

AI-Driven Air Quality Index Prediction and Real-Time Health Advisory System Using Machine Learning

Tanisha Gautam¹, Prachi Sharma¹, Danish Kaul¹, Sherry Verma²

¹ School of Engineering and Technology, Sushant University, Gurgaon, Haryana, India.

² Associate Professor, School of Engineering and Technology, Sushant University, Gurgaon, Haryana, India, ORCID: 0009-0003-0616-8162.

Abstract

Air pollution represents one of the most pressing environmental and public health challenges globally, with the World Health Organization estimating approximately 7 million premature deaths annually attributable to ambient and household air pollution. This paper presents an AI-driven system for predicting the Air Quality Index (AQI) from pollutant concentration measurements and delivering personalized health advisories through a full-stack web application. Five machine learning regression models, Linear Regression, Decision Tree, Random Forest, Multi-Layer Perceptron (MLP), and XGBoost, are trained and comparatively evaluated on a 10,000-sample dataset generated using India's Central Pollution Control Board (CPCB) breakpoint formulae across eight major Indian cities. The XGBoost model achieved the highest performance with $R^2 = 0.9938$, $MAE = 2.40$, and $MAPE = 1.54\%$ on the held-out test set. The system integrates real-time pollutant data from multiple external APIs (WAQI, OpenWeatherMap) through a multi-tier fallback architecture, and delivers actionable health advisories categorized across five AQI severity levels. The complete system is deployed as a FastAPI backend with a React-based frontend, demonstrating that machine learning can effectively bridge the gap between raw environmental sensor data and actionable public health guidance.

Keywords: Air Quality Index, Machine Learning, XGBoost, Random Forest, Health Advisory System, Air Pollution Prediction, CPCB, Real-Time Monitoring, FastAPI, Environmental Computing

1. Introduction

Air pollution has emerged as a defining environmental crisis of the 21st century, with profound implications for public health, economic productivity, and quality of life. The World Health Organization reports that 99% of the global population breathes air that exceeds WHO guideline limits, contributing to approximately 7 million premature deaths annually from diseases including stroke, heart disease, chronic obstructive pulmonary disease, lung cancer, and acute respiratory infections [1]. India faces a disproportionate burden: 14 of the world's 15 most polluted cities are located in India, and ambient air pollution contributes to an estimated 1.67 million deaths annually in the country [2].

The Air Quality Index (AQI) serves as the primary standardized metric for communicating air quality conditions to the public. India's Central Pollution Control Board (CPCB) defines AQI through a piecewise linear function that maps concentrations of six criteria pollutants, PM_{2.5}, PM₁₀, NO₂, SO₂, CO, and O₃, to a unified index scale ranging from 0 (Good) to 500+ (Hazardous) [3]. While monitoring stations measure these pollutants, significant gaps exist in spatial coverage, temporal resolution, and the translation of AQI values into actionable health guidance for the general population.

Machine learning offers a promising approach to address these limitations. By learning complex, nonlinear relationships between pollutant concentrations and AQI values, ML models can provide rapid predictions that supplement and extend the reach of physical monitoring infrastructure [4, 5]. However, most existing studies focus exclusively on model development in isolation, without addressing the critical requirements of real-time data integration, health advisory generation, and end-to-end deployment as accessible applications [6].

This paper presents an end-to-end AI-driven system that addresses the complete pipeline from raw pollutant data to actionable health advisories. The contributions of this paper are: (a) a comparative evaluation of five regression models for AQI prediction on a CPCB-formula-validated synthetic dataset; (b) a novel multi-tier real-time data acquisition architecture with API fallback and unit conversion mechanisms; (c) a rule-based health advisory engine

that translates predicted AQI values into personalized, category-specific health recommendations; and (d) a complete full-stack deployment demonstrating production-ready ML system architecture.

The remainder of this paper is organized as follows. Section 2 reviews related work on air quality prediction, machine learning for environmental monitoring, and health advisory systems. Section 3 details the methodology including dataset generation, model architectures, and system design. Section 4 presents experimental results. Section 5 discusses implications, limitations, and future directions. Section 6 concludes.

2. Literature Review

2.1 Air Quality Prediction Using Machine Learning

Machine learning approaches for air quality forecasting have gained significant traction over the past decade. Kumar and Goyal [4] employed multiple linear regression and artificial neural networks for AQI prediction in Delhi, demonstrating that nonlinear models substantially outperform linear approaches for capturing the complex interactions between pollutant species. Their work established the foundation for ML-based AQI prediction in Indian urban environments.

Zheng et al. [5] developed a comprehensive urban air quality inference framework using heterogeneous data sources including meteorological observations, traffic patterns, and point-of-interest data. Their spatiotemporal approach achieved significant improvements over single-source models, highlighting the importance of multi-modal data integration, a principle that informs the real-time data acquisition architecture presented in this paper.

Sharma et al. [6] conducted a comparative study of decision trees, random forests, and gradient boosting methods for AQI classification in Indian cities, finding that ensemble methods consistently outperform individual learners. Their findings directly motivate the inclusion of Random Forest and XGBoost in the model comparison framework of this study.

2.2 Ensemble Learning and Gradient Boosting

The theoretical foundations of ensemble learning provide critical context for the model architectures employed in this study. Breiman [7] introduced the Random Forest algorithm, demonstrating that aggregating predictions from multiple decorrelated decision trees through bootstrap aggregation (bagging) substantially reduces variance while maintaining low bias. The Random Forest's inherent resistance to overfitting and its ability to provide feature importance rankings make it particularly well-suited for environmental prediction tasks.

Chen and Guestrin [8] proposed XGBoost, a scalable gradient boosting framework that introduces regularization techniques to the boosting paradigm. XGBoost's key innovations, including column subsampling, shrinkage, and efficient handling of sparse data, have made it the algorithm of choice for a wide range of structured prediction tasks. The application of XGBoost to air quality prediction represents a natural extension of its demonstrated superiority in regression benchmarks.

2.3 Neural Network Approaches

The application of neural networks to environmental prediction has a rich history. Kaur and Gupta [9] provide a comprehensive review of artificial neural networks, tracing their development from simple perceptrons to complex deep learning architectures. Their survey highlights how fundamental principles of neural computation, learning from data, generalizing to unseen inputs, and adapting through iterative optimization, are directly applicable to the environmental prediction task investigated in this study.

Kaur and Gupta [10] propose a constructive neural network framework that addresses the challenge of determining optimal network topology through performance-driven incremental construction. This concept of starting from minimal configurations and iteratively adding complexity bears direct relevance to the Multi-Layer Perceptron architecture evaluated in this study.

The bipolar sigmoid algorithm proposed by Kaur and Gupta [11] for constructive neural networks demonstrates that careful selection of activation functions can significantly impact training efficiency and convergence. Their extended

bipolar sigmoid algorithm [12] further enhances performance through additional optimization techniques, establishing the principle that systematic refinement of algorithmic components yields measurable performance gains.

2.4 Real-Time Environmental Monitoring Systems

The integration of intelligent computing with real-time monitoring systems has opened new avenues for environmental applications. Dargan, Gupta, and Singh [13] present a blockchain-based energy management system model that combines intelligent algorithms with distributed ledger technology for transparent monitoring and decision-making. Their approach illustrates the convergence of AI and secure distributed systems for environmental applications.

Dargan and Gupta [14] explore blockchain integration for IoT communication, proposing architectural frameworks that combine cryptographic security with intelligent data processing. The concepts of reliable data acquisition and verification discussed in their work parallel the multi-tier API fallback strategy implemented in this paper's real-time data pipeline.

Dewan, Duhan, and Gupta [15] demonstrate how intelligent systems can be combined with blockchain technology to address trust and security concerns in applications requiring high integrity. Their work on secure electronic voting illustrates the broader trend of applying computational intelligence to systems where data reliability is paramount, a consideration directly relevant to environmental monitoring systems where sensor data quality is critical.

2.5 Health Impact Assessment and Advisory Systems

The translation of air quality measurements into health advisories represents a crucial but understudied component of environmental monitoring systems. The CPCB framework [3] defines five AQI categories, Good, Satisfactory, Moderate, Poor, Very Poor, and Severe, each associated with specific health implications and recommended actions. However, existing monitoring platforms typically present only numerical AQI values without personalized, actionable guidance [16].

3. Methodology

3.1 Dataset Generation and Validation

The dataset used in this study comprises 10,000 synthetically generated air quality records across eight major Indian cities: Delhi, Mumbai, Chennai, Kolkata, Bangalore, Hyderabad, Bhopal, and Pune. Each record contains six pollutant concentration measurements (PM_{2.5}, PM₁₀, NO₂, SO₂, CO, O₃) and a mathematically computed AQI value.

The AQI for each record is calculated using the CPCB sub-index formula:

Sub-index formula:

$$I_p = ((I_{HI} - I_{LO}) / (BP_{HI} - BP_{LO})) \times (C_p - BP_{LO}) + I_{LO}$$

Where I_p is the sub-index for pollutant p , C_p is the measured concentration, BP_{HI} and BP_{LO} are the concentration breakpoints bounding C_p , and I_{HI} and I_{LO} are the corresponding index breakpoints. The overall AQI is the maximum sub-index across all pollutants:

$$AQI = \max(I_{PM2.5}, I_{PM10}, I_{NO2}, I_{SO2}, I_{CO}, I_{O3})$$

Table 1 presents the CPCB breakpoint values used for dataset generation.

Table 1: CPCB Breakpoint Values for AQI Sub-Index Calculation

Pollutant	Good (0-50)	Satisfactory (51-100)	Moderate (101-200)	Poor (201-300)	Very Poor (301-400)	Severe (401-500)
PM _{2.5} (µg/m ³)	0-30	31-60	61-90	91-120	121-250	251-380
PM ₁₀ (µg/m ³)	0-50	51-100	101-250	251-350	351-430	431-550
NO ₂ (µg/m ³)	0-40	41-80	81-180	181-280	281-400	401-500

SO ₂ (µg/m ³)	0-40	41-80	81-380	381-800	801-1600	1601-2000
CO (mg/m ³)	0-1.0	1.1-2.0	2.1-10.0	10.1-17.0	17.1-34.0	34.1-50.0
O ₃ (µg/m ³)	0-50	51-100	101-168	169-208	209-748	749-1000

The use of synthetically generated data with CPCB-validated AQI labels ensures mathematical consistency between pollutant concentrations and AQI values, eliminating the label noise that commonly affects real-world air quality datasets where AQI values may be computed using different methodologies or contain measurement errors.

3.2 Data Preprocessing Pipeline

The preprocessing pipeline is implemented as a scikit-learn Pipeline object to ensure consistent transformation during both training and inference:

Step 1: Missing Value Imputation: A SimpleImputer with median strategy handles missing or null values in pollutant measurements. Median imputation is preferred over mean imputation due to its robustness to outliers common in environmental data.

Step 2: Feature Scaling: A StandardScaler applies Z-score normalization (zero mean, unit variance) to all features. This step is critical for models sensitive to feature magnitude, particularly SVR and MLP, while being benign for tree-based models.

Step 3: Model Training: The preprocessed features are passed to the regression model for training.

The dataset is split into 80% training (8,000 samples) and 20% testing (2,000 samples) using a fixed random seed (42) for reproducibility.

3.3 Machine Learning Models

Five regression models are comparatively evaluated, spanning linear, tree-based, ensemble, gradient boosting, and neural network paradigms:

Table 2: Model Configurations

Model	Key Hyperparameters
Linear Regression	Default (OLS baseline)
Decision Tree	random_state=42
Random Forest	n_estimators=100, random_state=42, n_jobs=-1
MLP Neural Network	hidden_layers=(128,64,32), activation=ReLU, solver=Adam, max_iter=500
XGBoost	n_estimators=200, learning_rate=0.05, max_depth=6

Linear Regression serves as the baseline, modeling AQI as a linear combination of pollutant concentrations. **Decision Tree Regressor** captures nonlinear relationships through recursive binary partitioning. **Random Forest** aggregates 100 decorrelated decision trees via bagging [7]. **MLP Neural Network** employs three hidden layers (128→64→32 neurons) with ReLU activation and Adam optimizer, representing the deep learning paradigm [9]. **XGBoost** implements regularized gradient boosting with 200 sequential trees [8].

3.4 Evaluation Metrics

Six metrics are employed for comprehensive model evaluation:

- **MAE (Mean Absolute Error):** Average absolute deviation between predicted and actual AQI values.
- **MSE (Mean Squared Error):** Average squared deviation, penalizing larger errors.

- **RMSE (Root Mean Squared Error):** Square root of MSE, in the same units as AQI.
- **R** Proportion of variance explained by the model (1.0 = perfect).² (**Coefficient of Determination**):
- **MAPE (Mean Absolute Percentage Error):** Percentage-based error metric for interpretability.
- **CV R** 5-fold cross-validated R^2 (**Cross-Validated R^2**):² on the training set, assessing generalization.

Additionally, per-AQI-category performance is evaluated to assess model accuracy across different pollution severity levels.

3.5 Health Advisory Engine

A rule-based health advisory engine maps predicted AQI values to five severity categories, each producing specific health recommendations aligned with CPCB and US EPA guidelines:

Table 3: AQI Categories and Health Advisory Framework

AQI Range	Category	Advisory Actions
0-50	Good	No precautions needed; ideal for outdoor activities
51-100	Moderate	Sensitive individuals should limit prolonged outdoor exertion
101-200	Unhealthy (Sensitive)	Reduce outdoor activities; consider wearing masks; use air purifiers
201-300	Very Unhealthy	Avoid outdoor exercise; wear N95 masks; close windows and doors
301+	Hazardous	Emergency conditions, stay indoors; seek medical attention if symptomatic

3.6 System Architecture

The system is deployed as a three-tier architecture comprising a React frontend, FastAPI backend, and ML model service. The backend exposes RESTful API endpoints for AQI prediction, historical data retrieval, city statistics, and real-time pollutant data acquisition.

Real-time data is obtained through a multi-tier fallback strategy:

- **Tier 1 (Primary):** WAQI (World Air Quality Index) API providing ground-station measurements
- **Tier 2 (Fallback):** OpenWeatherMap Air Pollution API providing satellite/modelled data

The system implements unit conversion between measurement standards (ppb $\leftrightarrow$ $\mu\text{g}/\text{m}^3$) using molecular weight conversion factors, and includes sanity checking for anomalous sensor readings (e.g., rejecting temperature values below -20°C or above 55°C). A novel AQI-to-concentration reverse mapping using US EPA breakpoints enables conversion of WAQI sub-index values back to raw pollutant concentrations.

4. Results

4.1 Model Comparison

Table 4 presents the comprehensive performance comparison of all five models on the held-out test set (2,000 samples).

Table 4: Comparative Performance of Machine Learning Models on Test Set

Model	MAE	MSE	RMSE	R ²	MAPE	Accuracy%	CV R ² (±std)
XGBoost	2.40	40.80	6.39	0.9938	1.54%	98.46%	0.9889 ± 0.0020
Random Forest	2.17	80.43	8.97	0.9878	1.37%	98.63%	0.9877 ± 0.0026
Decision Tree	2.20	122.80	11.08	0.9814	1.34%	98.66%	0.9787 ± 0.0041
Deep Learning (MLP)	5.63	155.57	12.47	0.9764	2.95%	97.05%	0.9728 ± 0.0047
Linear Regression	26.75	1090.83	33.03	0.8347	13.15%	86.85%	0.8290 ± 0.0035

XGBoost achieved the highest R² (0.9938) and lowest RMSE (6.39), outperforming all other models. The performance hierarchy reveals a clear pattern: gradient boosting (XGBoost) > bagging ensemble (Random Forest) > single tree (Decision Tree) > neural network (MLP) > linear model. All ensemble and tree-based models achieved R² > 0.97, confirming that the relationship between pollutant concentrations and CPCB-formula AQI is predominantly captured by nonlinear, threshold-based decision boundaries that tree architectures naturally model.

Linear Regression, serving as the baseline, achieved R² = 0.8347, notably lower than all nonlinear models. This 16 percentage point gap confirms that the CPCB breakpoint formula introduces substantial nonlinearity that linear models cannot adequately capture, consistent with findings by Kumar and Goyal [4].

The MLP neural network (R² = 0.9764), while substantially outperforming linear regression, underperformed relative to tree-based approaches. This is consistent with the observation that neural networks typically require larger datasets and more extensive hyperparameter optimization to compete with well-tuned ensemble methods on tabular regression tasks [9, 10].

4.2 Cross-Validation Analysis

The 5-fold cross-validation R² scores provide insight into model stability and generalization capacity. XGBoost achieved a CV R² of 0.9889 ± 0.0020, exhibiting the lowest standard deviation among all models. This narrow confidence band indicates robust generalization across data partitions. The close agreement between test-set R² (0.9938) and CV R² (0.9889), a gap of only 0.0049, suggests minimal overfitting.

Random Forest similarly demonstrated strong generalization with CV R² = 0.9877 ± 0.0026. The Decision Tree exhibited the highest variance (CV R² = 0.9787 ± 0.0041), consistent with the known instability of individual decision trees that ensemble methods are designed to mitigate [7].

4.3 Per-Category Performance

Table 5 presents the model performance broken down by AQI health category, evaluated using the best model (XGBoost).

Table 5: XGBoost Performance by AQI Health Category

AQI Category	Range	Test Samples (n)	MAE	R ²
Good	0-50	0	N/A	N/A
Moderate	51-100	29	14.86	-0.8330
Unhealthy (Sensitive)	101-200	533	4.56	0.8873

Very Unhealthy	201-300	616	2.41	0.9326
Hazardous	301+	813	0.56	0.9980

The model demonstrates a clear improvement in prediction accuracy at higher pollution levels: Hazardous conditions are predicted with exceptional precision ($MAE = 0.56$, $R^2 = 0.9980$), while the Moderate category exhibits weaker performance ($MAE = 14.86$, $R^2 = -0.8330$). This pattern is attributable to two factors: (1) the Moderate category contains significantly fewer training samples (29 vs. 813 for Hazardous), and (2) the CPCB breakpoint formula produces narrower AQI ranges for lower pollution levels, making boundary predictions more challenging.

The negative R^2 for the Moderate category indicates that, for this small subset, the model's predictions have higher variance than a constant mean predictor. However, given the extremely small sample size ($n=29$), this metric should be interpreted cautiously.

4.4 Sample Predictions

Table 6 presents representative sample predictions from the XGBoost model, illustrating the close agreement between actual and predicted AQI values across the pollution spectrum.

Table 6: Sample Predictions, XGBoost Model

Actual AQI	Predicted AQI	Absolute Error
349.7	350.1	0.4
345.9	346.0	0.1
302.5	302.4	0.1
268.7	267.4	1.3
137.5	136.9	0.6
303.8	303.7	0.1
185.0	185.1	0.1
330.7	330.2	0.5
392.6	392.5	0.1

The majority of predictions exhibit absolute errors below 1.0 AQI units, demonstrating that the XGBoost model effectively learns the CPCB breakpoint formula's piecewise linear structure.

5. Discussion

5.1 Interpretation of Results

The experimental results demonstrate that ensemble and gradient boosting methods are highly effective for AQI prediction from pollutant concentrations. XGBoost's superior performance ($R^2 = 0.9938$) can be attributed to its regularized boosting framework, which sequentially corrects residual errors while preventing overfitting through L1/L2 regularization [8]. The close performance of Random Forest ($R^2 = 0.9878$) confirms the effectiveness of bagging-based ensembles for this task.

The performance hierarchy observed in this study, XGBoost > Random Forest > Decision Tree > MLP > Linear Regression, aligns with the broader machine learning literature on tabular data, where gradient boosted trees consistently outperform neural networks without extensive feature engineering and hyperparameter optimization [8]. This finding is particularly relevant for environmental monitoring applications where model interpretability and deployment efficiency are valued alongside predictive accuracy.

The constructive neural network approaches proposed by Kaur and Gupta [10, 11, 12] offer an illuminating parallel to the ensemble methods evaluated in this study. Just as their constructive algorithms iteratively build optimal network topologies from minimal starting points, ensemble methods like Random Forest and XGBoost construct complex predictive functions by aggregating simpler base learners. Both paradigms demonstrate that iterative, performance-driven construction is a broadly effective optimization strategy.

5.2 Real-Time Data Integration

The multi-tier API fallback architecture addresses a critical challenge in operational air quality monitoring: the unreliability of any single data source. Ground-station networks (accessed via WAQI) provide the most accurate measurements but suffer from spatial sparsity and frequent sensor outages. Satellite-derived estimates (via OpenWeatherMap) offer broader coverage but reduced accuracy.

The system's unit conversion layer handles the heterogeneity of measurement standards across data sources. WAQI reports pollutant values as US EPA sub-indices, requiring reverse engineering through the EPA breakpoint formula to recover raw concentrations. OpenWeatherMap reports concentrations in $\mu\text{g}/\text{m}^3$, requiring conversion to ppb using molecular weight factors:

- NO_2 : $1 \text{ ppb} \times 1.88 = 1 \mu\text{g}/\text{m}^3$
- SO_2 : $1 \text{ ppb} \times 2.62 = 1 \mu\text{g}/\text{m}^3$
- O_3 : $1 \text{ ppb} \times 1.96 = 1 \mu\text{g}/\text{m}^3$
- CO : $1 \text{ ppb} \times 1.145/1000 = 1 \text{ mg}/\text{m}^3$

The implementation of sensor anomaly detection, rejecting physically implausible values such as temperatures below -20°C in tropical Indian cities, adds robustness to the data pipeline. This approach aligns with the principles of reliable data acquisition discussed by Dargan and Gupta [14] in the context of IoT communication systems.

5.3 Health Advisory System

The rule-based health advisory engine represents a deliberate design choice prioritizing reliability and interpretability over complexity. While ML-based advisory generation could theoretically personalize recommendations based on individual health profiles, the deterministic rule-based approach ensures consistent, medically validated guidance that can be audited and verified by public health authorities.

The advisory framework generates 4-7 specific, actionable recommendations per AQI category, progressing from general guidance ("Ideal conditions for outdoor activities") at low AQI levels to urgent instructions ("Emergency conditions, stay indoors; seek immediate medical attention if experiencing breathing difficulty") at hazardous levels. This graduated response structure aligns with international best practices for environmental health communication [3, 16].

5.4 Implications for Intelligent Computing Systems

The findings of this study have broader implications for the design of intelligent environmental computing systems. The integration of ML prediction with real-time data acquisition and health advisory generation demonstrates a complete intelligent computing pipeline that transforms raw sensor data into actionable knowledge. This approach aligns with the vision of intelligent computing systems discussed by Kaur and Gupta [9], where computational intelligence is applied to solve real-world problems.

The principles of reliable data verification implemented in the real-time pipeline resonate with the blockchain-based approaches of Dewan et al. [15] and Dargan et al. [13, 14]. The immutable logging of predictions, data source provenance, and advisory outputs on a distributed ledger could enhance the trustworthiness of AI-driven environmental monitoring systems, particularly in regulatory contexts where accountability is essential.

5.5 Limitations

This study has several important limitations:

(1) **Synthetic dataset**, While the CPCB-formula-based dataset ensures mathematical consistency, it may not capture the full complexity of real-world atmospheric chemistry, including secondary pollutant formation, photochemical reactions, and meteorological interactions.

(2) **Feature scope**, The model uses only six pollutant features. Incorporating meteorological variables (temperature, humidity, wind speed, precipitation) and temporal features (hour, season) could improve predictive accuracy for real-world deployment.

(3) **Spatial generalization**, The model is trained on data from eight Indian cities and may not generalize to cities with different pollution profiles, industrial compositions, or geographic characteristics.

(4) **Advisory validation**, The health advisory engine has not been clinically validated through epidemiological studies or medical review.

(5) **Single-point prediction**, The system predicts current AQI from current pollutant concentrations; it does not perform temporal forecasting of future AQI values.

6. Conclusion

This paper presented an end-to-end AI-driven system for Air Quality Index prediction and real-time health advisory generation, evaluated through a comparative analysis of five machine learning models on a CPCB-formula-validated dataset of 10,000 samples across eight major Indian cities.

The experimental results establish a clear performance hierarchy among the evaluated models. Linear Regression, serving as the baseline, achieved $R^2 = 0.8347$, confirming that the CPCB breakpoint formula introduces substantial nonlinearity requiring more sophisticated modeling approaches. The Decision Tree ($R^2 = 0.9814$) and Random Forest ($R^2 = 0.9878$) demonstrated the natural affinity of tree-based architectures for the piecewise linear structure of the AQI formula. The MLP neural network ($R^2 = 0.9764$), while effective, underperformed relative to tree-based methods, consistent with the literature on neural networks for tabular prediction tasks [9, 10]. XGBoost achieved the highest overall performance with $R^2 = 0.9938$, $MAE = 2.40$, and $MAPE = 1.54\%$, demonstrating that regularized gradient boosting effectively captures the complex, threshold-based relationships in AQI computation.

Per-category analysis revealed that model accuracy improves with pollution severity: the Hazardous category was predicted with exceptional precision ($MAE = 0.56$, $R^2 = 0.9980$), while the Moderate category showed weaker performance due to limited training representation. This finding has important practical implications, the model is most accurate precisely in the conditions where accurate prediction is most critical for public health.

The system's multi-tier real-time data acquisition architecture, with automatic fallback between WAQI ground-station data and OpenWeatherMap satellite estimates, addresses the practical challenge of data source reliability in operational deployment. The health advisory engine translates predicted AQI values into specific, actionable recommendations across five severity categories, bridging the gap between raw environmental measurements and public health guidance.

The parallels between the ML pipeline presented in this paper and the constructive neural network frameworks of Kaur and Gupta [10, 11, 12] suggest a unifying principle: iterative, data-driven optimization from simple components to complex predictive systems is effective across different levels of computational intelligence. Combined with the secure data management frameworks proposed through blockchain integration [13, 14, 15], these approaches point toward a future of more reliable, transparent, and actionable intelligent environmental monitoring systems.

Future work should investigate: (a) incorporating temporal features and LSTM/Transformer architectures for AQI time-series forecasting; (b) integrating IoT sensor networks for hyperlocal pollution monitoring; (c) developing mobile applications with push notification alerts for real-time health advisories; (d) applying model interpretability techniques (SHAP, LIME) for pollution source attribution; (e) extending the system to international air quality standards (US EPA, European EEA); and (f) conducting clinical validation of the health advisory framework through epidemiological studies.

References

- [1] World Health Organization. "Ambient (Outdoor) Air Pollution." WHO Fact Sheet, 2022. [https://www.who.int/news-room/fact-sheets/detail/ambient-\(outdoor\)-air-quality-and-health](https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health)
- [2] Balakrishnan, K., Dey, S., Gupta, T., et al. "The Impact of Air Pollution on Deaths, Disease Burden, and Life Expectancy Across the States of India." *The Lancet Planetary Health* 3(1) (2019): e26-e39.

- [3] Central Pollution Control Board. "National Air Quality Index." CPCB, Ministry of Environment, Forest and Climate Change, Government of India (2014).
- [4] Kumar, A., and Goyal, P. "Forecasting of Air Quality in Delhi Using Principal Component Regression Technique." *Atmospheric Pollution Research* 2(4) (2011): 436-444.
- [5] Zheng, Y., Liu, F., and Hsieh, H. P. "U-Air: When Urban Air Quality Inference Meets Big Data." *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2013): 1436-1444.
- [6] Sharma, E., Deo, R. C., Prasad, R., and Parisi, A. V. "A Hybrid Air Quality Early-Warning Framework: An Hourly Forecasting Model with Online Sequential Extreme Learning Machine and Empirical Mode Decomposition Algorithm." *Science of The Total Environment* 709 (2020): 135934.
- [7] Breiman, L. "Random Forests." *Machine Learning* 45(1) (2001): 5-32.
- [8] Chen, T., and Guestrin, C. "XGBoost: A Scalable Tree Boosting System." *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2016): 785-794.
- [9] Kaur, J., and Gupta, N. "Artificial Neural Network: A Review." *International Journal of Technical Research & Science* (n.d.): 1-4.
- [10] Kaur, J., and Gupta, N. "Constructive Neural Network: A Framework." *International Journal of Engineering and Advanced Technology* 9(2) (2019). <https://doi.org/10.35940/ijeat.b3304.129219>
- [11] Kaur, J., and Gupta, N. "Bipolar Sigmoid Algorithm for Designing Constructive Neural Network." *International Journal on Emerging Technologies* 11(2) (2020): 991-996.
- [12] Kaur, J., and Gupta, N. "Extended Bipolar Sigmoid Algorithm for Enhancing Performance of Constructive Neural Network." *International Journal on Emerging Technologies* 11(2) (2020): 1034-1038.
- [13] Dargan, J., Gupta, N., and Singh, L. "Blockchain-Based Energy Management System: A Proposed Model." In *2021 International Conference on Technological Advancements and Innovations (ICTAI)*, pp. 510-514. IEEE (2021). <https://doi.org/10.1109/ICTAI53825.2021.9673233>
- [14] Dargan, J., and Gupta, N. "Integration of Blockchain for IoT Communication." In *Adaptive Power Quality for Power Management Using Smart Technologies*, pp. 319-343. CRC Press (2023).
- [15] Dewan, S., Duhan, L., and Gupta, N. "Secure Electronic Voting System Based on Mobile App and Blockchain." *Recent Advances in Computer Science and Communications* 14(9) (2021). <https://doi.org/10.2174/2666255813999200821161058>
- [16] Morawska, L., Thai, P. K., Liu, X., et al. "Applications of Low-Cost Sensing Technologies for Air Quality Monitoring and Exposure Assessment." *Environment International* 116 (2018): 286-299.
- [17] Pedregosa, F., Varoquaux, G., Gramfort, A., et al. "Scikit-learn: Machine Learning in Python." *Journal of Machine Learning Research* 12 (2011): 2825-2830.
- [18] Wei, J., Wang, X., Schuurmans, D., et al. "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models." *Advances in Neural Information Processing Systems* 35 (2022): 24824-24837.
- [19] Brown, T. B., Mann, B., Ryder, N., et al. "Language Models are Few-Shot Learners." *Advances in Neural Information Processing Systems* 33 (2020): 1877-1901.
- [20] Liu, P., Yuan, W., Fu, J., et al. "Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing." *ACM Computing Surveys* 55(9) (2023): 1-35.