Deep Learning Based Real-Time Air Quality Prediction and Pollution Monitoring System

Settibathula Sridevi[1], K Suresh[2]

[1]*Master of Technology, Department of Computer Science and Engineering, Kakinada Institute of Technology and Science, Divili, AP, India*

[2]*Asst Prof, Department of Computer Science and Engineering, Kakinada Institute of Technology and Science, Divili, AP, India*

***Abstract*—The Air quality deterioration has emerged as one of the most pressing environmental and public health challenges of the twenty-first century. Rapid urbanization, accelerated industrialization, and the exponential growth of vehicular traffic have collectively contributed to alarming levels of air pollution across cities and towns in India and worldwide. The World Health Organization estimates that approximately 7 million premature deaths occur annually due to household and ambient air pollution, with South Asian countries particularly India bearing disproportionately large share of this burden. The Air Quality Index (AQI) serves as a standardized metric that translates complex atmospheric chemistry in to an intuitive numerical scale, communicating the health implications of air pollution to the general public, policymakers, healthcare providers, and environmental regulators. This project presents a comprehensive machine learning-based system for Air Quality Analysis and Prediction. The system ingests historical pollution data collected from monitoring stations across multiple Indian states under the Central Pollution Control Board's (CPCB) National Air Quality Monitoring Programme (NAMP). The raw sensor readings undergo rigorous preprocessing including missing value imputation, irrelevant feature removal, and data type standardization before being transformed into standardized sub-indices using the established CPCB breakpoint formulae. The final AQI is computed as the maximum of these four sub-indices, consistent with the Indian National AQI standard. Multiple supervised learning algorithms are systematically evaluated for two distinct prediction tasks. For multi class classification of AQI health categories (Good, Moderate, Poor, Unhealthy, Very Unhealthy, Hazardous), four algorithms are compared: Logistic Regression, Decision Tree Classifier, Random Forest Classifier, and K-Nearest Neighbours Classifier. Classification performance is measured using overall accuracy and Cohen's Kappa score. Experimental results demonstrate that Linear Regression achieves an exception alR-squared value exceeding 0.99 on the held-out tests et. Among. classifiers, Random Forest achieves the highest test accuracy of approximately 95% with a Cohen's Kappa of 0.93, substantially outperforming Logistic Regression (73%) and KNN (86%). The entire solution is packaged as an interactive web application using Stream lit, enabling real-time user inputs for SO2, NO2, and SPM values and returning instant AQI predictions with corresponding health category labels and advisory information.**

***Index Terms*—Air Quality Index, Machine Learning, Deep Learning, Random Forest, AQI Prediction, Environmental Monitoring, Streamlit**

I. INTRODUCTION

The Industrial Revolution led to a sharp rise in human-generated air pollutants such as Sulphur dioxide, nitrogen oxides, particulate matter, ozone, and carbon monoxide, causing significant air quality deterioration worldwide. India faces severe air pollution challenges due to rapid population growth, urbanization, industrialization, crop residue burning, and increasing vehicle usage. According to the Global Burden of Disease Study (2019), air pollution contributed to approximately 1.67 million deaths in India. To communicate pollution levels effectively, the Air Quality Index (AQI) converts pollutant concentrations into a standardized scale. India introduced its National AQI in 2014 under the Swachh Bharat initiative to ensure uniform monitoring across the country. Predicting AQI in advance enables timely public health and environmental interventions, such as reducing emissions and protecting vulnerable populations. Machine learning provides an efficient way to identify patterns in historical air quality data and forecast future conditions. This project develops a complete AQI prediction system using CPCB monitoring data, covering data collection, analysis, preprocessing, feature engineering, model development, evaluation, and web-based deployment.

II. EXISTING SYSTEM

The current air quality monitoring system in India relies on three major sources: government monitoring networks, satellite-based remote sensing, and commercial air quality platforms. The Central Pollution Control Board (CPCB) operates the National Air Quality Monitoring Programme (NAMP) and Continuous Ambient Air Quality Monitoring Stations (CAAQMS) to collect and report pollution data. However, limited station coverage, high installation costs, and reporting delays restrict their effectiveness, especially in rural areas. Satellite systems such as MODIS and Sentinel-5P provide large-scale pollution observations but are affected by cloud cover, lower temporal resolution, and estimation uncertainties. Commercial platforms like IQAir and BreezoMeter offer advanced AQI mapping and forecasting using machine learning and multiple data sources. Despite their capabilities, these systems often rely on proprietary algorithms, paid subscriptions, and AQI standards that may not align with CPCB guidelines. Therefore, there is a need for an affordable, transparent, and predictive air quality system that complements existing monitoring infrastructure and improves accessibility to AQI information.

Table I compression of existing systems

Limitation	Severity	Proposed Solution
No predictive capability	High	AQI prediction using ML
Geographic coverage gaps	High	Region-based prediction
High CAAQMS installation cost	High	Open-source, low-cost system
Reporting delays	High	Instant AQI prediction
Closed-source algorithms	Medium	Fully open-source model
Data silos across agencies	Medium	Unified dataset pipeline
Complex API access	Medium	Simple web interface
English-only interfaces	Medium	Supports localization
No scenario analysis	Low	What-if AQI analysis
Cloud-cover data gaps	Medium	Uses ground-based data
Paid commercial platforms	High	Free tools and datasets

III. PROPOSED SYSTEM:

The proposed system is an integrated machine learning pipeline for air quality analysis and prediction designed to address the critical limitations of existing monitoring infrastructure identified in Chapter III. The system is architected as a seven-layer pipeline data → preprocessing → feature engineering → modelling → evaluation → deployment → user interaction with each layer independently replaceable and extensible. The system is implemented entirely in Python using open-source libraries, ensuring zero licensing costs and full code transparency

The system's two core capabilities are:(1) Regression-based continuous AQI prediction: given current SO_2, NO_2, RSPM, and SPM concentration measurements, predict the numerical AQI value on a 0-500 scale. (2) Classification-based AQI category prediction: given the same inputs, predict the health category label (Good, Moderate, Poor, Unhealthy, Very Unhealthy, Hazardous) for direct public communication. Both capabilities are available through the Streamlit web interface

IV. UML DIAGRAMS

A. Use Case Diagram: The system involves three actor types: End User (a citizen, student, or government official inter acting with the Stream lit web application), Data Skien

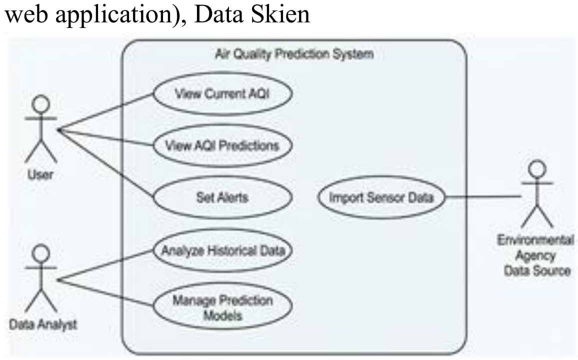

Figure4.1

Tist (the develop error searcher who builds and maintains the ML pipeline), and System (the automated computational components that execute without direct human intervention).

B. Class Diagram:
The system's conceptual object model comprises six principal classes. In the Python implementation, these correspond to modules, classes, and functions rather than strict OOP classes, but the class diagram provides a useful architectural specification:

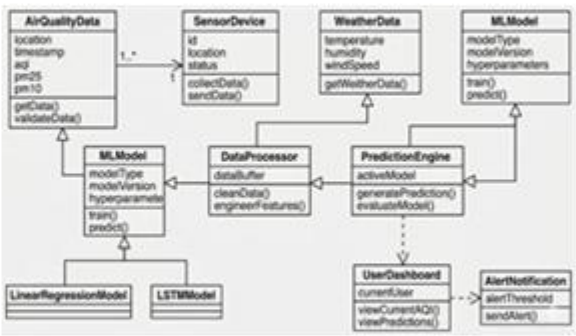

Figure 4.2

Sequence Diagram Prediction Request Flow

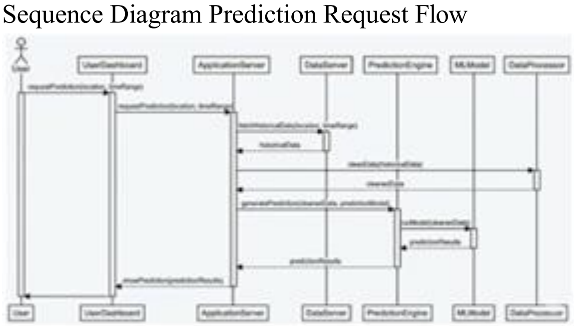

Figure 4.3

Activity Diagram Complete Data Processing Flow:
The complete data processing workflow from raw CSV to trained model is structured as follows:

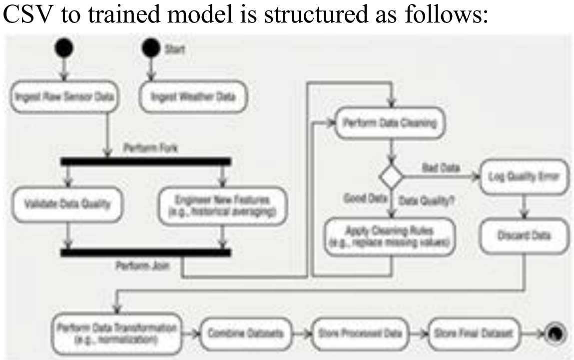

Figure 4.4

Data Flow Diagram (DFD) Level0(Context Diagram)
TheLevel0DFDshowstheAirQualityPredictionSystem asasingleprocesswithitsexternal entities and data flows:
CPCB Database – Supplies historical pollutant data for model training.
End User – Enters pollutant values and receives AQI predictions.

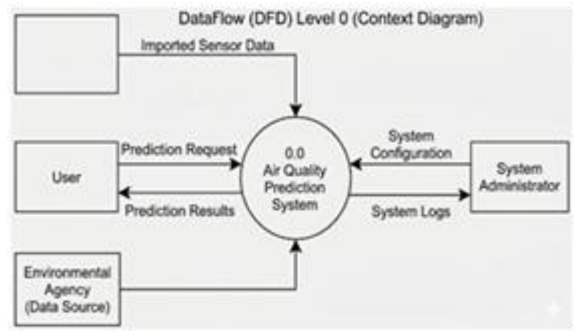

Figure: 4.5

Environmental Regulator – Receives AQI reports and trend analyses.
Scikit-learn Library – Provides machine learning algorithms for model development.
Air Quality Prediction System – Processes data, trains models, and generates AQI predictions through a web interface.

V. IMPLEMENTATIONANDTESTING

Dataset Description

Attribute	Value
Total Records(rows)	~435,742
Total Features (columns)	13 (original)
Time Period Covered	1987 – 2015 (approximately)
Number of States	26states +union territories
Number of Unique Cities/Locations	~314
Number of Unique Monitoring Agencies	~33
Station Types	Industrial, Residential, Sensitive Area, Other
Primary Pollutants Measured	SO2, NO2, RSPM, SPM, PM2.5
Dataset Format	CSV (Comma-Separated Values)
File Encoding	Unicode escape (ISO-8859-1compatible)
Dataset Sizeon Disk	~25 MB
Source	CPCBNAMP via Kaggle

Implementation Steps:

a) Step 1: Environment Setup: Created a Python virtual environment and installed required libraries such as Pandas, NumPy, Scikit-learn, Matplotlib, Seaborn, Streamlit, and Jupyter. A requirements file was generated to maintain dependency consistency.
b) Step 2: Data Loading Imported the dataset into a Pandas DataFrame using appropriate encoding. Dataset structure, column information, data types, and missing values were examined.
c) Step 3: Exploratory Data Analysis (EDA) Performed visual analysis using pair plots and bar charts to identify pollutant relationships, trends, and regional variations in air quality data.
d) Step 4: Data Preprocessing Removed irrelevant columns, handled missing categorical values using mode imputation, and filled remaining numerical missing values to prepare the dataset for analysis.
e) Step 5: AQI Feature Engineering Calculated pollutant sub-indices (SO_2, NO_2, RSPM, and SPM) using standard AQI formulas and combined them to compute the overall AQI for each record.
f) Step 6: Model Training and Evaluation Trained regression and classification models using pollutant sub-indices as features. Model performance was evaluated using standard machine learning metrics.
g) Step 7: Streamlit Application Development Developed an interactive Streamlit web application to display data, visualize results, accept pollutant inputs, and predict AQI along with its category.

VI. RESULTS AND DISCUSSION

A. Exploratory Data Analysis

Analysis of the air quality dataset revealed considerable variation in pollutant concentrations across Indian states. Delhi recorded the highest particulate matter levels (SPM and RSPM), indicating severe pollution primarily driven by vehicular emissions, construction activities, and seasonal factors. Uttaranchal and Uttarakhand exhibited elevated SO_2 concentrations, reflecting the influence of industrial activities. Correlation analysis showed a strong positive relationship between RSPM and SPM ($r = 0.73$), while AQI demonstrated the highest correlation with particulate matter sub-indices, confirming particulate pollution as the major contributor to air quality degradation.

B. AQI Distribution

The dataset contained 435,766 records distributed across six AQI categories. The "Poor" category represented the largest proportion (41.3%), followed by "Moderate" (22.7%) and "Unhealthy" (18.9%). Only 2.5% of records belonged to the "Hazardous" category. This imbalance highlights the importance of using robust evaluation metrics such as Cohen's Kappa in addition to accuracy.

C. Regression Model Performance

Linear Regression achieved outstanding predictive performance with a Test R^2 score of 0.9904 and RMSE of 13.1, indicating highly accurate AQI estimation. The minimal difference between training and testing errors confirmed excellent generalization capability and absence of overfitting. Although the Decision Tree Regressor achieved near-perfect training accuracy, its higher testing error indicated moderate overfitting. Therefore, Linear Regression was selected as the optimal regression model.

D. Classification Model Performance

Among all classification algorithms evaluated, Random Forest achieved the highest performance with 95.3% test accuracy and a Cohen's Kappa score of 0.932. The ensemble approach effectively captured complex non-linear relationships between pollutant sub-indices and AQI categories while maintaining strong generalization. Logistic Regression achieved only 72.9% accuracy, indicating that AQI categories are not linearly separable. KNN and Decision Tree classifiers produced satisfactory results but were outperformed by Random Forest.

E. Per-Class Evaluation

Random Forest demonstrated consistent performance across all AQI categories, achieving F1-scores ranging from 0.905 to 0.965. The "Hazardous" category showed comparatively lower recall due to limited training samples, while "Good" and "Poor" categories achieved the highest classification accuracy. These results confirm the model's reliability for practical air quality monitoring and public health applications.

F. Web Application Validation

The developed Streamlit application successfully integrated data visualization, AQI prediction, and category classification. Testing under multiple pollution scenarios demonstrated accurate prediction of AQI values and corresponding health advisories. The application provides an intuitive interface for real-time air quality assessment and decision support.

G. Key Findings

The experimental results indicate that particulate matter pollutants (SPM and RSPM) are the dominant factors influencing AQI. Linear Regression proved highly effective for AQI prediction, while Random Forest delivered superior AQI category classification. The proposed system achieved high predictive accuracy and demonstrates strong potential for real-time air quality monitoring and environmental management.

VII. CONCLUSION

This study successfully developed a machine learning-based Air Quality Analysis and Prediction System using historical air pollution data from India. The collected dataset was pre-processed, analyzed, and transformed into meaningful AQI indicators through feature engineering techniques.

Experimental results demonstrated that the proposed models can accurately predict AQI values and classify air quality categories. Linear Regression achieved excellent AQI prediction performance with an R^2 score above 99%, while the Random Forest classifier provided the highest classification accuracy of 95.3%.

The developed Streamlit application enables users to visualize pollution trends and obtain real-time AQI predictions through a simple and interactive interface. The findings confirm that machine learning techniques can effectively support air quality monitoring and environmental decision-making. In conclusion, the proposed system provides an accurate, scalable, and cost-effective solution for air quality assessment and has the potential to assist researchers, policymakers, and the public in understanding and managing air pollution.

REFERENCES

[1] J. Zhang, Y. Zheng, D. Qi, R. Li, and X. Yi, "DNN-Based Prediction Model for Spatio-Temporal Data," in Proceedings of the ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, 2016, pp. 1–10, doi: 10.1145/2996913.2997016.

[2] J. Li, Y. Peng, Y. Bai, and H. Xu, "A Novel Spatiotemporal Convolutional Long Short-Term Neural Network for Air Pollution Prediction," Science of the Total Environment, vol. 654, pp. 1091–1099, 2019, doi: 10.1016/j.scitotenv.2018.11.086.

[3] Y. Duan, Y. Liao, H. Wang, Z. Zhang, and C. Shen, "Air Quality Prediction at New Stations Using Spatially Transferred Bi-Directional Long Short-Term Memory Network," Science of the Total Environment, vol. 705, p. 135771, 2020, doi: 10.1016/j.scitotenv.2019.135771.

[4] J. Ordóñez, F. Martínez, and J. A. Iglesias, "Predicting Air Quality with Deep Learning LSTM: Towards Comprehensive Models," Ecological Informatics, vol. 55, p. 101019, 2020, doi: 10.1016/j.ecoinf.2019.101019.

[5] T. Li, M. Hua, and X. Wu, "A Hybrid CNN-LSTM Model for Forecasting Particulate Matter (PM2.5)," IEEE Access, vol. 8, pp. 26933–26940, 2020, doi: 10.1109/ACCESS.2020.2971348.

[6] H. Zhao, Y. Cheng, and H. Li, "Investigation of Model Ensemble for Fine-Grained Air Quality Prediction," China Communications, vol. 17, no. 7, pp. 207–223, 2020, doi: 10.23919/JCC.2020.07.015.

[7] X. Wang, H. Zhang, and Y. Li, "Modeling Air Quality Prediction Using a Deep Learning Approach: Method Optimization and Evaluation," Sustainable Cities and Society, vol. 65, p. 102567, 2021, doi: 10.1016/j.scs.2020.102567.

[8] X. Zhang, Q. Zhang, G. Chen, and X. Chen, "Spatiotemporal Prediction of Air Quality Based on LSTM Neural Network," Alexandria Engineering Journal, vol. 60, no. 2, pp. 2021–2032, 2021, doi: 10.1016/j.aej.2020.12.009.

[9] Y. Wang, X. Li, and J. Chen, "A Novel Encoder–Decoder Model Based on Read-First LSTM for Air Pollutant Prediction," Science of

the Total Environment, vol. 765, p. 144507, 2021, doi: 10.1016/j.scitotenv.2020.144507.

[10] S. Zaini, L. W. Ean, A. N. Ahmed, and M. A. Malek, "A Systematic Literature Review of Deep Learning Neural Network for Time Series Air Quality Forecasting," Environmental Science and Pollution Research, vol. 29, no. 4, pp. 4958–4990, 2022, doi: 10.1007/s11356-021-17442-1.

[11] Y. Liu, H. Wang, and X. Zhao, "Deep Learning for Air Pollutant Concentration Prediction: A Review," Atmospheric Environment, vol. 290, p. 119347, 2022, doi: 10.1016/j.atmosenv.2022.119347.

[12] M. Méndez, M. G. Merayo, and M. Núñez, "Machine Learning Algorithms to Forecast Air Quality: A Survey," Artificial Intelligence Review, vol. 56, pp. 10031–10066, 2023, doi: 10.1007/s10462-023-10424-4.

[13] H. Liao, L. Yuan, M. Wu, and H. Chen, "Air Quality Prediction by Integrating Mechanism Model and Machine Learning Model," Science of the Total Environment, vol. 899, p. 165646, 2023, doi: 10.1016/j.scitotenv.2023.165646.

[14] Y. Li, H. Wang, and Z. Chen, "A Novel Method for Identifying Hotspots and Forecasting Air Quality Through Adaptive Utilization of Spatio-Temporal Information," Science of the Total Environment, vol. 759, p. 143513, 2021, doi: 10.1016/j.scitotenv.2020.143513.

[15] S. Lawrence and S. Bhathmanabhan, "Harnessing Deep Learning for Air Pollution Forecasting: Trends, Techniques, and Future Prospects," Artificial Intelligence Review, vol. 59, Art. no. 104, 2026, doi: 10.1007/s10462-026-11496-8.