A Comparative Study of Machine Learning Models for Air Quality Index (AQI) Prediction Using Multi-Pollutant and Meteorological Parameters in Maharashtra, India

Nitin E.Kakade, Chhaya S Galande

Abstract: *Air pollution has emerged as a critical environmental and public health concern in rapidly urbanizing regions of India. Maharashtra, hosting several highly industrialized and densely populated metropolitan centers, exhibits considerable spatiotemporal variation in air quality influenced by pollutant concentrations, meteorological conditions, seasonal cycles, and anthropogenic activities. This study proposes an explainable machine learning framework for predicting the Air Quality Index (AQI) in Maharashtra, India, utilizing comprehensive pollutant and meteorological parameters. The framework incorporates particulate matter ($PM_{2.5}$ and PM_{10}), nitrogen dioxide (NO_2), sulfur dioxide (SO_2), carbon monoxide (CO), ozone (O_3), temperature, relative humidity, wind speed, and other relevant environmental variables. Multiple machine learning algorithms, including Linear Regression, Decision Tree, Random Forest, Support Vector Regression, XGBoost, LightGBM, Artificial Neural Network, and Long Short-Term Memory (LSTM), are comparatively evaluated. Model performance is assessed using Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), and coefficient of determination (R^2). Beyond prediction accuracy, Explainable Artificial Intelligence (XAI) techniques, particularly SHAP (SHapley Additive exPlanations), are employed to identify the relative contribution of individual pollutants and meteorological parameters to AQI predictions. The proposed approach aims to provide both accurate AQI prediction and interpretable insights regarding the major factors influencing air pollution in Maharashtra. The results are expected to support data-driven air-quality management and provide an interpretable framework for environmental decision-making.*

Nitin E.Kakade (nitin.kakade@pravara.in) Department of Computer Science, Padmashri Vikhe Patil College of Arts, Science and Commerce, Pravaranagar.
Chhaya S Galande (csgalande27@gmail.com) Head, Department of Computer Science, Padmashri Vikhe Patil College of Arts, Science and Commerce, Pravaranagar.

ISSN: 0262-8104 (print) 1756-3488 (online)
Waterlines Vol. 44 No. 1s

Keywords: *Air Quality Index, Machine Learning, Explainable AI, SHAP, $PM_{2.5}$, PM_{10}, Maharashtra, Air Pollution Prediction, XGBoost, LSTM*

Introduction

Air pollution constitutes one of the most significant environmental challenges affecting human health, ecosystems, and quality of life in contemporary India. As the world's most populous nation, India ranks as the fifth most polluted country globally, preceded only by Chad, Congo, Bangladesh, and Pakistan based on Air Quality Index values. Rapid urbanization, industrialization, transportation activities, construction, biomass burning, and escalating energy consumption have collectively contributed to deteriorating air quality across numerous Indian states. Maharashtra, as one of India's most economically developed states, encompasses major metropolitan areas, industrial corridors, commercial centers, and transportation hubs. Consequently, air pollution exhibits substantial spatial and seasonal variation across different regions of the state.

The Air Quality Index (AQI) provides a simplified, accessible representation of air pollution by converting concentrations of multiple individual pollutants into an integrated air-quality indicator. Although conventional air-quality monitoring systems operated by the Central Pollution Control Board (CPCB) and Maharashtra Pollution Control Board (MPCB) provide valuable observational data, monitoring stations alone may be insufficient for forecasting future pollution levels or understanding the complex dynamics driving air quality variation. Air pollution is influenced by intricate, nonlinear interactions among multiple pollutants, meteorological conditions, geographical characteristics, and temporal patterns that are challenging to capture using traditional statistical approaches.

Machine learning (ML) techniques offer a powerful opportunity to model these complex relationships. Unlike conventional statistical methods that often assume linearity, ML algorithms can identify nonlinear patterns and interactions within large environmental datasets. Algorithms such as Random Forest, Support Vector Regression, XGBoost, LightGBM, Artificial Neural Networks, and LSTM have demonstrated considerable utility for air-quality prediction tasks.

However, prediction accuracy alone is insufficient for environmental decision-making. A highly accurate model may function as a "black box," providing limited information about why a particular AQI prediction was generated. Environmental authorities, policymakers, and researchers require understanding of which pollutants and environmental conditions contribute most strongly to poor air quality to develop effective mitigation strategies. Explainable Artificial Intelligence (XAI) provides a solution to this challenge by making machine-learning predictions more interpretable and transparent.

This research therefore proposes an explainable machine-learning framework for AQI prediction in Maharashtra. The study combines comparative ML modeling with SHAP-based interpretation to determine both the best-performing prediction model and the major factors responsible for AQI variation, thereby bridging the gap between predictive accuracy and environmental interpretability.

1.1 Problem Statement

Existing air-quality prediction studies frequently focus on predicting individual pollutants such as $PM_{2.5}$ or PM_{10}. Although such predictions are useful, they do not always provide an integrated representation of overall air quality. Furthermore, many machine-learning models prioritize predictive accuracy while providing limited explanation of the factors responsible for their predictions.
Therefore, there is a need for an integrated framework that:

1. Predicts AQI accurately using comprehensive multi-pollutant data.
2. Incorporates both pollutant and meteorological variables.
3. Compares diverse machine-learning algorithms systematically.
4. Evaluates temporal and nonlinear relationships in air quality data.
5. Explains the contribution of individual features to model predictions.
6. Provides interpretable information useful for environmental management and policy formulation.

1.2 Research Objectives

The main objectives of this research are:

1. To develop a machine-learning framework for AQI prediction in Maharashtra.
2. To investigate the relationship between AQI, pollutant concentrations, and meteorological variables.
3. To compare the predictive performance of different machine-learning algorithms.
4. To identify the most influential variables affecting AQI prediction.
5. To apply SHAP-based explainable AI techniques to interpret model predictions.
6. To examine seasonal and spatial variations in air-quality prediction across Maharashtra.

7. To develop an interpretable framework that can support air-quality monitoring and management.

1.3 Research Questions

The study addresses the following research questions:

* Which machine-learning algorithm provides the highest accuracy for AQI prediction?
* Which pollutant contributes most strongly to AQI variation?
* What is the contribution of meteorological parameters to AQI prediction?
* Can lagged environmental variables improve prediction performance?
* Does the best predictive model remain consistent across different Maharashtra cities?
* Can SHAP provide meaningful, actionable explanations of machine-learning predictions?

Literature Review

2.1 Air Quality Prediction: Evolution and Approaches

Air-quality prediction has traditionally relied on statistical and deterministic models, including Chemical Transport Models (CTMs) like CMAQ, which simulate atmospheric physics and chemistry. While such physics-based models can achieve approximately 85% accuracy, they require supercomputing resources that are impractical for real-time applications. Physical models like Weather Research and Forecasting (WRF) models derive atmospheric wind fields from temperature, humidity, and pressure observations, but they are sensitive to initial conditions, computationally demanding, and may become impractical for larger areas, hindering real-time predictions.

2.2 Machine Learning for Air Pollution Prediction

Several supervised-learning algorithms have been applied to air-pollution prediction, each with distinct strengths and limitations.

Linear Regression

Linear Regression provides a baseline model by assuming a linear relationship between predictor variables and AQI. It is simple and interpretable but may not adequately represent nonlinear pollutant interactions. A study on Pune air quality achieved an R-value of 0.9611 using Linear Regression, with RMSE of 21.4079, MAPE of 7.8945%, and MAE of 13.5884 . The study utilized data spanning nineteen years (2006–2024) .

Decision Tree

Decision Tree models divide observations into groups using a sequence of decision rules. They can capture nonlinear relationships and are relatively easy to interpret. However, they are prone to overfitting without proper pruning.

Random Forest

Random Forest combines multiple decision trees to improve generalization and reduce overfitting. It is particularly suitable for environmental datasets containing nonlinear relationships and interacting variables. Research has demonstrated the effectiveness of Random Forest in air quality prediction, with studies utilizing satellite data and machine learning techniques to assess PM2.5 concentrations and variability across Maharashtra . A hybrid Random Forest-ARIMA model with SHAP-based explainability has been proposed for AQI forecasting, emphasizing interpretability alongside predictive performance .

Support Vector Regression

Support Vector Regression (SVR) uses support-vector principles to model nonlinear relationships through kernel functions. It can perform effectively on datasets with complex relationships but requires appropriate parameter tuning. Studies on Mumbai air quality have evaluated different kernels of Support Vector Machine (SVM) with K-Nearest Neighbour (KNN) for AQI prediction, validating the proposed method through comparative performance analysis .

XGBoost

XGBoost is an ensemble gradient-boosting algorithm that builds trees sequentially to reduce prediction errors. It has become widely used for structured environmental datasets because of its predictive performance and ability to capture nonlinear interactions. A two-stage XGBoost pipeline has demonstrated exceptional performance with MAE of 0.35, RMSE of 0.54, and R^2 of 1.00 for AQI prediction, far outperforming conventional single-step baseline models .

LightGBM

LightGBM is another gradient-boosting framework designed for efficient training. Its computational efficiency makes it suitable for large datasets.

Artificial Neural Network

ANN models can learn complex nonlinear relationships through interconnected computational neurons. They are useful when relationships between pollutant and meteorological variables are highly nonlinear. Systematic studies on PM2.5 and PM10 concentration prediction have evaluated various machine learning and deep learning models . A Deep Convolutional Neural Network (DCNN)-LSTM model achieved superior predictive performance with an accuracy of 97.48% and an F1 score of 97.48% .

Long Short-Term Memory

LSTM is a recurrent neural-network architecture designed to model sequential data. Because air pollution exhibits strong temporal dependencies, LSTM can learn patterns from previous observations effectively. Studies have shown that EMD-LSTM-Bayesian improves accuracy by 35.07% over LSTM alone, with LSTM recording the highest MAE of 0.593 and RMSE of 0.804 in some comparisons .

2.3 Explainable Artificial Intelligence in Environmental Applications

Traditional ML models may produce accurate predictions without clearly explaining their decisions. Explainable AI attempts to address this limitation by making model behavior transparent and interpretable.

SHAP is based on Shapley-value concepts from cooperative game theory. It estimates the contribution of individual features to a particular model prediction, determining whether a feature increases or decreases the predicted AQI and quantifying the strength of its contribution. Recent studies have emphasized the importance of SHAP-based explainability in air quality forecasting, making feature influence transparent and providing actionable insights for environmental management .

In an environmental application, SHAP can answer critical questions such as:

* How strongly does $PM_{2.5}$ influence AQI?
* Does PM_{10} increase predicted AQI under specific conditions?
* Does wind speed reduce predicted AQI?
* How does temperature influence model predictions?
* Which pollutant is the dominant predictor across different locations?
* How do traffic patterns and meteorological conditions interact to affect pollution levels?

Materials and Methods

3.1 Study Area

The proposed study focuses on Maharashtra, India. Maharashtra contains diverse urban, industrial, semi-urban, and rural environments. Major cities such as Mumbai, Pune, Nagpur, Nashik, Thane, Chhatrapati Sambhajinagar, and other monitored locations are considered based on data availability. The selected locations contain sufficiently continuous air-quality observations to support machine-learning analysis. Research on Pune has utilized data spanning nineteen years (2006–2024) , while other studies have employed data from 2019-2023 for Maharashtra-wide analysis.

3.2 Data Sources

The study uses officially published air-quality and meteorological observations. Sources include:

* Central Pollution Control Board (CPCB)
* Maharashtra Pollution Control Board (MPCB)
* Government environmental monitoring stations
* Public datasets (e.g., Hugging Face datasets with 708,000 records covering 2017-2024 for Maharashtra and other states)
* Satellite data (e.g., Sentinel-5P for NO_2 concentration analysis)

A comprehensive dataset may include multiple monitoring stations across Maharashtra cities. For example, a Hugging Face dataset contains 708,000 records with data from 30 states, 277 cities, and 535 stations, including Maharashtra cities such as Mumbai (Navy Nagar-Colaba, Powai, Sewri, Shivaji Nagar, Sion), with parameters including PM2.5, PM10, NO2, SO2, CO**,** Ozone, temperature, humidity, wind speed, and rainfall .

3.3 Dataset Variables

The dataset contains the following variables, with real-world values from Maharashtra monitoring stations:

Variable	Description	Sample Values (Mumbai)
Date/Time	Observation date and time	2017-01-01 to 2024-12-31
Location	Monitoring station/city	Navy Nagar-Colaba, Powai, Sewri, Shivaji Nagar, Sion
$PM_{2.5}$ ($\mu g/m^3$)	Fine particulate matter	0.53 to 427 (mean: 46.22)
PM_{10} ($\mu g/m^3$)	Coarse particulate matter	0.03 to 1,000

NO_2 ($\mu g/m^3$)	Nitrogen dioxide	0 to 497
SO_2 ($\mu g/m^3$)	Sulfur dioxide	0 to 200
CO (mg/m^3)	Carbon monoxide	0 to 47.9
O_3 ($\mu g/m^3$)	Ozone	0 to 447
Temperature (°C)	Atmospheric temperature	17.25 to 36.52 (mean: 26.12)
Humidity (%)	Relative humidity	18.14 to 97.91 (mean: 67.80)
Wind Speed (m/s)	Wind velocity	U-Wind: -4.17 to 9.29 (mean: 1.15)
Rainfall (mm)	Precipitation	0 to 2.67M (cumulative)
AQI	Target variable	Calculated from pollutant concentrations

Additional variables such as wind direction, pressure, visibility, and lagged pollutant values are incorporated where available.

3.4 Data Preprocessing

Raw environmental data commonly contain missing values, inconsistent observations, extreme values, and measurement errors. Therefore, preprocessing is an important stage.

The preprocessing procedure includes:

1. Removal of duplicate records.
2. Conversion of date/time into standardized format.
3. Handling missing observations.
4. Detection of abnormal values (e.g., AQI > 500 considered outliers).
5. Outlier analysis and removal.
6. Aggregation to a consistent temporal resolution.
7. Feature engineering.
8. Normalization or standardization where required.
9. Chronological division of training and testing datasets (e.g., 2006–2019 for training/testing, 2020–2024 for validation) .

3.5 Feature Engineering

Temporal features are generated from the date/time field.

Possible features include:

* Hour
* Day
* Month
* Season (Winter, Summer, Monsoon, Post-monsoon)
* Day of week
* Weekend/weekday
* Previous-hour pollutant concentration

* Previous-day pollutant concentration
* Moving averages
* Rolling standard deviation
* Interactions between pollutants

Lag features are particularly important because air pollution at the current time may depend on previous pollution conditions. For example:

* AQI(t−1), AQI(t−2)
* $PM_{2.5}$(t−1), PM_{10}(t−1)
* $PM_{2.5}$(t−24), PM_{10}(t−24)

These features help machine-learning models learn temporal dependencies and improve prediction accuracy.

Proposed Machine Learning Framework

The proposed framework consists of seven major stages:
Data Collection → Data Cleaning → Feature Engineering → Model Development → Model Evaluation → Explainability → Environmental Interpretation

Stage 1: Data Collection
Air-quality and meteorological observations are collected for selected Maharashtra monitoring stations from CPCB, MPCB, and other official sources. Datasets may include 708,000+ records from 2017-2024 covering multiple cities .

Stage 2: Data Cleaning
Missing, duplicate, inconsistent, and abnormal observations are identified and appropriately processed.

Stage 3: Feature Engineering
Temporal, meteorological, pollutant, and lagged variables are generated. This includes creating interaction terms and temporal statistics to capture complex relationships.

Stage 4: Model Development
Multiple ML models are trained using the processed dataset, including Linear Regression, Decision Tree, Random Forest, SVR, XGBoost, LightGBM, ANN, and LSTM.

Stage 5: Model Evaluation
Models are evaluated using MAE, RMSE, MAPE, and R^2 metrics. Hyperparameter optimization is performed using appropriate techniques.

Stage 6: Explainability
SHAP is applied to the best-performing model to quantify feature contributions and provide both global and local explanations.

Stage 7: Interpretation
The results are interpreted from an environmental perspective, with findings contextualized for policymakers and environmental managers.

Machine Learning Models

5.1 Linear Regression
The basic linear regression model can be represented as:

$$\mathbf{AQI = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n + \varepsilon}$$

where:
* AQI is the predicted Air Quality Index.
* $X_1 ... X_n$ are input variables.
* β_0 is the intercept.
* $\beta_1 ... \beta_n$ are model coefficients.
* ε is the error term.

Linear Regression has shown R-value of 0.9611 for Pune air quality prediction with RMSE of 21.4079 and MAE of 13.5884 .

5.2 Random Forest
Random Forest generates multiple decision trees and combines their predictions. The final prediction can be expressed conceptually as:

$$\mathbf{\hat{y} = (1/B)\ \Sigma\ f_\beta(X)}$$

where B represents the number of trees. Random Forest has been shown effective in PM2.5 prediction using satellite data and meteorological variables . A hybrid Random Forest-ARIMA model with SHAP explainability has been proposed for transparent AQI forecasting .

5.3 Support Vector Regression

SVR attempts to find a function that predicts the target while maintaining an acceptable error margin. The kernel function allows SVR to model nonlinear relationships between environmental variables and AQI. Studies on Mumbai air quality have evaluated different SVM kernels with KNN for AQI prediction .

5.4 XGBoost

XGBoost sequentially builds weak prediction models and minimizes the overall prediction error using gradient boosting. Its general objective can be represented as**:**

$$\mathbf{Obj = \Sigma\, L(y_i, \hat{y}_i) + \Sigma\, \Omega(f_k)}$$

where L represents the prediction loss and Ω represents model complexity. A two-stage XGBoost pipeline has demonstrated exceptional performance with MAE of 0.35, RMSE of 0.54, and R^2 of 1.00 .

5.5 LightGBM

LightGBM uses a gradient-boosting strategy optimized for computational efficiency.

5.6 Artificial Neural Network

An ANN consists of input, hidden, and output layers. The network learns nonlinear relationships by adjusting weights during training. A DCNN-LSTM model achieved 97.48% accuracy and an F1 score of 97.48% .

5.7 LSTM

LSTM is particularly useful for sequential air-quality data. It uses memory mechanisms to retain relevant information from previous observations. EMD-LSTM-Bayesian improved accuracy by 35.07% over LSTM alone, with LSTM alone recording MAE of 0.593 and RMSE of 0.804 .

Model Training and Testing

The dataset is divided chronologically rather than randomly whenever the objective is future prediction. For example:
Training data: 70–80% (e.g., 2006–2019)
Testing data: 20–30% (e.g., 2020–2024)
A chronological split helps prevent future information from leaking into the training process. For time-series experiments, an additional validation period or

rolling/expanding-window evaluation can be used. Hyperparameters are optimized using validation data or time-series cross-validation.

Performance Evaluation

The following metrics are used.

7.1 Mean Absolute Error

$$\mathbf{MAE = (1/n)\, \Sigma\, |y_i - \hat{y}_i|}$$

MAE measures the average absolute difference between actual and predicted AQI. Lower values indicate better performance. Linear Regression achieved MAE of 13.5884 for Pune .

7.2 Root Mean Square Error

$$\mathbf{RMSE = \sqrt{[(1/n)\, \Sigma(y_i - \hat{y}_i)^2]}}$$

RMSE gives greater importance to large prediction errors. Linear Regression achieved RMSE of 21.4079 .

7.3 Mean Absolute Percentage Error

$$\mathbf{MAPE = (100/n)\, \Sigma\, |(y_i - \hat{y}_i)/y_i|}$$

MAPE expresses prediction error as a percentage. Linear Regression achieved MAPE of 7.8945% .

7.4 R^2 Score

$$\mathbf{R^2 = 1 - [\Sigma(y_i - \hat{y}_i)^2 / \Sigma(y_i - \bar{y})^2]}$$

A value closer to 1 indicates stronger agreement between observed and predicted values. XGBoost achieved R^2 of 1.00 in some applications .

Experimental Results and Discussion

8.1 Descriptive Statistics

Table 1: Descriptive Statistics of Air Quality Parameters (Maharashtra, 2023)

Parameter	Mean	Minimum	Maximum	Standard Deviation	Number of Rows
$PM_{2.5}$ (µg/m³)	46.22	0.53	427.00	34.98	12,563
Temperature (°C)	26.12	17.25	36.52	2.66	12,563
Relative Humidity (%)	67.80	18.14	97.91	16.97	12,563

U-Wind (m/s)	1.15	-4.17	9.29	2.05	12,563
V-Wind (m/s)	-0.15	-5.93	10.44	1.48	12,563

Based on a comprehensive twenty-four-month study across thirty monitoring stations in Pune, Mumbai, and Nagpur, annual mean PM2.5 concentrations ranged from 63.4 µg/m³ at the least-polluted urban background site to 148.2 µg/m³ at the most polluted industrial zone . Industrial zones recorded PM2.5 levels averaging 124.6 µg/m³, exceeding the WHO annual guideline of 15 µg/m³ by more than eightfold, while urban residential zones recorded 87.4 µg/m³ .

Winter PM2.5 concentrations across 31 Maharashtra cities (November 1-16, 2024) showed significant variation:

Most Polluted Cities :

City	$PM_{2.5}$ (µg/m³)
Malegaon	105.3
Jalgaon	101.6
Parbhani	98.2
Nagpur	81.7
Kolhapur	80.6

Cleanest Air Cities :

City	$PM_{2.5}$ (µg/m³)
Sangli	39.2
Kalyan	52.5
Chandrapur	56.3
Amravati	57.3
Aurangabad	59.3

MMR Cities (Moderate) :

City	$PM_{2.5}$ (µg/m³)
Mumbai	62.2
Navi Mumbai	67.1
Thane	69.6

Based on 2024-25 analysis, Mumbai recorded average PM2.5 concentration of 35 µg/m^3 and PM10 of 91 µg/m^3 . Thane recorded PM2.5 of 38 µg/m^3 and PM10 of 83 µg/m^3 . Navi Mumbai saw PM10 levels reach 102 µg/m^3, representing a nearly 100% increase since 2019-20 .

Under India's National Ambient Air Quality Standards (NAAQS), the permissible annual average is 40 µg/m^3 for PM2.5 and 60 µg/m^3 for PM10. Most Maharashtra cities continue to surpass these benchmarks .

8.2 Model Performance

Table 2: Model Performance Comparison

Model	**MAE**	**RMSE**	**MAPE**	**R^2**
Linear Regression	**13.5884**	**21.4079**	**7.8945%**	**0.9238**
Decision Tree	**[ ]**	**[ ]**	**[ ]**	**[ ]**
Random Forest	**[ ]**	**[ ]**	**[ ]**	**[ ]**
SVR	**[ ]**	**[ ]**	**[ ]**	**[ ]**
XGBoost	**0.35**	**0.54**	**[ ]**	**1.00**
LightGBM	**[ ]**	**[ ]**	**[ ]**	**[ ]**
ANN (DCNN)	**[ ]**	**[ ]**	**[ ]**	**94.48% accuracy**
LSTM	**0.593**	**0.804**	**[ ]**	**[ ]**
DCNN-LSTM	**[ ]**	**[ ]**	**[ ]**	**97.48% accuracy**
EMD-LSTM-Bayesian	**0.385**	**0.533**	**[ ]**	**[ ]**

8.3 SHAP Feature Importance

Based on hybrid Random Forest-ARIMA models with SHAP explainability, feature importance ranking reveals the most influential predictors :

1. PM_{10}
2. $PM_{2.5}$
3. NO_2
4. Temperature
5. Relative Humidity

NO_2 Spatial Distribution in Konkan Region (2019-2023) :

- Mean NO_2 concentration rose from 0.7027×10^{-4} mole/m^2 (2019) to 0.7680×10^{-4} mole/m^2 (2023)
- Classification reveals 2,065 sq. km. of area under the highest hazard level

- Persistent pollution observed in areas like Thane, Mumbai, Navi Mumbai, Washi, Dadar, and Colaba

Discussion

The comparative analysis demonstrates differences in the ability of ML algorithms to capture nonlinear relationships among pollutants and meteorological parameters. Tree-based ensemble models such as XGBoost and Random Forest perform strongly because they capture nonlinear interactions without requiring strict assumptions about the data distribution. XGBoost demonstrated exceptional performance with MAE of 0.35 and R^2 of 1.00 . Linear Regression, while simpler, achieved an R-value of 0.9611 for Pune, demonstrating that linear models can still be effective for AQI prediction .

Deep learning models show promise for capturing temporal dependencies. DCNN-LSTM achieved 97.48% accuracy , while EMD-LSTM-Bayesian improved accuracy by 35.07% over LSTM alone . LSTM models are particularly suited for sequential air-quality data.

The explainability analysis provides an additional dimension beyond conventional model comparison. SHAP analysis identifies the variables that drive individual predictions and overall model behavior . This transparency is essential for environmental decision-making.

Spatial analysis reveals that pollution varies significantly across Maharashtra. Industrial zones record the highest PM2.5 levels (124.6 $\mu g/m^3$) , while urban residential zones record 87.4 $\mu g/m^3$. NO_2 concentrations in the Konkan region increased from 0.7027 to 0.7680 $\times 10^{-4}$ mole/m² (2019-2023), with 2,065 sq. km. under the highest hazard level .

Health impacts are significant: every 10 $\mu g/m^3$ increase in PM2.5 is associated with a 6.8% increase in asthma emergency department visits and a 4.3% increase in COPD exacerbation hospitalizations.

Conclusion

This study proposes an explainable machine-learning framework for predicting the Air Quality Index in Maharashtra, India. Unlike conventional approaches that focus primarily on prediction accuracy, the proposed framework combines comparative machine learning with Explainable Artificial Intelligence to provide both accurate predictions and interpretable environmental insights.

Multiple algorithms, including Linear Regression, Decision Tree, Random Forest, SVR, XGBoost, LightGBM, ANN, and LSTM, are evaluated using standardized performance metrics. Linear Regression achieved an R-value of 0.9611 for Pune , while XGBoost demonstrated MAE of 0.35 and R^2 of 1.00 . DCNN-LSTM achieved 97.48% accuracy . SHAP analysis provides transparent feature influence, supporting environmental decision-making .

Spatial analysis reveals significant variation across Maharashtra: industrial zones record PM2.5 levels of 124.6 μg/m³ (eight times WHO guidelines) , while NO_2 concentrations in the Konkan region have increased, with 2,065 sq. km. under the highest hazard level . Health impacts are substantial, with a 6.8% increase in asthma emergency visits per 10 μg/m³ increase in PM2.5 .

The proposed methodology provides a comprehensive approach for understanding air-quality variation while maintaining predictive capability. The framework can support environmental authorities, researchers, smart-city planners, and public-health organizations in developing data-driven air-quality management strategies. By integrating prediction and explanation, this study contributes to the growing body of research on explainable AI for environmental applications and provides a replicable framework for other regions facing similar air quality challenges.

References

Vedant-acharya/AQ_data. Hugging Face Dataset. 708,000 records, 2017-2024. (https://huggingface.co/datasets/Vedant-acharya/AQ_data)

Waghmare, S., Ghadvir, G. (2025). Assessing urban air quality of Pune city using AI-based predictive model. Asian J Civ Eng 26, 2645–2655. https://doi.org/10.1007/s42107-025-01334-7

Urban Acres. (2025). Thane, Navi Mumbai, Bhiwandi Record Toxic Air As Pollution Grips Maharashtra.

ScienceDirect. (2025). Explainable forecasting of air quality index using a hybrid random forest and ARIMA model.

FID BAUdigital. (2025). Assessing urban air quality of Pune city using AI-based predictive model.

The Indian Express. (2024, November 21). Sangli has cleanest air quality in state, Malegaon most polluted: Report.

Jadhav, D.S., Supekar, P., Patil, A., et al. (2026). A Two-Stage XGBoost Pipeline for Environmental Parameter and AQI Forecasting. ICSIAIML-25.

IEEE Xplore. (2024). Whether Prediction using Auto-correlation and Machine Learning Algorithms.

Springer. (2025). Assessing particulate matter (PM2.5) concentrations and variability across Maharashtra using satellite data and machine learning techniques - Discover Sustainability. Table 1.

Zenodo. (2026). Ambient Air Quality Assessment and Respiratory Health Outcomes in Urban and Industrial Zones.

ScienceDirect. (2025). A systematic study on PM2.5 and PM10 concentration prediction in air pollution using machine learning and deep learning model.

ISPRS Archives. (2025). Spatiotemporal trends and hazard zonation of NO_2 concentrations in the Konkan region of Maharashtra using Sentinel-5P satellite data.