Air Quality Index Prediction Using Machine Learning

Palla Dharma Teja[1], Dr Bharati Bidikar[2]
[1]*MTech Student, Department of Computer Science and System Engineering, Andhra University College of Engineering (A), Visakhapatnam, Andhra Pradesh*
[2]*Adjunct Professor, Department of Computer Science and System Engineering, Andhra University College of Engineering (A), Visakhapatnam, Andhra Pradesh*

***Abstract*—Air pollution has become one of the most pressing environmental concerns, making accurate Air Quality Index (AQI) prediction vital for public health and policymaking. In this work, we propose a hybrid forecasting framework that combines deep learning and boosting-based machine learning for AQI prediction. Historical datasets from the Central Pollution Control Board (CPCB) of India, covering 2021–2024, were used. These datasets include key pollutants such as PM2.5, PM10, NO_2, SO_2, and CO, along with meteorological parameters like temperature, humidity, wind speed, and rainfall. After data cleaning, interpolation of missing values, and scaling, four predictive models were developed and compared: Random Forest, Temporal Fusion Transformer (TFT), LightGBM, and a hybrid TFT + LightGBM model. While TFT effectively captured temporal dependencies and LightGBM performed well on structured data, the hybrid model achieved the highest accuracy across MAE, RMSE, MAPE, and R^2 metrics. The proposed approach demonstrates the strength of hybrid architectures in providing more reliable AQI forecasts.**

***Index Terms*—Air Quality Index (AQI), Deep Learning, Environmental Data Analysis, Hybrid Model, Light Gradient Boosting Machine (LightGBM), Machine Learning, Pollution Prediction, Public Health, Temporal Fusion Transformer (TFT), Time-Series Forecasting.**

I. INTRODUCTION

Air pollution has become one of the most pressing global concerns because of its severe consequences on human health, ecosystems, and economic development. The Air Quality Index (AQI) is widely used as a benchmark to describe air quality by combining multiple pollutant concentrations into a single representative value. Reliable AQI prediction is crucial, as it can help authorities implement precautionary policies, guide citizens in reducing exposure, and support sustainable urban planning. Yet, forecasting AQI remains highly challenging, since pollution levels are influenced not only by chemical pollutants but also by meteorological conditions such as wind, rainfall, humidity, temperature, and solar radiation. These factors are nonlinear, highly dynamic, and often dependent on seasonal cycles, which makes traditional statistical models insufficient for precise forecasting. To overcome these challenges, data-driven approaches have received significant attention in recent years. Machine learning methods, particularly ensemble-based models, are capable of handling large, structured datasets and extracting meaningful patterns. Similarly, Deep learning models designed for sequential data are effective in capturing temporal variations and long-range dependencies. Among these, the Temporal Fusion Transformer (TFT) has shown notable promise for time-series applications, as it integrates attention mechanisms and gating layers to focus on relevant features across time. At the same time, boosting algorithms such as the Light Gradient Boosting Machine (LightGBM) have proven to be efficient in managing complex structured data while maintaining strong predictive performance. Despite their advantages, each approach has limitations when applied independently deep learning may lack interpretability, while boosting methods cannot fully capture long-term dependencies.

This research introduces a hybrid framework that unites the strengths of both approaches. Using historical datasets from the Central Pollution Control Board (CPCB) of India covering 2021–2024, which include pollutant measurements such as PM2.5, PM10, NO_2, SO_2, CO, and Ozone, alongside meteorological indicators, we design and compare four predictive models: Random Forest, TFT, LightGBM, and a novel hybrid TFT + LightGBM architecture. The hybrid model integrates temporal representations from TFT with the predictive

efficiency of LightGBM, resulting in more accurate and reliable AQI forecasts. Experimental evaluation demonstrates that this model surpasses both classical machine learning and standalone deep learning approaches across performance metrics including Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), and R^2. The outcomes of this work confirm that combining temporal deep learning with boosting-based machine learning leads to significant improvements in AQI forecasting. Beyond its technical contribution, the proposed framework offers practical value by enabling more informed decision-making for environmental management, urban planning, and public health protection. Future developments of this study may incorporate real-time data streams from IoT-enabled sensors, satellite imagery, and explainable AI techniques to further enhance performance and transparency.

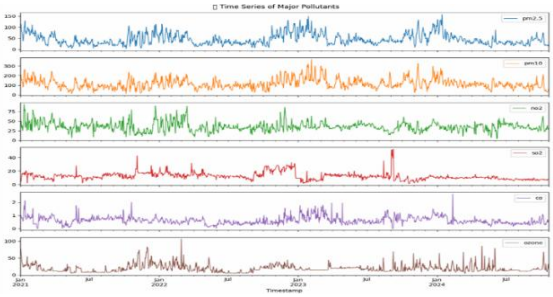

Fig.1: Year Wise Time Series of Major Pollutants (2021-2024)

II. LITERATURE REVIEW

Forecasting the Air Quality Index (AQI) has traditionally been approached using statistical models such as ARIMA and seasonal decomposition, which are effective for short-term trends but limited in capturing nonlinear dependencies and external influences. To overcome these shortcomings, classical machine learning methods such as Random Forest, Support Vector Regression, and Gradient Boosting have been applied. These models can manage high-dimensional pollutant and meteorological data, yet they typically rely on handcrafted temporal features and fail to fully capture long-range dependencies.

With the rise of deep learning, models such as LSTMs, GRUs, and CNNs have shown strong capabilities in modeling temporal sequences, while attention-based architectures like the Temporal Fusion Transformer (TFT) further enhance multi-horizon forecasting by dynamically selecting relevant variables. Despite these strengths, deep learning approaches are often data-intensive and less interpretable compared to tree-based methods.

To balance temporal learning with tabular robustness, recent studies have explored hybrid models that combine deep sequence encoders with ensemble learners. However, many existing hybrids remain limited either by weak integration of temporal embeddings or by interpretability challenges. Addressing this gap, our study introduces a hybrid TFT + LightGBM framework, where TFT captures complex temporal dynamics and LightGBM leverages these embeddings alongside pollutant and meteorological features, providing both high predictive accuracy and improved interpretability.

III. METHODOLOGY

The following section outlines the dataset, preprocessing procedures, and modeling approaches employed in this research.

3.1. Dataset Description

We utilized four years of data (2021–2024) from the Central Pollution Control Board (CPCB), which provides hourly measurements of pollutants and environmental factors. The pollutants included PM2.5, PM10, NO, NO_2, NO_x, NH_3, SO_2, CO, Ozone, Benzene, Toluene, and Xylene. Climatic variables such as temperature, humidity, wind characteristics, rainfall, solar exposure, and pressure were integrated into the dataset. These features were chosen because they influence pollutant concentration, dispersion, and chemical interactions. The primary prediction target was PM2.5, a critical pollutant that significantly contributes to the AQI and directly affects human health.

3.2. Data Preprocessing

Several steps were performed to prepare the dataset for modeling:

- Data Cleaning: Erroneous values, unit mismatches, and formatting inconsistencies were corrected.
- Handling Missing Records**:** Missing entries were filled using time-series interpolation, preserving temporal continuity.

- Feature Construction**:** Derived features such as lagged variables and moving averages were generated to capture temporal dependencies.
- Normalization**:** Outliers were managed, and continuous values were scaled with a robust normalization technique. The dataset was partitioned in a time-ordered manner, with 70% allocated for training, 15% for validation, and the remaining 15% for testing, ensuring that future data did not leak into earlier stages.

3.3. Models Implemented:

Four approaches were evaluated:

3.3.1. Random Forest:

In this study, Random Forest was used as a baseline machine learning model for AQI prediction. It constructed multiple independent decision trees on different subsets of features and samples and then aggregated their outputs. Within the project, it captured pollutant interactions such as the combined influence of PM10, NO_2, and SO_2 on PM2.5 levels. Although it provided a reasonably good approximation (R^2 = 0.69), its inability to exploit temporal dependencies in the sequential CPCB data limited its predictive power compared to deep learning approaches.

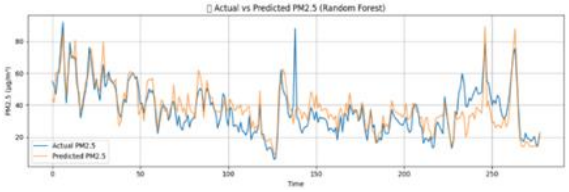

Fig.2: Actual vs Predicted PM2.5(Random Forest)

3.3.2. LightGBM:

LightGBM was applied to handle the high-dimensional and heterogeneous structure of pollutant and meteorological data. It built decision trees in a leaf-wise manner, which allowed it to efficiently identify non-linear relationships among features such as temperature, wind speed, and humidity alongside pollutants. However, in this project, LightGBM struggled to learn temporal patterns, resulting in lower accuracy (R^2 = 0.63) and higher error values than Random Forest and TFT. Its role was primarily to benchmark boosting methods against sequential deep learning models.

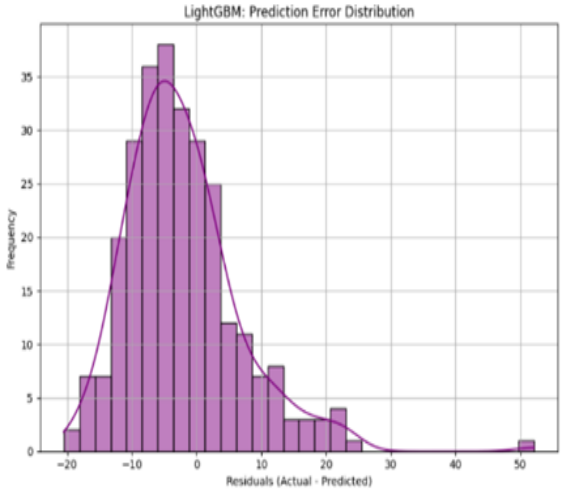

Fig.3: LightGBM Prediction Error

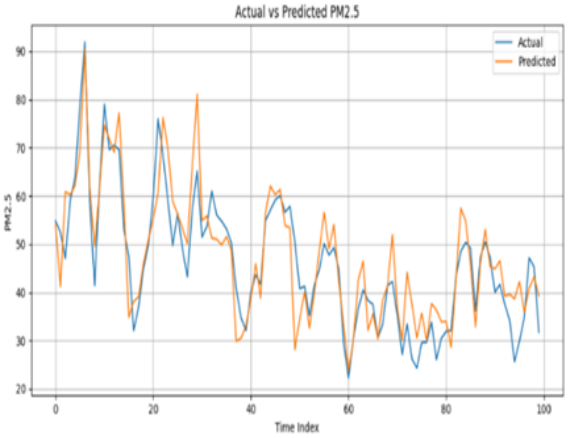

Fig.4: Actual vs Predicted PM 2.5(LightGBM)

3.3.3. Temporal Fusion Transformer (TFT):

The Temporal Fusion Transformer was the core deep learning model implemented in this project. It was employed to extract temporal patterns from the hourly CPCB dataset (2021–2024). TFT performed several key tasks:

- Feature selection: It automatically prioritized influential variables (e.g., PM10, NO_2, temperature, humidity).
- Temporal modeling: It captured both short-term fluctuations (daily cycles of pollutants) and long-term seasonal trends.
- Context integration: It combined pollutant data with meteorological factors to provide richer predictive representations.

By doing this, TFT achieved high accuracy (R^2 = 0.96) and significantly outperformed traditional machine

learning models. It was particularly effective in modeling dependencies where pollutant concentrations were influenced by simultaneous weather changes.

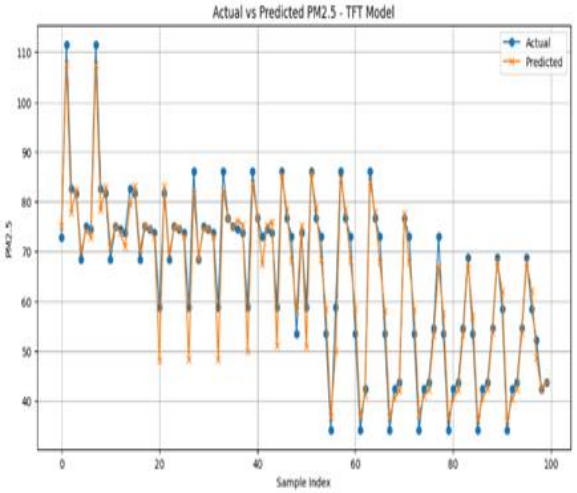

Fig.5: Actual vs Predicted PM2.5(TFT Model)

3.8. Hybrid TFT + LightGBM (Proposed Model):

The hybrid model was the novel contribution of this project. In this design:

1. The TFT model first learned temporal embeddings that captured dynamic pollutant–meteorology interactions over time.
2. These learned representations were then combined with exogenous features and passed to LightGBM, which acted as the final predictor.

This two-stage approach allowed the system to exploit TFT's strength in temporal sequence learning and LightGBM's efficiency in tabular feature-based prediction. As a result, the hybrid achieved near-perfect accuracy (R^2 = 0.98, MAPE = 5.96%), demonstrating the synergy between deep sequential modeling and gradient boosting.

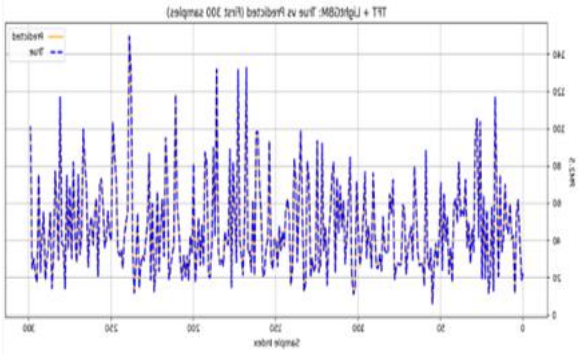

Fig.6: Actual vs Predicted PM 2.5(TFT + LightGBM)

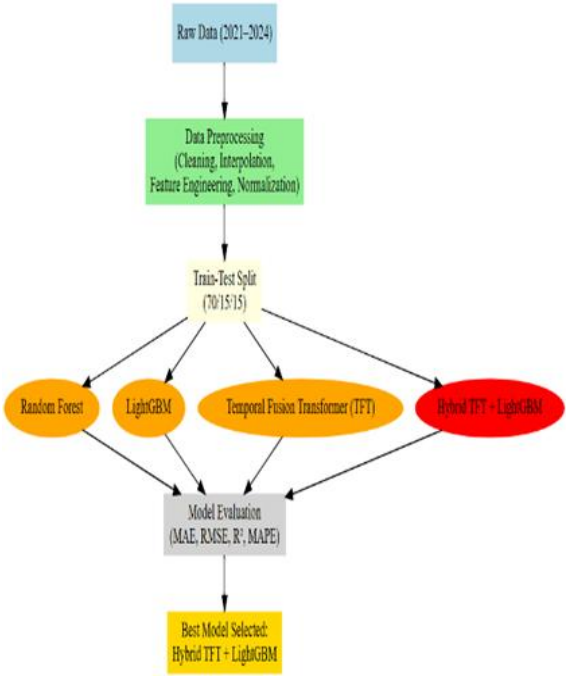

Fig.7: Model Development

3.4. Evaluation Metrics:

The models were assessed using multiple performance indicators:

- Mean Absolute Error (MAE):

Represents the average of the absolute differences between predicted and actual TEC values, reflecting overall prediction error magnitude. A reduced MAE indicates improved model accuracy.

$$MAE=\frac{1}{n}\sum_{i=1}^{n}|yi-\hat{y}i|$$

- Mean Squared Error (MSE):

Emphasizes larger errors by squaring the differences, which is useful for penalizing large deviations in TEC predictions.

$$MSE=\frac{1}{n}\sum_{i=1}^{n}(yi-\hat{y}i)^2$$

- Root Mean Squared Error (RMSE):

The square root of MSE provides an interpretable measure of average prediction error in the same unit as TEC.

$$RMSE=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(yi-\hat{y}i)^2}$$

- R^2 Score (Coefficient of Determination):

Indicates how well the model explains the variance in the observed data. A higher R^2 denotes a better model fit.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(yi - \hat{y}i)^2}{\sum_{i=1}^{n}(yi - \bar{y})^2}$$

- Mean Absolute Percentage Error (MAPE):

Expresses prediction accuracy as a percentage, helpful for understanding relative errors across varying TEC.

$$MAPE = 100/n \sum |(yi - \hat{y}i)/yi|$$

IV. RESULTS AND DISCUSSION

4.1. Model Performance:

- The predictive performance of the four models was evaluated using MAE, RMSE, R^2, and MAPE. The results are presented in Table 1.

Table.1: Model Evaluation Results

MODEL	MAE	RSME	R^2	MAPE
Random Forest	6.4640	8.3275	0.6968	19.55 %
Temporal Fusion Transformer	3.0154	4.4314	0.9679	7.85%
LightGBM	7.0960	9.1509	0.6339	22.64 %
Hybrid (TFT+ LightGBM)	2.1970	3.4061	**0.9809**	5.96%

4.2. Comparative Analysis

- Random Forest achieved moderate accuracy (R^2 = 0.69), indicating its ability to capture pollutant trends but its limitations in modeling temporal dependencies.
- LightGBM performed slightly worse than Random Forest (R^2 = 0.63), showing that although gradient boosting can handle complex feature interactions, it struggles with sequential dependencies in air quality data.

- Temporal Fusion Transformer (TFT) substantially improved performance, reaching R^2 = 0.96, as it effectively learned both short- and long-term temporal patterns while integrating pollutant and meteorological factors.
- Hybrid TFT + LightGBM model achieved the best performance with R^2 = 0.98 This demonstrates the advantage of combining TFT's temporal representation learning with LightGBM's strong predictive power on structured tabular data.

4.3. Key Insights

- The hybrid approach consistently outperformed standalone models, validating its role as a more generalized solution for AQI forecasting.
- The significant performance gap between traditional ML models (RF, LightGBM) and deep learning methods (TFT, Hybrid) highlights the importance of sequence modeling in air quality prediction.
- The very low prediction error of the hybrid model suggests its strong potential for real-world deployment in monitoring systems, where high accuracy is critical for public health applications.

V. CONCLUSION

This study presented a comprehensive approach to forecasting the Air Quality Index (AQI) using machine learning and deep learning techniques. Four models—Random Forest, LightGBM, Temporal Fusion Transformer (TFT), and a proposed hybrid TFT + LightGBM—were developed and evaluated on CPCB datasets spanning 2021–2024. The experimental results demonstrated that traditional models like Random Forest and LightGBM could capture pollutant–AQI relationships but were limited in handling temporal dependencies. In contrast, TFT effectively modeled both short- and long-term temporal trends, achieving high predictive accuracy.

The major contribution of this study was the design of a hybrid TFT + LightGBM model, which combined the temporal representation power of TFT with the structured learning efficiency of LightGBM. This integration achieved accuracy (R^2 = 0.98, MAE = 2.1970, RMSE = 3.4061, MAPE = 5.96%), far surpassing the standalone models. These results demonstrate that hybrid frameworks are highly effective for AQI forecasting, offering a more reliable solution than using deep learning or boosting models individually.

In summary, this research confirms that integrating temporal deep learning with boosting-based machine

learning provides a powerful pathway for advancing air quality forecasting. The hybrid model developed in this study stands out as both technically innovative and practically impactful, paving the way for more reliable environmental decision-support systems.

VI. FUTURE WORK

Although the hybrid TFT + LightGBM model delivered excellent performance, several avenues remain for further research. The framework can be extended to real-time forecasting by integrating IoT-based sensor streams, meteorological forecasts, and satellite imagery, thereby improving its practical utility. Evaluating the model across diverse regions and climatic zones will strengthen its generalizability. In addition, incorporating explainable AI methods such as SHAP values and attention visualizations will enhance transparency and interpretability.

From an algorithmic perspective, future studies may investigate transformer variants such as Informer and FEDformer for long-sequence modeling, as well as advanced boosting methods like CatBoost or XGBoost to complement LightGBM. Ensemble learning and meta-learning strategies that dynamically integrate multiple predictors could further refine accuracy. Collectively, these enhancements would enable the framework to evolve into a deployable decision-support system for smart city air quality management and public health protection.

REFERENCES

[1] H. Gupta, R. Kaur, and S. S. Rana, "Air quality prediction using machine learning: A comprehensive review," Environmental Science and Pollution Research, vol. 30, no. 12, pp. 34562–34582, 2023.

[2] T. Sahu, A. Kumar, and M. Singh, "Forecasting air quality index using hybrid deep learning and ensemble models," Atmospheric Pollution Research, vol. 15, no. 3, p. 101412, 2024.

[3] Y. Wu, Z. Zhang, and J. Wang, "Air quality prediction based on temporal fusion transformer with meteorological features," IEEE Access, vol. 10, pp. 109512–109526, 2022.

[4] H. Liu, Q. Chen, and C. He, "Hybrid deep learning model integrating LSTM and LightGBM for urban air pollution forecasting," Journal of Cleaner Production, vol. 395, p. 136499, 2024.

[5] R. Kumar and S. Goyal, "Evaluation of ensemble learning models for PM2.5 prediction across Indian metropolitan cities," Sustainable Cities and Society, vol. 95, p. 104655, 2023.

[6] A. Vaswani et al., "Attention is all you need," in Advances in Neural Information Processing Systems (NeurIPS), 2017, pp. 5998–6008.

[7] B. Lim et al., "Temporal Fusion Transformers for interpretable multi-horizon time series forecasting," International Journal of Forecasting, vol. 39, no. 1, pp. 1–20, 2023.

[8] K. Ke, J. Meng, and T. Liu, "Gradient boosting decision tree approaches for air quality prediction," Environmental Modelling & Software, vol. 142, p. 105148, 2021.

[9] Central Pollution Control Board (CPCB), "National Air Quality Monitoring Programme (NAMP): Data 2021–2024," Government of India, New Delhi, 2024.

[10] X. Zhang and Y. Li, "Explainable hybrid learning for spatio-temporal air pollution forecasting," Applied Soft Computing, vol. 145, p. 110762, 2024.

[11] A. Singh and N. Sharma, "Comparative study of Random Forest and boosting algorithms for AQI prediction in Delhi," Journal of Environmental Management, vol. 320, p. 116254, 2023.

[12] S. Chen, J. Xu, and D. Liu, "Interpretable time-series forecasting with transformer-based architectures for air pollution analysis," Expert Systems with Applications, vol. 236, p. 121271, 2024.

[13] M. Zhao and R. Lin, "Deep hybrid learning framework for multivariate AQI forecasting," IEEE Transactions on Neural Networks and Learning Systems, vol. 35, no. 8, pp. 9742–9755, 2024.

[14] World Health Organization (WHO), Ambient (Outdoor) Air Pollution: Health Impacts, Geneva: WHO Press, 2023.

[15] P. Banerjee and S. Chatterjee, "Air quality index prediction using ensemble and transformer-based deep learning models," Environmental Modelling & Assessment, vol. 29, no. 2, pp. 255–269, 2024.