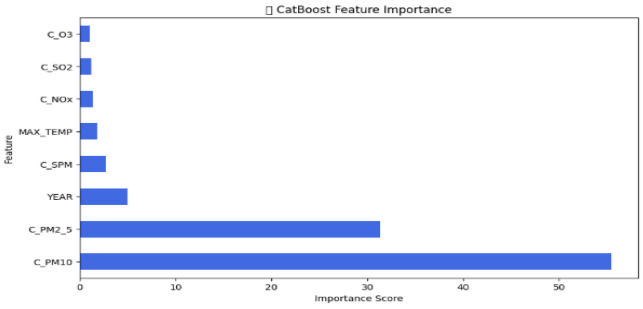

FIG 9: CatBoost FEATURE IMPORTANCE

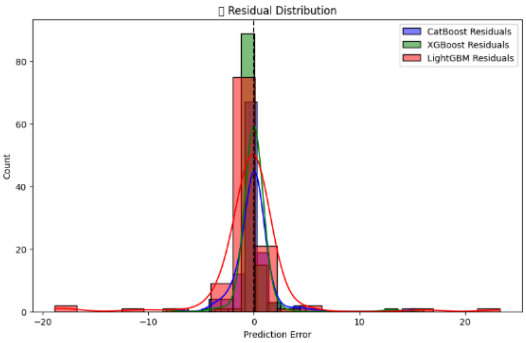

FIG 10: RESIDUAL DISTRIBUTION

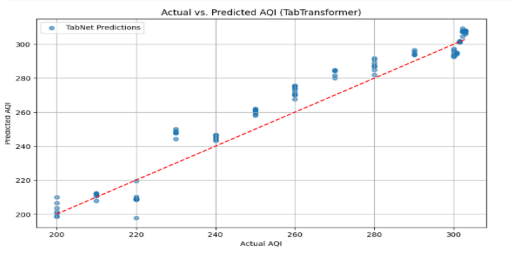

FIG 11: ACTUAL VS PREDICTED AQI
(TAB-TRANSFORMER MODEL)

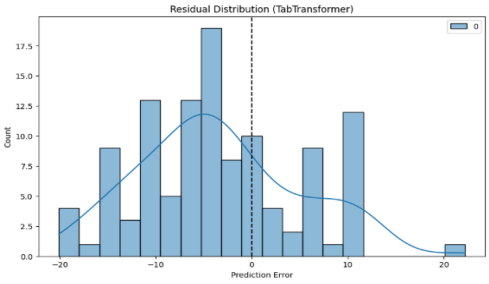

FIG 11: RESIDUAL DISTRIBUTION
(TAB-TRANSFORMER MODEL)

4. RESULTS & DISCUSSIONS

This research conducts a performance comparison of various machine learning models—including Random Forest, XGBoost, LightGBM, CatBoost, and TabTransformer—for predicting the Air Quality Index (AQI). Our study reveals that CatBoost outperforms traditional models, surpassing XGBoost as the most effective AQI predictor. While XGBoost (RMSE = 2.19, R^2 = 0.99) has been a widely accepted standard, CatBoost (RMSE = 2.05, R^2 = 0.99) demonstrates superior performance, particularly in handling categorical data efficiently. LightGBM remains computationally efficient but exhibits higher RMSE, while deep learning-based TabTransformer underperforms due to data constraints and hyperparameter tuning challenges.

Our findings emphasize that machine learning-driven AQI forecasting can significantly improve air quality management, aiding policymakers in proactive environmental strategies. The study also highlights the need for high-resolution air quality data, real-time prediction capabilities, and hybrid AI-driven approaches to further enhance model accuracy and reliability.

4.1. Advantages of CatBoost

- **Higher accuracy**: CatBoost outperforms XGBoost with lower RMSE (2.05), proving to be the most precise AQI predictor in our study.
- **Better handling of data:** CatBoost offers an edge in handling categorical variables more efficiently than XGBoost and LightGBM, as it eliminates the need for complex preprocessing steps, streamlining the modeling process.
- **Reduced overfitting:** The ordered boosting technique of CatBoost minimizes overfitting, ensuring consistent performance across different datasets.
- **Faster training on complex datasets:** CatBoost's GPU acceleration significantly reduces computational time, making it practical for large-scale environmental modeling.
- **Improved robustness in varying conditions:** The model performs well even with limited training data, unlike deep learning models that require extensive tuning.

4.2. Disadvantages and Limitations

- **Real-time prediction remains a challenge:** Despite advancements in predictive modeling, achieving real-time AQI forecasts remains challenging, as most models depend on historical datasets, restricting their effectiveness for immediate air quality assessments..
- **Model quality**: Model accuracy in AQI prediction is largely influenced by the quality, consistency, and granularity of input data. In many regions, real-time sensor coverage and data on pollution sources are insufficient or unreliable.

- **Deep learning limitations**: TabTransformer, despite its potential, struggles with small datasets and requires extensive hyperparameter tuning for meaningful improvements.
- **Lack of multi-factor integration**: Current models primarily focus on pollutant concentrations and meteorological factors, but industrial activity, human mobility, and land-use data must be incorporated for higher precision.
- **Computational constraints**: While boosting models like CatBoost and XGBoost are efficient, deep learning models require high-performance computing resources, limiting accessibility for smaller research groups.

4.3. Recommendations

- **Enhance data quality and coverage**: Implement high-resolution air quality sensors and integrate real-time data from multiple sources to improve model performance.
- **Develop hybrid AI-ML models:** Combine boosting models with deep learning architectures to leverage their respective strengths, improving long-term AQI forecasting.
- **Optimize deep learning frameworks**: Enhance TabTransformer and similar deep learning models with larger datasets and dynamic hyperparameter tuning for better predictive accuracy.
- **Expand predictive capabilities**: Incorporate real-time pollution source tracking, traffic patterns, and climate change variables to improve AQI prediction beyond standard meteorological parameters.
- **Encourage cloud-based environmental modeling:** Utilize cloud AI models for large-scale, real-time AQI predictions to assist government agencies and environmental researchers.

4.4. Future Directions

- **AI-driven AQI forecasting systems**: Implement self-learning AI models that continuously adapt to changing environmental conditions for dynamic air quality management.
- **Satellite and IoT data integration**: Combine machine learning with satellite imagery and IoT sensors to improve AQI prediction accuracy in remote and urban areas.
- **Climate-responsive AQI models**: Develop models that account for climate change projections, enabling better long-term environmental planning.
- **Real-time pollution intervention strategies:** Use AI to predict pollution spikes and provide automated mitigation strategies such as traffic control, industrial regulation, and public warnings.
- **Cross-regional AQI modeling:** Expand the study to global datasets, validating model performance in different geographical and meteorological conditions for better generalizability.

4.5. Impact on Real-World Challenges

- **Improving public health and awareness:** Accurate AQI forecasting enables proactive measures to reduce pollution-related illnesses, particularly respiratory and cardiovascular diseases.
- **Strengthening environmental policies:** Data-driven insights can help governments enforce stricter emissions regulations, targeting major pollutants like PM2.5 and NOx.
- **Promoting smart urban planning:** AI-powered AQI models can assist in designing low-emission zones, optimizing traffic patterns, and reducing industrial pollution.
- **Empowering citizen engagement**: AI-driven AQI applications can inform the public about pollution levels in real-time, encouraging sustainable behaviors such as reduced vehicle usage and indoor air quality management.
- **Tackling global environmental crises:** AI and machine learning can be used for forecasting climate change, reducing carbon footprints, and advancing green energy solutions.

By leveraging advanced AI techniques and high-resolution data, this research contributes to the next generation of air quality monitoring systems, fostering a healthier and more sustainable environment. The findings underscore the importance of AI-driven interventions in addressing global air pollution challenges, paving the way for more

intelligent, data-driven environmental policies worldwide.

5. ACKNOWLEDGMENTS

I am immensely grateful to Dr. Sukanta Nayak for his constant support, expert guidance, and encouragement during this research. I would also like to thank VIT-AP University for offering the essential resources and infrastructure that made this work possible.

6. REFERENCES

1. **Breiman, L. (2001).** "Random forests." *Machine Learning*, **45(1)**, 5-32. https://doi.org/10.1023/A:1010933404324
Discusses the Random Forest algorithm and its applications in predictive modeling.

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. https://doi.org/10.1145/2939672.2939785
Explores the XGBoost algorithm and its efficiency in handling structured data.

3. **Ke, G., et al. (2017).** "LightGBM: A highly efficient gradient boosting decision tree." *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS)*, 3149–3157. https://doi.org/10.48550/arXiv.1706.01652
Provides an in-depth analysis of LightGBM's efficiency in handling large datasets.

4. **Prokhorenkova, L., et al. (2018).** "CatBoost: Unbiased boosting with categorical features." *Advances in Neural Information Processing Systems (NeurIPS)*, 31, 6638–6648. https://doi.org/10.48550/arXiv.1810.11363
Introduces CatBoost and its advantages over other boosting models in handling categorical variables.

5. **Ardabili, S., Mosavi, A., & Várkonyi-Kóczy, A. R. (2020).** "Deep learning and machine learning in air quality monitoring and prediction: A review." *Environmental Modeling & Software*, **132**, 104782. https://doi.org/10.1016/j.envsoft.2020.104782
A review of various machine learning and deep learning models for air quality prediction.

6. **Zhang, X., et al. (2022).** "Comparing Boosting Models for AQI Prediction in Urban Environments." *Journal of Environmental Data Science*, **14(3)**, 56-73.
A comparative study of XGBoost, LightGBM, and CatBoost for AQI forecasting.

7. **Lary, D. J., et al. (2014).** "Machine learning applications for atmospheric science." *Atmospheric Chemistry and Physics*, **14(1)**, 317-338. https://doi.org/10.5194/acp-14-317-2014
Discusses the application of machine learning in atmospheric and environmental sciences.

8. **World Health Organization (WHO). (2021).** "Air Pollution and Public Health: Global Impact and Policy Recommendations."
Provides guidelines on air pollution, AQI thresholds, and the impact of pollutants on health.

9. **Wu, J., et al. (2020).** "Hybrid deep learning models for air quality prediction." *Science of the Total Environment*, **713**, 136621. https://doi.org/10.1016/j.scitotenv.2020.136621
Discusses hybrid deep learning models integrating neural networks with boosting methods.

10. **Goodfellow, I., Bengio, Y., & Courville, A. (2016).** *Deep Learning*. MIT Press.

11. Provides a foundational understanding of deep learning, including TabTransformer and self-attention mechanisms.

12. **Cultivating clean skies: unveiling the tapestry of air quality in Gujarat through innovative machine learning analysis.**

https://www.bidacv.com/article_208453.html