RESEARCH ARTICLE OPEN ACCESS

Air Quality Index Prediction Using Python

Swetha. B*, Dr. R. Praba**
*(B. sc. Information Technology, Dr. N.G.P. Arts and Science College, Coimbatore, Tamil Nadu
Email: swethasubramaniyan006@gmail.com)
** (B. sc. Information Technology, Dr. N.G.P. Arts and Science College, Coimbatore, Tamil Nadu
Email : praba2426@gmail.com)

----------**********----------

Abstract:

Air pollution is a critical environmental challenge that significantly affects public health, economic productivity, and overall quality of life, particularly in rapidly urbanizing regions. The Air Quality Index (AQI) provides a standardized mechanism to translate complex pollutant concentration measurements into an easily interpretable scale that communicates health risk levels to the public. With the growing availability of environmental monitoring data, predictive modelling has become essential for forecasting future AQI levels and supporting proactive environmental management. Python has emerged as a leading platform for AQI prediction due to its extensive ecosystem for data processing, statistical analysis, machine learning, deep learning, and visualization. Ensemble learning models such as Random Forest and Extra Trees regressors demonstrate high predictive accuracy and robustness against noisy environmental data. In addition, hybrid deep learning architectures combining Convolutional Neural Networks (CNN) and Bidirectional Long Short-Term Memory (Bi-LSTM) networks effectively capture complex spatial and temporal dependencies in air quality datasets. Modern systems emphasize practical deployment using lightweight web frameworks to ensure accessibility and real-time usability. Future advancements are expected through attention mechanisms, transformer architectures, physics-informed learning, uncertainty quantification, and multi-modal data integration to enhance prediction robustness and generalization across diverse geographical regions.

Keywords — **Air Quality Index, Python, Machine Learning, Deep Learning, Random Forest, LSTM, Stream lit, Spatiotemporal Prediction.**

----------**********----------

I. INTRODUCTION

Air pollution caused by rapid urbanization, industrial growth, fossil fuel combustion, and vehicular emissions has become a major global health concern. According to the World Health Organization, prolonged exposure to polluted air increases the risk of respiratory infections, chronic obstructive pulmonary disease, lung cancer, and cardiovascular disordersThe Air Quality Index (AQI) was developed to simplify complex pollutant concentration data into a single numerical value that reflects overall air quality and associated health risks. AQI typically incorporates major pollutants such as PM2.5, PM10, NO_2, SO_2, CO, and O_3. By converting raw concentrations into standardized categories (Good, Moderate, Unhealthy, etc.), AQI enables policymakers and the public to make informed decisions. While real-time monitoring systems provide current pollution levels, forecasting future AQI values is equally important. Predictive models support early warning systems, hospital preparedness, traffic regulation, and environmental policy planning. Python-based predictive systems have become central to modern air quality analytics due to their flexibility and scalability.

A. The role of python in air quality index

Python plays a central role in air quality research because of its simplicity, flexibility, and powerful

ecosystem. Its readable syntax supports of this review collaboration among interdisciplinary teams, while libraries like Pandas, NumPy, and SciPy enable efficient handling of large environmental datasets. For modelling, frameworks such as Scikit-learn, TensorFlow, Kera's, and Py-Torch support both traditional and deep learning approaches. Visualization tools like Matplotlib, Seaborn help interpret results, while deployment frameworks such as Flask and Stream lit allow AQI models to be delivered as practical web applications.

B. Scope and organization

This review provides a comprehensive overview of research on predicting Air Quality Index (AQI) using Python, covering the full development pipeline from data collection and preprocessing to modelling, evaluation, and deployment. It compares classical machine learning and modern deep learning approaches, highlighting their strengths and practical suitability for different prediction scenarios. The paper is structured to guide readers through data preparation methods, model categories, evaluation techniques, deployment strategies, and emerging challenges, ultimately offering a clear roadmap for understanding and advancing AQI prediction systems.

II. DATA ACQUISITION AND PREPROCESSING

A. *Data Sources for Air Quality Monitoring*

Air Quality Index (AQI) prediction relies on historical pollutant data collected from government monitoring networks such as India's Central Pollution Control Board, the United States Environmental Protection Agency, and the European Environment Agency. Researchers also use aggregated public datasets available on platforms like Kaggle, which combine multi-city and multi-year measurements. These datasets include key pollutants such as PM2.5, PM10, NO_2, SO_2, CO, and O_3 along with AQI labels for model training.

The overall AQI is typically calculated based on the pollutant with the highest health risk rather than the average concentration, ensuring that the most harmful exposure is reflected. Therefore, accurate, consistent, and multi-pollutant data collection is essential for building reliable AQI prediction systems.

B. Data Cleaning and Missing Value Imputation

Raw air quality datasets frequently contain missing values, outliers, and inconsistencies caused by sensor faults, maintenance gaps, or transmission errors, making data cleaning a crucial preprocessing step. Records with missing AQI targets are usually removed because they cannot support supervised learning, while missing predictor values are filled using imputation techniques. Simple approaches like mean or median replacement are easy to implement but may distort natural variability and ignore time patterns. More advanced, time-aware strategies such as interpolation, forward or backward filling, and regression-based imputation better preserve temporal trends and cross-pollutant relationships. Proper handling of anomalies and missing data reduces noise, maintains dataset integrity, and ultimately improves the stability and predictive accuracy of AQI forecasting models.

Fig 1: AQI Data Acquisition and Preprocessing

C. Outlier Detection and Treatment

Outliers in air quality data may stem from genuine extreme events such as wildfires or dust storms, or from measurement errors like sensor malfunctions and calibration issues. Statistical methods such as the Z-score and interquartile range (IQR) approaches provide objective criteria for detecting unusual values, but they must be applied carefully since removing valid extremes can bias models toward normal conditions and reduce their ability to forecast during pollution episodes.

Treatment strategies depend on the origin of the outlier: valid extreme events should be retained, potentially with event indicators to help models distinguish routine variations from exceptional circumstances, while measurement errors should be corrected or removed. In cases where correction is not feasible, capping extreme values at plausible thresholds can reduce distortion while preserving meaningful signals, ensuring models remain both robust and environmentally relevant.

D. Feature Engineering for AQI Prediction

Feature engineering is essential for AQI prediction because it helps models capture meaningful temporal and environmental patterns beyond raw data. Features like hour of day, day of week, season, and holidays reflect human activity cycles, while lag values and rolling averages add historical context and smooth fluctuations. Interaction features, such as temperature with sunlight or wind with pollutant dispersion, highlight combined effects that drive air quality changes.

Feature selection then refines the dataset by removing redundant inputs and focusing on the most predictive ones. Methods like correlation checks, mutual information, and recursive elimination streamline the feature set, while domain expertise ensures chosen features have real-world relevance. Together, these steps make AQI models more efficient, accurate, and reliable in capturing pollution dynamics.

III. TRADITIONAL MACHINE LEARNING APPROACHES

A. Linear Regression and Regularized Variants

Linear regression is a simple baseline model for AQI prediction that assumes a linear relationship between predictors and the target variable. It expresses AQI as a weighted sum of input features, with coefficients estimated by minimizing squared residuals. While straightforward, it provides interpretable results, showing how each pollutant contributes to AQI, which can inform both public understanding and policy decisions.

Regularized variants improve performance when predictors are highly correlated or feature spaces are large. Ridge regression shrinks coefficients to reduce variance, Lasso regression performs feature selection by driving some coefficients to zero, and Elastic Net combines both approaches to balance shrinkage and selection. These methods help control complexity, prevent overfitting, and enhance predictive accuracy, making them especially useful in AQI modelling with many potential features.

B. Tree-Based Ensemble Methods

Tree-based ensemble methods are highly effective for AQI prediction because they capture complex, non-linear relationships between pollutants and meteorological variables. While single decision trees are interpretable, they often overfit, so ensembles like Random Forest average predictions from multiple trees to reduce variance and improve robustness. Random Forests handle non-linearities automatically, resist outliers, require little tuning, and provide feature importance scores that highlight key drivers of air quality, making them both powerful and practical.

Gradient Boosting methods such as XG-Boost, Light, and Cat Boost build trees sequentially to correct errors, often achieving higher accuracy but requiring careful tuning to avoid overfitting. Extra-Trees adds further randomization in split selection, which can be effective in noisy datasets. Studies consistently show that these ensemble methods outperform simpler models, with pollutants like PM2.5, CO, and NO_2 emerging as dominant predictors, underscoring their strength in environmental prediction tasks.

C. Support Vector Regression

Support Vector Regression (SVR) adapts Support Vector Machine principles for regression by mapping input features into a higher-dimensional space using kernel functions. It fits a hyperplane within an epsilon-insensitive margin, balancing model complexity with prediction accuracy. This allows SVR to capture both linear and non-linear relationships, making it useful for AQI prediction where pollutant interactions can be complex.

The choice of kernel function is key: linear kernels suit simple relationships, polynomial kernels capture interactions of varying degrees, and radial basis function (RBF) kernels handle highly non-linear patterns. While flexible, SVR requires careful tuning of kernel parameters, regularization, and epsilon values, and can be computationally intensive on

large datasets. When optimized, it delivers strong predictive performance and robustness in modelling air quality.

IV. DEEP LEARNING APPROACHES

A. Neural Network Fundamentals for Time Series

Deep learning has become a powerful tool for AQI prediction because it can automatically learn complex patterns from raw time series data. Neural networks use layers of interconnected neurons with non-linear activations, optimized through backpropagation, to minimize prediction error. Unlike traditional models, they can capture subtle interactions and long-range dependencies without manual feature engineering, making them well-suited for environmental data where pollutant dynamics are multi-scale and complex.

These models also integrate diverse data sources such as ground measurements, satellite observations, meteorological fields, and even text reports into unified frameworks. Recurrent and attention-based architectures allow them to leverage information from distant past observations, while their flexibility supports multimodal inputs. This makes deep learning especially effective for AQI forecasting in contexts where relationships are not fully understood or data is heterogeneous, offering both accuracy and adaptability.

B. Recurrent Neural Networks and LSTM

Deep learning has become a powerful tool for AQI prediction because it can automatically learn complex patterns from raw time series data. Neural networks use layers of interconnected neurons with non-linear activations, optimized through backpropagation, to minimize prediction error. Unlike traditional models, they can capture subtle interactions and long-range dependencies without manual feature engineering, making them well-suited for environmental data where pollutant dynamics are multi-scale and complex.

These models also integrate diverse data sources such as ground measurements, satellite observations, meteorological fields, and even text reports into unified frameworks. Recurrent and attention-based architectures allow them to leverage information from distant past observations, while their flexibility supports multimodal inputs. This makes deep learning especially effective for AQI forecasting in contexts where relationships are not fully understood or data is heterogeneous, offering both accuracy and adaptability.

C. Hybrid CNN-Bi-LSTM Architectures

Hybrid CNN-Bil-STM architectures combine convolutional layers with bidirectional LSTMs to improve AQI prediction. CNNs are effective at detecting local temporal patterns such as sharp increases during rush hours or gradual pollution buildup, while also capturing broader cycles like weekly or seasonal trends. They act as feature extractors, reducing dimensionality and providing compact representations of relevant patterns regardless of timing.

Bil-STM layers then process these features in both forward and backward directions, capturing dependencies from past and future states. This bidirectional context is valuable for modelling pollution peaks, which often show symmetric rise and fall patterns. Together, CNNs and Bil-STMs capture both local variations and long-range dependencies, producing accurate AQI forecasts without extensive manual feature engineering and effectively learning complex spatiotemporal dynamics directly from data.

V. MODEL EVALUATION AND DEPLOYMENT

A. Evaluation Metrics for Regression Performance

AQI prediction models are evaluated using regression metrics that measure the difference between predicted and actual values. Mean Absolute Error (MAE) gives the average absolute difference in AQI units, making it simple and robust to outliers. Mean Squared Error (MSE) penalizes larger errors more heavily, while Root Mean Squared Error (RMSE) brings results back to AQI units, combining interpretability with sensitivity to large deviations.

R-squared (R^2) measures the proportion of variance explained by the model, offering a scale-independent performance indicator. While useful for comparing models, R^2 must be interpreted carefully since it can be inflated by irrelevant predictors or overfitting. Together, these metrics provide a balanced view of accuracy, error sensitivity, and explanatory power in AQI prediction.

B. Cross-Validation Strategies for Time Series

Time series cross-validation differs from standard k-fold because temporal order must be preserved to avoid data leakage. Random shuffling used in traditional methods breaks this order, giving models unrealistic access to future information. Rolling window cross-validation solves this by training on expanding past data and validating on the next period, closely simulating real-world forecasting but at higher computational cost.

Other strategies include time series split, which creates multiple train-test splits while maintaining order, and purged walk-forward validation, which adds a gap between training and validation sets to reduce leakage from short-term autocorrelation. These approaches provide more realistic performance estimates, ensuring AQI prediction models are evaluated under conditions that reflect true forecasting challenges.

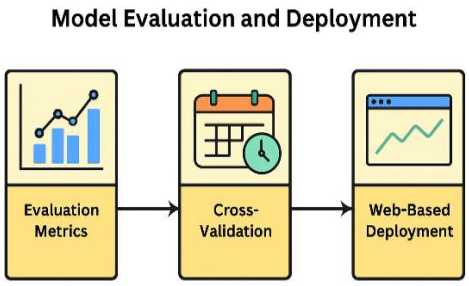

Fig 1: Model Evaluation and Deployment

D. Web-Based Deployment with Stream lit

Stream lit is widely used for deploying AQI prediction models because it allows interactive web applications to be built directly in Python without requiring web development expertise. A typical app lets users input pollutant concentrations or select a location and time, then loads a pre-trained model, processes inputs, and generates predictions. Interactive elements like sliders, dropdowns, and date pickers make it easy to explore scenarios and understand how changes in pollutant levels affect air quality.

Its visualization capabilities enhance usability by showing time series plots of historical and predicted AQI, maps of spatial patterns, comparisons across models, and feature importance charts. These visualizations help users interpret and trust predictions, supporting informed decisions for policymakers, healthcare providers, and the public. Streamlet's simplicity and minimal coding needs make it an attractive choice for researchers and agencies to share predictive models effectively.

VI. CHALLENGES AND FUTURE DIRECTIONS

A. *Current Limitations in AQI Prediction*

Current AQI prediction models face challenges with data quality and availability, especially in regions with sparse monitoring networks or incomplete records. Limited or low-quality data can reduce accuracy and prevent models from generalizing to new conditions. Spatial heterogeneity also complicates predictions, since air quality varies significantly across short distances due to local sources, topography, and weather, making station-based models less reliable for unmonitored areas.

Extreme events like wildfires, dust storms, or industrial accidents are difficult to predict because models are usually trained on normal conditions and struggle during rare, high-impact episodes. Additionally, temporal non-stationarity changes in emissions, meteorology, or atmospheric chemistry over time due to policies, development, or climate change—means models that assume stable relationships can become outdated, requiring continuous retraining and adaptation to maintain accuracy.

B. Emerging Research Directions

Emerging research in AQI prediction focuses on advanced deep learning techniques and richer data integration. Attention mechanisms and transformer architectures are being explored to capture complex dependencies across long time horizons, offering advantages for very long-range forecasting. Multi-modal integration is another key direction, combining ground measurements with satellite data, meteorological reanalysis, and even social media inputs to improve accuracy and robustness, especially in regions with limited monitoring infrastructure.

Other promising approaches include physics-informed neural networks that embed atmospheric chemistry and transport equations into models, ensuring physically consistent predictions. Transfer learning allows knowledge from data-rich regions to be adapted to data-sparse areas, while uncertainty quantification methods provide prediction intervals instead of point estimates, helping users assess confidence in forecasts. Together, these directions aim to make AQI prediction more accurate, generalizable, and reliable for real-world decision-making.

VII. CONCLUSION

This paper reviewed AQI prediction methods using Python, ranging from traditional machine learning to advanced deep learning. Ensemble tree-based models like Random Forest and Extra Trees consistently deliver strong accuracy, balancing interpretability and robustness, while hybrid deep learning architectures such as CNN-Bil-STM achieve the highest precision by learning both spatial and temporal patterns. Feature importance analyses highlight PM2.5, CO, and NO_2 as dominant pollutants, guiding monitoring and control strategies.

Deployment frameworks like Stream-lit make these models accessible through interactive web applications, supporting policymakers, healthcare providers, and the public with clear forecasts and visualizations. As air pollution remains a global health challenge, future research directions—including attention mechanisms, multi-modal integration, physics-informed learning, and uncertainty quantification promise to further enhance prediction accuracy, reliability, and practical impact.

ACKNOWLEDGEMENT

The authors express their sincere gratitude to all researchers and organizations whose publicly available air quality datasets, tools, and frameworks made this study possible. We acknowledge the valuable contributions of environmental monitoring agencies and open data platforms that support research in air quality prediction. Special appreciation is extended to academic mentors and peers for their guidance, feedback, and encouragement throughout the development of this work. Their support significantly contributed to the successful completion of this research.

REFERENCES

[1] World Health Organization, *Air Pollution and Health*, WHO Report, 2023.
[2] Environmental monitoring datasets from Central Pollution Control Board, Government of India.
[3] Air quality datasets from United States Environmental Protection Agency, EPA Open Data Portal.
[4] European air monitoring data from European Environment Agency.
[5] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
[6] M. Abadi et al., "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015.
[7] A. Geron, *Hands-On Machine Learning with Scikit-Learn, Kera's, and TensorFlow*, O'Reilly, 2019.
[8] Datasets sourced from Kaggle public repositories.
[9] Interactive deployment framework documentation from Stream lit.
[10] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, MIT Press, 2016.

