

Comparative Evaluation of Machine Learning Models for Short-Term Air Quality Index Prediction Using Historical Pollutant and Temporal Features in Srinagar, Jammu and Kashmir, India

Mohamad Musadiq¹ Dr. Shaheena Salam²

¹ Department of Computer Science, University of Kashmir

² Islamia College of Science and Commerce

Abstract - Air pollution monitoring and prediction are important for understanding short-term variations in atmospheric conditions and for supporting timely air-quality assessment. Machine learning has increasingly been applied to air-quality prediction because of its ability to capture nonlinear relationships among pollutant concentrations and temporal patterns. Previous studies have investigated statistical, machine-learning, ensemble, and deep-learning approaches for predicting pollutant concentrations and the Air Quality Index (AQI), with AQI and PM_{2.5} among the most frequently modeled targets [1–3].

This study presents a comparative framework for short-term AQI prediction in Srinagar, Jammu and Kashmir, India, using historical pollutant observations and temporal information. The data were obtained from the Rajbagh, Srinagar - JKPCC air-quality monitoring station located at 34.066206° N latitude and 74.81982° E longitude. The initial dataset contained 7,420 hourly observations. Following pollutant selection, missing-value treatment, AQI calculation, removal of incomplete records, and exclusion of invalid AQI observations, 6,262 observations were retained for analysis. Five pollutants, namely PM_{2.5}, PM₁₀, SO₂, CO, and O₃, were considered. AQI values were calculated using a CPCB-oriented pollutant sub-index methodology based on concentration breakpoints and categorized into six classes: Good, Satisfactory, Moderate, Poor, Very Poor, and Severe [4].

Two feature configurations were investigated. Experiment A, the Pollution-History Model, incorporated current pollutant concentrations, pollutant lags, historical AQI values, pollutant changes, and rolling statistical features. Experiment B, the Pollution + Temporal Model, extended this representation by incorporating calendar and cyclical temporal variables. Six classification algorithms and five regression algorithms were comparatively evaluated using five-fold chronological time-series validation.

The Decision Tree achieved 96.61% accuracy and a weighted F1-score of 95.61% in Experiment B, while Bagging achieved 96.55% accuracy and a weighted F1-score of 95.55%. The persistence classification baseline achieved 89.50% accuracy. For numerical AQI prediction, the persistence baseline produced an MAE of 14.69, RMSE of 51.12, and $R^2 = 0.4581$, compared with R^2 values ranging from 0.3070 to 0.3388 for the evaluated machine-learning regression models. Feature-importance analysis consistently identified PM_{2.5} as the dominant predictor, followed by PM₁₀ and pollutant-history features. The findings demonstrate the predictive value of historical pollutant information for short-term AQI-category prediction in Srinagar and highlight the importance of benchmarking machine-learning models against simple persistence-based forecasting approaches.

Key Words: Air Quality Index, Air Pollution, Machine Learning, AQI Prediction, PM_{2.5}, Time-Series Prediction, Random Forest, Decision Tree, Bagging, Air Quality Forecasting

1.INTRODUCTION

Air pollution is an important environmental concern associated with urban activity, transportation, industrial processes, energy consumption, and other anthropogenic sources. Continuous monitoring of atmospheric pollutants provides an essential basis for assessing existing air-quality conditions; however, monitoring alone does not indicate how air quality may evolve over subsequent time periods. Short-term prediction systems can complement monitoring by providing estimates of future air-quality conditions and supporting timely environmental assessment.

The Srinagar urban area of Jammu and Kashmir, India, provides the geographical setting for the present study. The analysis uses observations collected from the Rajbagh, Srinagar - JKPCC air-quality monitoring station, located at 34.066206° N latitude and 74.81982° E longitude. By focusing on a specific monitoring location within Srinagar, the study investigates the extent to which recent pollutant observations and temporal patterns can be used to predict near-future AQI conditions in the local environment. This location-specific focus also distinguishes the present work from studies based on aggregated observations across multiple Indian cities.

The Air Quality Index provides a simplified representation of ambient air pollution by converting pollutant measurements into an overall numerical index and corresponding qualitative category. In India, the Central Pollution Control Board (CPCB) developed the National Air Quality Index framework using pollutant-specific concentration breakpoints and six broad categories: Good, Satisfactory, Moderately polluted, Poor, Very Poor, and Severe [4]. The CPCB framework derives pollutant-specific sub-indices and uses these to determine the overall AQI. The increasing availability of continuous air-quality monitoring data has created opportunities for machine-learning-based prediction. Machine-learning methods can model nonlinear relationships between pollutant variables and prediction targets without requiring an explicitly specified functional relationship. Méndez et al. reviewed 155 studies on machine-learning-based air-quality forecasting and reported that AQI was among the most frequently predicted air-quality measures, while pollutant variables represented a major class of predictors [1].

Air-quality prediction is inherently a temporal problem because pollutant concentrations observed during one period may be related to concentrations observed during preceding periods. Historical pollutant information can therefore provide useful predictive information for subsequent observations. Liu et al. reviewed conventional machine-learning, deep-learning,

ensemble, hybrid, and spatiotemporal approaches for air-quality time-series forecasting and emphasized the importance of data processing and temporal modeling strategies [2].

Recent research has also emphasized the importance of matching the evaluation methodology to the actual prediction setting. Tang et al. identified feature engineering, class imbalance, realistic validation strategies, and model interpretation as important considerations in machine-learning-based air-quality research [3]. These considerations are particularly relevant for short-term prediction because randomly mixing observations from different points in time may not adequately represent the chronological nature of the forecasting task.

Several studies have investigated AQI prediction and air-pollution modeling in India. Maltare and Vahora evaluated SARIMA, support-vector-machine, and LSTM approaches for AQI prediction in Ahmedabad, demonstrating the application of statistical and machine-learning methods to Indian AQI data [5]. Kumar and Pande investigated machine-learning-based air-pollution prediction across multiple Indian cities and compared several algorithms using air-quality observations and standard performance measures [6]. Natarajan et al. developed an optimized machine-learning framework for AQI prediction across major Indian cities and compared decision-tree regression with methods including KNN, Random Forest, and support-vector regression [7].

More advanced temporal and hybrid methods have also been investigated. Udristioiu et al. developed hybrid approaches for particulate matter and AQI prediction and forecasting using sensor observations [8]. Wang et al. proposed an interval-based ensemble framework for AQI time-series forecasting that incorporated temporal decomposition and ensemble learning [9]. Deep-learning approaches have also received increasing attention in air-quality prediction, with Zhou et al. reviewing deep-learning architectures developed for PM_{2.5} concentration prediction and highlighting the importance of temporal and nonlinear modeling approaches [11]. Gao et al. compared multiple machine-learning algorithms for PM_{2.5} prediction and demonstrated that predictive performance can vary according to the model, prediction task, and forecasting setting [10].

Comparative machine-learning approaches have also been applied to environmental prediction problems in the Srinagar context. Mohd et al. compared multiple data-mining algorithms for rainfall prediction using historical weather observations from Srinagar, including J48, Random Forest, Naive Bayes, Logistic Regression, IBk, and Bagging [12]. Their study demonstrates the applicability of comparative predictive modeling to temporally varying environmental data within the Srinagar region. The present study extends this comparative perspective to air-quality prediction, focusing specifically on pollutant history and temporal information rather than meteorological variables.

Despite the growing body of air-quality prediction research, many studies incorporate meteorological variables, multiple monitoring locations, or advanced deep-learning architectures. The present study addresses a more constrained question: **to what extent can short-term AQI conditions at a specific monitoring station in Srinagar be predicted using historical pollutant observations and temporal features alone?** Rather than relying on a single algorithm, the study comparatively evaluates several classical machine-learning models under two feature configurations. The analysis also considers both categorical AQI prediction and numerical AQI prediction,

providing two complementary perspectives on predictive performance.

The main objectives of the study are therefore:

1. To construct AQI values and AQI categories for the Srinagar monitoring location using a CPCB-oriented pollutant sub-index methodology [4].
2. To develop historical pollutant, AQI-history, trend, rolling-statistical, and temporal features for short-term prediction.
3. To compare multiple machine-learning classification algorithms for AQI-category prediction.
4. To evaluate numerical AQI prediction using machine-learning regression algorithms.
5. To compare machine-learning performance against a persistence baseline.
6. To identify the pollutant and historical features that contribute most strongly to the predictions of tree-based models.

2. RELATED WORK

A. Machine Learning for Air-Quality Prediction

Machine learning has become an established methodology for predicting atmospheric pollution. Méndez et al. conducted a systematic survey covering 155 publications between 2011 and 2021 and found that AQI was among the most commonly predicted air-quality measures, while PM_{2.5} represented one of the most frequently predicted individual pollutants. [1] Their review also identified pollutant variables and weather variables as major predictor groups.

Liu et al. reviewed intelligent modeling strategies for air-quality time-series prediction and categorized approaches into conventional predictors, deep-learning models, ensemble methods, hybrid models, and related optimization approaches. The review highlighted the importance of appropriate data processing and temporal modeling when dealing with air-quality time series. [2]

The application of machine learning to air-quality data has continued to expand. Zhou et al. reviewed 118 studies involving deep-learning architectures for PM_{2.5} concentration prediction and emphasized the growing use of temporal and nonlinear approaches for particulate-matter forecasting. [11] Gao et al. compared decision trees, Random Forest, support-vector models, KNN, neural networks, Gaussian-process regression, and other approaches for PM_{2.5} prediction. Their results illustrated that model performance can vary considerably according to the prediction task and that machine-learning systems may behave differently when predicting future conditions rather than reproducing historical relationships. [10]

B. AQI Prediction in India

Research on Indian air-quality datasets has investigated several combinations of classical machine learning, statistical modeling, and deep learning.

Maltare and Vahora investigated AQI prediction for Ahmedabad and evaluated SARIMA, support-vector-machine models with different kernels, and LSTM models using CPCB-derived data. Their study emphasized the importance of preprocessing and model selection when predicting AQI. [5] Kumar and Pande investigated air-pollution prediction using six years of observations covering 23 Indian cities. Their study used exploratory data analysis, feature selection, resampling,

and multiple machine-learning algorithms to investigate air-quality prediction. [6]

Natarajan et al. investigated AQI prediction across major Indian cities using an optimization-based feature-selection approach together with decision-tree regression. Their comparison included KNN, Random Forest, and support-vector regression, demonstrating the continued use of classical machine-learning algorithms for AQI prediction. [7]

These studies demonstrate the applicability of machine learning to Indian air-quality data, while also showing that model performance depends on the selected predictors, temporal setting, preprocessing strategy, and evaluation framework.

C. Comparative Environmental Prediction Studies

Comparative machine-learning studies have also been applied to environmental prediction problems in the Srinagar context. Mohd et al. compared multiple data-mining algorithms for rainfall prediction using historical weather observations from Srinagar and evaluated algorithms including J48, Random Forest, Naive Bayes, Logistic Regression, IBk, and Bagging. Their study illustrates the usefulness of comparative algorithm evaluation for environmental prediction problems involving temporally varying observations. [12]

The present study adopts a similarly comparative philosophy while focusing on air quality rather than rainfall. The predictor domain is also different, with the present analysis emphasizing pollutant history and temporal variables rather than meteorological measurements.

3. DATA COLLECTION AND PREPROCESSING

A. Study Area and Monitoring Station

The study was conducted using air-quality observations from the Rajbagh, Srinagar - JKPC monitoring station in Srinagar, Jammu and Kashmir, India. The station is located at 34.066206° N latitude and 74.81982° E longitude. The station-level dataset forms the basis of the temporal air-quality analysis presented in this study.

B. Initial Dataset

The original dataset contained 7,420 records and 10 columns. The recorded fields consisted of start and end timestamps together with PM2.5, PM10, NO, NO2, NH3, SO2, CO, and O3 measurements. The initial pollutant data contained substantial missing values, particularly for NO, NO2, and NH3. The raw dataset covered observations from July 2025 through July 2026. The first records contained hourly measurements, with each observation representing a one-hour monitoring interval.

	From Date	To Date	PM2.5	PM10	SO2	CO	OZONE
0	01-07-2025 00:00	01-07-2025 01:00	33.40	69.70	2.69	0.24	NaN
1	01-07-2025 01:00	01-07-2025 02:00	28.28	59.08	2.69	0.27	NaN
2	01-07-2025 02:00	01-07-2025 03:00	33.63	70.25	2.69	0.26	NaN
3	01-07-2025 03:00	01-07-2025 04:00	37.00	77.30	2.68	0.27	NaN
4	01-07-2025 04:00	01-07-2025 05:00	40.55	84.73	2.68	0.23	NaN

Figure 1: Raw hourly air-quality observations showing timestamp, PM2.5, PM10, SO2, CO, and O3.

C. Pollutant Selection

NO, NO2, and NH3 were removed because they contained substantially fewer valid observations compared with the other pollutant variables. The retained pollutant variables were:

- PM2.5

- PM10
- SO2
- CO
- O3

The retained variables form the basis of AQI computation and subsequent machine-learning models.

D. Missing-Value Treatment

The initial missing-value counts for the selected pollutant variables were 178 observations for PM2.5, 206 for PM10, 1,006 for SO2, 67 for CO, and 379 for O3.

Missing values for PM2.5 and PM10 were replaced using their respective arithmetic means. Similarly, missing CO values were filled using the mean CO concentration. SO2 and O3 observations containing missing values were removed because both variables were included in the AQI calculation.

After these preprocessing operations, 6,264 observations remained with complete values for the five selected pollutants. Two further observations subsequently produced an invalid AQI because the pollutant-availability condition required for AQI calculation was not met. These two records were excluded, yielding a final modeling dataset of **6,262 observations**.

E. Descriptive Statistics

The cleaned pollutant measurements exhibited substantial variation and positively skewed distributions. The mean concentrations of PM2.5, PM10, SO2, and CO were approximately 28.43, 61.53, 3.30, and 0.305, respectively. The maximum recorded concentrations reached 839.07 for PM2.5, 847.90 for PM10, 128.69 for SO2, and 2.50 for CO.

These distributions also contained substantial outliers, particularly for PM2.5 and PM10.

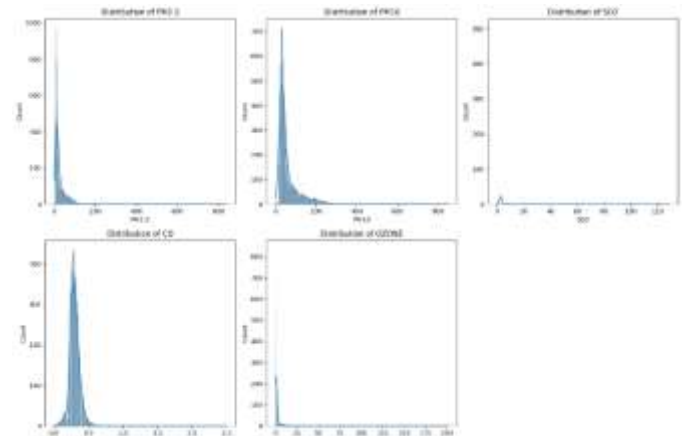


Figure 2: Histograms of PM2.5, PM10, SO2, CO, and O3.

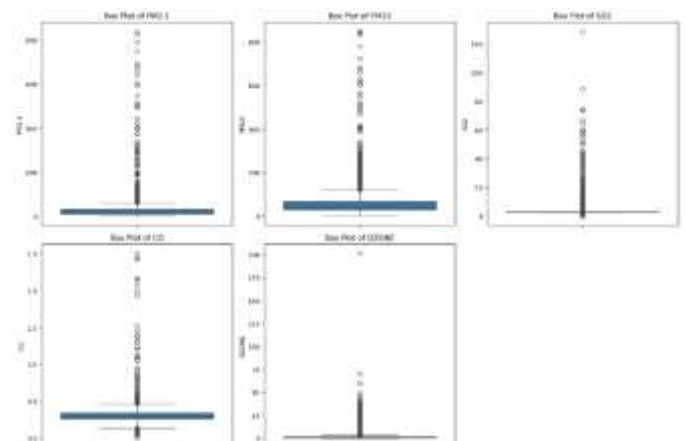


Figure 3 here: Box plots of the five pollutant variables illustrating dispersion and extreme observations.

A correlation analysis was also conducted to examine relationships among the pollutant variables.



Figure 4: Correlation heatmap of PM2.5, PM10, SO2, CO, and O3.

4. AIR QUALITY INDEX CALCULATION

A. CPCB-Oriented AQI Framework

The AQI was calculated using pollutant-specific concentration breakpoints and corresponding AQI ranges. The CPCB framework describes AQI as a transformation of individual pollutant concentrations into pollutant-specific sub-indices, with the overall index representing the worst relevant sub-index. [4]

The five pollutants used in the present implementation were PM2.5, PM10, SO2, CO, and O3. The six AQI categories used in the analysis were:

AQI range	Category	Encoded value
0–50	Good	0
51–100	Satisfactory	1
101–200	Moderate	2
201–300	Poor	3
301–400	Very Poor	4
401 and above	Severe	5

The six-category structure follows the Indian AQI terminology described by CPCB. [4]

B. Sub-Index Calculation

For pollutant p , the sub-index is determined through linear interpolation between the lower and upper concentration breakpoints:

$$I_p = \left(\frac{I_{high} - I_{low}}{B_{high} - B_{low}} \right) (C_p - B_{low}) + I_{low}$$

where:

C_p = measured concentration of pollutant p ,

B_{low} and B_{high} = lower and upper pollutant concentration breakpoints,

I_{low} and I_{high} = corresponding lower and upper AQI index values.

This linear interpolation approach is consistent with the CPCB description of pollutant sub-indices. [4]

When a concentration was below the lower breakpoint, the sub-index was assigned a value of zero. Concentrations above the highest defined breakpoint were capped at an AQI value of 500 in the implemented calculation.

C. Overall AQI

For each observation, the available pollutant sub-indices were calculated. The overall AQI was defined as the maximum pollutant sub-index:

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

provided that at least three pollutants were available and at least one of PM2.5 or PM10 was present. Observations that did not satisfy these criteria were marked invalid and subsequently removed from the final modeling dataset.

The resulting AQI values were mapped to the six categorical levels shown in Table 1.

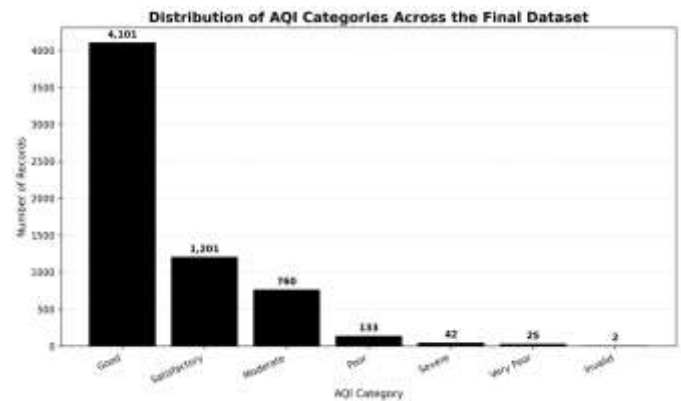


Figure 5: Distribution of AQI categories across the final dataset.

5. TEMPORAL FEATURE EXTRACTION

The timestamps were converted from string representation into datetime objects. From the starting timestamp, the following calendar variables were extracted:

Year, Month, Day, Hour, DayOfWeek

The extracted variables were intended to represent annual, monthly, daily, hourly, and weekly patterns in pollutant behavior.

The observations extended from 24 July 2025 at 00:00 to 1 July 2026 at 23:00 after preprocessing.

6. TIME-SERIES CHARACTERIZATION

Before constructing historical features, the temporal characteristics of the cleaned dataset were investigated.

There were no duplicate timestamps. The most common interval between consecutive observations was one hour, occurring for 6,101 consecutive pairs. However, the dataset also contained longer gaps. Across the complete expected hourly range, 8,232 hourly intervals were possible, while 1,970 intervals were absent. The first identified missing intervals included 10:00, 11:00, 12:00, 15:00, and 16:00 on 24 July 2025. The temporal characteristics demonstrate that the dataset is predominantly hourly but not perfectly continuous.



Figure 6: Distribution of time differences between consecutive observations and visualization of missing intervals.

The observations were sorted chronologically before the creation of historical features.

7. FEATURE ENGINEERING

A. Future AQI Category

For the short-term prediction formulation, a future AQI category was constructed by shifting the AQI-category sequence by one observation:

$$Y_{t+1} = AQI_Category_{t+1}$$

Thus, the target associated with an observation at time t corresponds to the following observation in the temporal sequence.

The final row has no subsequent observation and therefore produces a missing target.

B. Pollutant Lag Features

Historical pollutant values were incorporated using six lag intervals:

$$L = \{1, 2, 3, 6, 12, 24\}$$

For each pollutant X_t , the lagged variables were defined as:

$$X_{t-k} = X_t^{(k)}$$

where $k \in \{1, 2, 3, 6, 12, 24\}$.

The procedure was applied independently to PM2.5, PM10, SO2, CO, and O3. Thus, each pollutant contributed six historical lag variables.

These variables were designed to represent recent pollutant behavior and daily-scale temporal dependence.

C. Historical AQI Features

Historical AQI values were similarly represented through lag variables:

$$AQI_{t-1}, AQI_{t-2}, AQI_{t-3}, AQI_{t-6}, AQI_{t-12}, AQI_{t-24}$$

Historical AQI therefore provides the model with information about the recent overall state of air quality.

D. Pollutant Change Features

For each pollutant, a first-order change variable was calculated:

$$\Delta X_t = X_t - X_{t-1}$$

The resulting variables were:

$$PM2.5_{\Delta}, PM10_{\Delta}, SO2_{\Delta}, CO_{\Delta}, O3_{\Delta}$$

A positive value indicates an increase relative to the previous observation, while a negative value indicates a decrease.

The purpose of these variables was to capture short-term pollutant trends rather than relying only on absolute concentration levels.

E. Rolling Statistical Features

Rolling statistics were calculated for window lengths of 3, 6, 12, and 24 observations.

For pollutant X , the rolling mean was calculated as:

$$\mu_t^{(w)} = \frac{1}{w} \sum_{i=1}^w X_{t-i}$$

and the rolling standard deviation was calculated as:

$$\sigma_t^{(w)} = \sqrt{\frac{1}{w-1} \sum_{i=1}^w (X_{t-i} - \mu_t^{(w)})^2}$$

The rolling statistics were shifted so that the historical-window construction used preceding observations.

The rolling features included:

- 3-hour rolling mean and standard deviation
- 6-hour rolling mean and standard deviation
- 12-hour rolling mean and standard deviation
- 24-hour rolling mean and standard deviation

for each of the five pollutants.

F. Cyclical Temporal Features

Because time variables such as hour and month are cyclical, direct integer encoding may incorrectly represent their relationships. For example, 23:00 and 00:00 are temporally adjacent despite being numerically distant when represented as integers.

Therefore, hour was represented using:

$$Hour_{sin} = \sin\left(\frac{2\pi Hour}{24}\right)$$

$$Hour_{cos} = \cos\left(\frac{2\pi Hour}{24}\right)$$

Similarly, month was represented using:

$$Month_{sin} = \sin\left(\frac{2\pi Month}{12}\right)$$

$$Month_{cos} = \cos\left(\frac{2\pi Month}{12}\right)$$

These transformations provide continuous cyclic representations of periodic time variables.

8. EXPERIMENTAL FEATURE SETS

Two principal feature sets were constructed.

A. Experiment A: Pollution-History Model

Experiment A consisted entirely of pollutant and historical-air-quality information.

The feature groups were:

1. Current pollutant concentrations.
2. Pollutant lag variables.
3. Historical AQI variables.
4. Pollutant-change variables.
5. Rolling means.

6. Rolling standard deviations.

This configuration contained **86 features**.

B. Experiment B: Pollution + Temporal Model

Experiment B contained all features from Experiment A and added nine temporal variables:

Year, Month, Day, Hour, DayOfWeek, Hour_{sin}, Hour_{cos}, Month_{sin}, Month_{cos}

This produced **95 features**.

After removal of observations with missing values introduced by lagging and rolling operations, 25 rows were excluded for the classification feature construction.

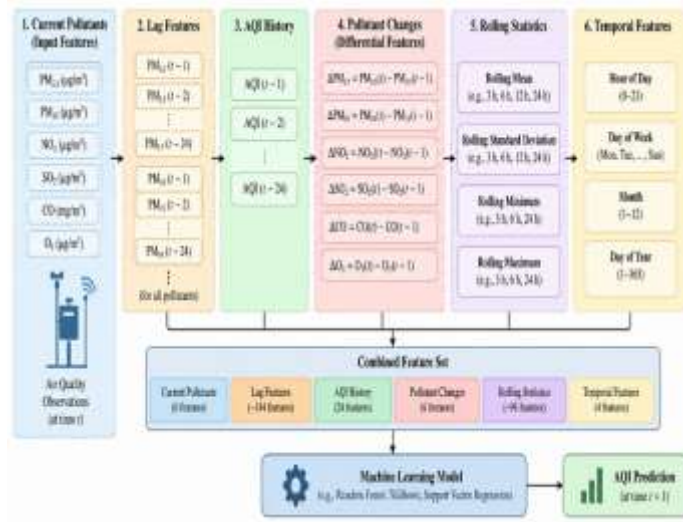


Figure 7: Complete feature-engineering architecture

9. MACHINE-LEARNING ALGORITHMS

Six classification algorithms were evaluated.

A. Random Forest

Random Forest is an ensemble learning method that constructs multiple decision trees using randomized samples and feature subsets. The classification output is obtained through aggregation of the predictions of the individual trees.

A Random Forest prediction can conceptually be represented as:

$$\hat{y} = \text{mode}\{T_1(X), T_2(X) \dots T_B(X)\}$$

where T_b denotes the prediction of tree b and B is the number of trees.

Random Forest is particularly suitable for high-dimensional tabular data because it can model nonlinear relationships and interactions without requiring linear assumptions.

B. Decision Tree

A decision tree recursively divides the feature space into increasingly homogeneous regions. Each internal node represents a feature-based decision and each terminal node corresponds to a predicted class.

Decision-tree models are particularly suitable when the target is related to threshold-based relationships. This is relevant to AQI because the AQI framework itself uses pollutant concentration breakpoints.

The implementation used a decision-tree classifier as the tree-based classification model; it was reported as the J48 decision-tree model in the comparative experimental framework.

C. Naive Bayes

Naive Bayes uses Bayes' theorem:

$$P(C | X) = \frac{P(X | C)P(C)}{P(X)}$$

With the conditional-independence assumption:

$$P(X | C) = \prod_{j=1}^n P(x_j | C)$$

and the predicted class is:

$$\hat{C} = \arg \max_c P(C) \prod_{j=1}^n P(x_j | C)$$

The Gaussian Naive Bayes implementation was used in the experiments.

D. Logistic Regression

Logistic regression estimates class probabilities using the logistic function:

$$P(Y = 1 | X) = \frac{1}{1 + e^{-z}}$$

where:

$$z = \beta_0 + \sum_{j=1}^n \beta_j x_j$$

For multiclass classification, the model extends this probability framework to multiple classes.

E. K-Nearest Neighbors

KNN predicts a sample based on its nearest training observations.

Euclidean distance was used as the underlying distance concept:

$$d(x, x_i) = \sqrt{\sum_{j=1}^n (x_j - x_{ij})^2}$$

The five-nearest-neighbor configuration was used for both classification and regression.

F. Bagging

Bagging, or Bootstrap Aggregating, creates multiple models using bootstrap samples of the training data and aggregates their predictions.

For classification:

$$\hat{y} = \text{mode}\{\hat{y}_1, \hat{y}_2, \dots, \hat{y}_B\}$$

The method can reduce model variance by combining several individual learners.

10. VALIDATION AND PERFORMANCE MEASURES

A. Chronological Validation

Because the data represent a time sequence, chronological validation was adopted for the primary comparative experiment. Five-fold TimeSeriesSplit was used so that the training portion of each fold preceded its corresponding testing portion.

The structure can be represented as:

$$\begin{aligned} \text{Train}_1 &\rightarrow \text{Test}_1 \\ \text{Train}_1 + \text{Train}_2 &\rightarrow \text{Test}_2 \end{aligned}$$

and so forth, with progressively expanding training sets. This approach is consistent with recommendations in the air-quality machine-learning literature that validation should reflect the intended real-world prediction setting rather than arbitrarily mixing temporal observations.

B. Accuracy

Classification accuracy was defined as:

$$\text{Accuracy} = \frac{1}{N} \sum_{i=1}^N I(y_i = \hat{y}_i)$$

where $I(\cdot)$ equals 1 when prediction and actual class match and 0 otherwise.

C. Precision

For a class:

$$\text{Precision} = \frac{TP}{TP + FP}$$

Weighted precision was used to account for the number of observations belonging to each class.

D. Recall

$$\text{Recall} = \frac{TP}{TP + FN}$$

Weighted recall was reported across the AQI classes.

E. F1-Score

The F1-score is the harmonic mean of precision and recall:

$$F1 = 2 \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Weighted F1 was used for the multi-class results.

F. Mean Absolute Error

For either ordinal classification codes or numerical AQI prediction:

$$MAE = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i|$$

For classification, the AQI categories were represented using their ordered codes from 0 to 5, allowing MAE to provide a measure of categorical distance.

G. Root Mean Squared Error

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2}$$

H. Relative Absolute Error

The relative absolute error was calculated as:

$$RAE = \frac{\sum_{i=1}^N |y_i - \hat{y}_i|}{\sum_{i=1}^N |y_i - \bar{y}|}$$

where $\bar{y}$ is the mean value of the observed target.

I. Regression Coefficient of Determination

For numerical AQI prediction:

$$R^2 = 1 - \frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{\sum_{i=1}^N (y_i - \bar{y})^2}$$

An R^2 value closer to 1 indicates that a larger proportion of variance in the observed AQI is explained by the predictions.

11. PERSISTENCE BASELINE

A persistence model was included as a simple benchmark.

For AQI-category prediction:

$$\hat{Y}_{t+1} = Y_t$$

For numerical AQI prediction:

$$\widehat{AQI}_{t+1} = AQI_t$$

The persistence approach is important because AQI often exhibits temporal continuity. A machine-learning model should therefore be evaluated not only according to its absolute performance but also relative to a simple assumption that the current air-quality state will persist into the following observation.

12. EXPERIMENTAL RESULTS

A. AQI Category Classification

The classification results obtained using the historical feature representation and temporal feature representation are presented in Table 2.

Table 2. Classification performance across the two feature configurations

Exp	Model	Accuracy	Precision	Recall	F1	MAE	RMSE	RAE
A	Random Forest	0.9523	0.9380	0.9523	0.9442	0.0699	0.3658	0.0958
A	Decision Tree	0.9627	0.9600	0.9627	0.9527	0.0608	0.3655	0.0834
A	Naive Bayes	0.7892	0.8016	0.7892	0.7854	0.3565	0.9574	0.4887
A	Logistic Regression	0.9192	0.9069	0.9192	0.9091	0.1030	0.4043	0.1412
A	KNN	0.8169	0.8044	0.8169	0.8096	0.2117	0.5402	0.2903
A	Bagging	0.9650	0.9629	0.9650	0.9552	0.0564	0.3424	0.0773
B	Random Forest	0.9578	0.9450	0.9578	0.9483	0.0637	0.3524	0.0873
B	Decision Tree	0.9661	0.9632	0.9661	0.9561	0.0576	0.3631	0.0789
B	Naive Bayes	0.7698	0.7781	0.7698	0.7712	0.2597	0.5796	0.3560
B	Logistic Regression	0.9167	0.9061	0.9167	0.9066	0.1057	0.4081	0.1449
B	KNN	0.7938	0.7934	0.7938	0.7906	0.2373	0.5692	0.3254
B	Bagging	0.9655	0.9626	0.9655	0.9555	0.0579	0.3615	0.0794
BL	Persistence	0.8950	0.8950	0.8950	0.8950	0.1286	0.4509	0.1792

The classification results indicate strong performance from the tree-based and ensemble approaches. In Experiment A, Bagging obtained an accuracy of 96.50% and weighted F1-score of 95.52%, while the Decision Tree obtained an accuracy

of 96.27%. The reported results are based on five chronological validation folds.

When temporal variables were added in Experiment B, the Decision Tree achieved 96.61% accuracy and a weighted F1-score of 95.61%. Bagging achieved 96.55% accuracy and a weighted F1-score of 95.55%, while Random Forest achieved 95.78% accuracy.

The Naive Bayes model produced the lowest classification performance in both feature configurations, with accuracy declining from 78.92% in Experiment A to 76.98% in Experiment B.

The persistence baseline achieved 89.50% accuracy and a weighted F1-score of approximately 89.50%. Thus, the reported tree-based and ensemble classifiers provided higher classification scores than the persistence benchmark.

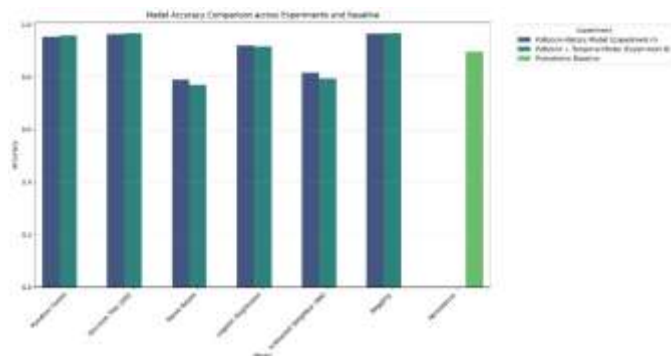


Figure 8: Accuracy comparison of all classification models and persistence baseline.



Figure 9: Weighted F1-score comparison across classification models.

13. CONFUSION-MATRIX ANALYSIS

Confusion matrices were generated for each classification model to examine the distribution of correct and incorrect predictions across AQI categories.

The confusion matrix represents actual categories along one axis and predicted categories along the other. Correct predictions are represented along the principal diagonal, while off-diagonal observations represent misclassifications.

For AQI prediction, the distinction between adjacent categories is particularly relevant. A prediction of Moderate instead of Satisfactory represents a smaller ordinal discrepancy than a prediction of Severe instead of Good.



Figure 10: Confusion matrix of the Decision Tree model for Experiment B.

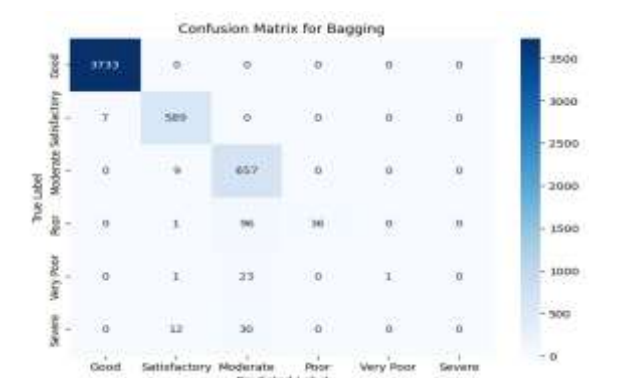


Figure 11: Confusion matrix of the Bagging model for Experiment B.

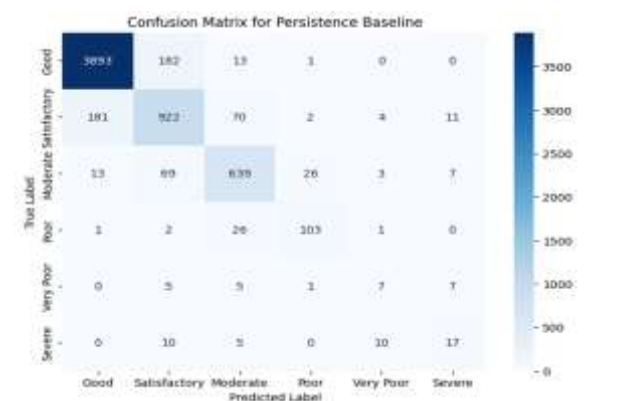


Figure 12: Confusion matrix of the persistence baseline.

The confusion matrices are presented with the six ordered classes:

Good, Satisfactory, Moderate, Poor, Very Poor, Severe

This ordering also corresponds to the numerical encoding:

0,1,2,3,4,5

respectively.

14. NUMERICAL AQI PREDICTION

In addition to categorical prediction, the study evaluated direct numerical prediction of AQI.

The numerical target was defined as the AQI value associated with the subsequent observation:

$$Y_{t+1} = AQI_{t+1}$$

Five regression models were evaluated:

- Random Forest Regressor
- Decision Tree Regressor
- Linear Regression
- K-Nearest Neighbor Regressor
- Bagging Regressor

The persistence baseline was again used for comparison.

Table 3. Numerical AQI prediction results

Exp.	Model	MAE	RMSE	R ²
A	Random Forest Regressor	20.4526	61.5176	0.3332
A	Decision Tree Regressor	21.7590	62.6538	0.3084
A	Linear Regression	55.6345	353.5646	-21.0245
A	KNN Regressor	21.7311	61.2612	0.3388
A	Bagging Regressor	20.9163	62.0424	0.3218
B	Random Forest Regressor	20.4989	61.5670	0.3322
B	Decision Tree Regressor	22.2254	62.7186	0.3070
B	Linear Regression	63.6341	339.2088	-19.2723
B	KNN Regressor	22.8423	61.5211	0.3332
B	Bagging Regressor	20.2986	61.8634	0.3257
BL	Persistence	14.6852	51.1176	0.4581

The numerical prediction results show a different behavior from categorical classification. The persistence baseline produced MAE = 14.6852, RMSE = 51.1176, and R^2 = 0.4581.

Among the machine-learning regression approaches, the highest reported R^2 in Experiment A was 0.3388 for KNN, while the lowest MAE in Experiment A was achieved by Random Forest at 20.4526. In Experiment B, Bagging achieved the lowest MAE among the machine-learning models at 20.2986.

The persistence baseline therefore remained a strong benchmark for direct numerical AQI prediction.

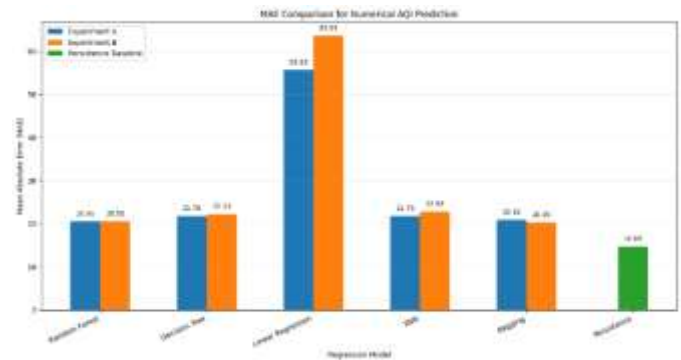


Figure 13: MAE comparison for numerical AQI prediction.

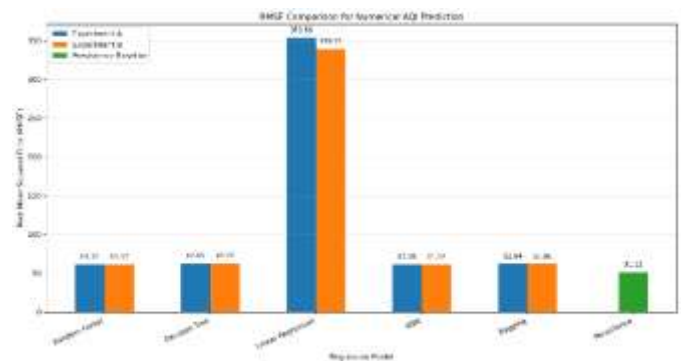


Figure 14: RMSE comparison for numerical AQI prediction.

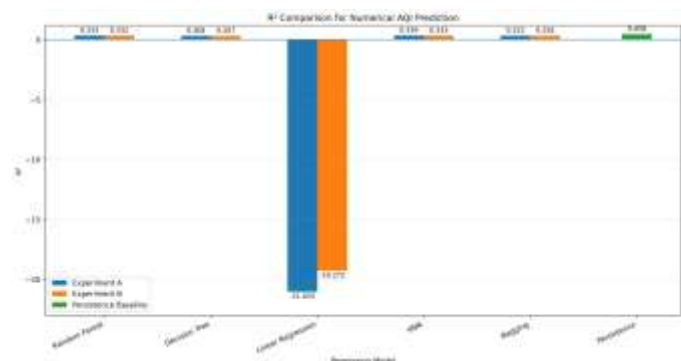


Figure 15: R² comparison for numerical AQI prediction.

15. FEATURE IMPORTANCE ANALYSIS

Feature importance was investigated for Random Forest and Decision Tree regression models to identify the variables contributing most strongly to the predictions.

A. Random Forest, Experiment A

The most important predictors reported for the Random Forest model were:

Feature	Importance
PM2.5	0.557357
PM10	0.046976
PM2.5 change	0.026920

PM2.5 rolling SD, 24 observations	0.025047
PM2.5 lag 1	0.023346
PM2.5 rolling mean, 12 observations	0.021822
PM2.5 rolling SD, 12 observations	0.021334
PM10 change	0.016229
AQI lag 1	0.015630
PM2.5 rolling mean, 24 observations	0.012362

B. Decision Tree, Experiment A

For the Decision Tree regression model, the largest importance contributions were:

Feature	Importance
PM2.5	0.617704
PM10	0.041715
PM2.5 change	0.041661
PM2.5 rolling SD, 12 observations	0.032416
SO2 rolling SD, 6 observations	0.031551
PM2.5 rolling mean, 24 observations	0.025579
PM10 lag 2	0.022294
O3 rolling mean, 12 observations	0.020711
AQI lag 1	0.018199
CO rolling SD, 24 observations	0.013157

PM2.5 again constituted the dominant feature in the model.

C. Experiment B

Adding temporal features produced a broadly similar importance pattern. For Random Forest, PM2.5 had an importance of 0.556130 and PM10 0.047888. PM2.5 change, rolling statistics, PM2.5 lag variables, PM10 change, and historical AQI variables followed.

For the Decision Tree model, PM2.5 accounted for 0.618056 of feature importance, followed by PM10 at 0.041508 and PM2.5 change at 0.041467. The temporal feature Hour_cos appeared among the top reported features with an importance value of 0.008933.

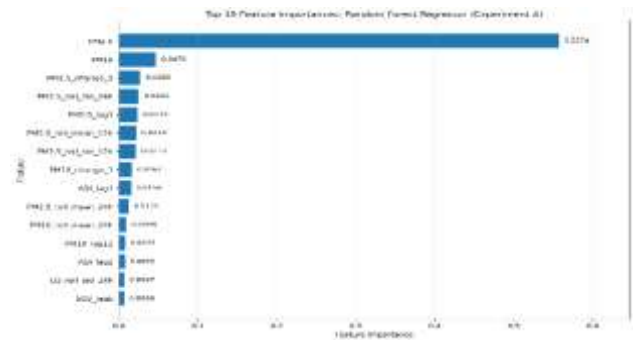


Figure 16: Top 15 feature importances for the Random Forest regression model in Experiment A.

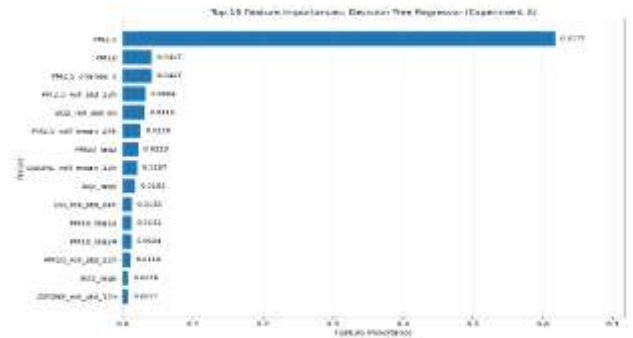


Figure 17: Top 15 feature importances for the Decision Tree regression model in Experiment A.

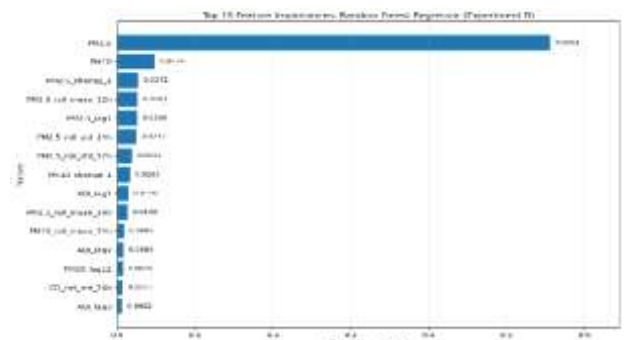


Figure 18: Top 15 feature importances for the Random Forest regression model in Experiment B.

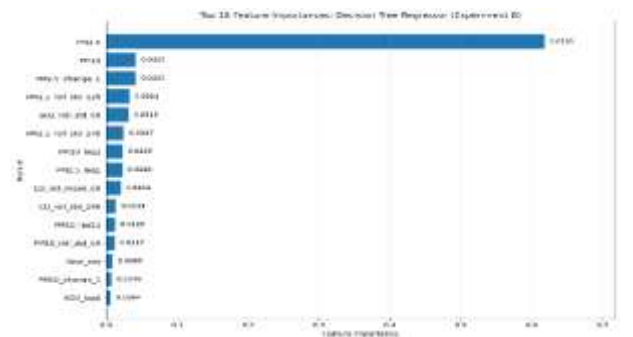


Figure 19: Top 15 feature importances for the Decision Tree regression model in Experiment B.

16. DISCUSSION

The experimental findings demonstrate distinct behavior between categorical AQI prediction and numerical AQI prediction.

For categorical prediction, the tree-based and ensemble algorithms exhibited strong predictive performance. Decision Tree, Random Forest, and Bagging achieved accuracies above 95% across both feature configurations. This performance is consistent with the ability of tree-based models to represent nonlinear threshold relationships between pollutant concentrations and AQI categories.

The strong performance of Decision Tree is particularly relevant to AQI categorization because the AQI framework itself is based on pollutant concentration breakpoints. The tree model can therefore construct hierarchical decision boundaries that are naturally suited to threshold-oriented target definitions.

Bagging produced similarly high results, with 96.50% accuracy in Experiment A and 96.55% in Experiment B. This indicates that ensemble aggregation provided strong classification performance while maintaining robustness across the two feature configurations.

Random Forest also performed strongly, achieving 95.23% accuracy using the pollution-history representation and 95.78% after temporal features were added.

The addition of temporal variables did not produce a universal improvement across every algorithm. For example, Random Forest and Decision Tree improved their classification accuracy in Experiment B, whereas Naive Bayes and KNN showed lower accuracy. This indicates that the value of additional features depends partly on the underlying assumptions of the learning algorithm.

The comparatively lower performance of Naive Bayes can potentially be interpreted in relation to its conditional-independence assumption. Air pollutants can exhibit strong relationships with one another, particularly among particulate-matter variables. Therefore, treating predictor variables as conditionally independent may limit the suitability of the model for this feature space.

The regression results provide a different perspective. The persistence baseline achieved a lower MAE and RMSE and a higher R^2 than each evaluated regression model. This indicates that the AQI sequence exhibits considerable short-term temporal continuity. In such settings, a model predicting that the future AQI will remain close to the current AQI can be difficult to outperform. This result is important because previous air-quality studies often focus primarily on the performance of complex machine-learning algorithms. A persistence benchmark establishes whether the additional complexity of a machine-learning system provides predictive information beyond the temporal inertia of the target series. The difference between classification and regression performance also illustrates an important distinction. Categorical prediction does not require a model to estimate the precise numerical AQI. An observation with an actual AQI of 120 and a prediction of 150 would receive no classification

penalty if both fall within the same category. Numerical regression, in contrast, penalizes the magnitude of such an error directly.

Feature importance provides additional evidence that particulate matter was central to the predictive structure of the dataset. PM_{2.5} consistently dominated the feature-importance rankings, accounting for approximately 55.6% of Random Forest importance and approximately 61.8% of Decision Tree importance in the reported experiments. PM₁₀ was generally the second most influential feature. [1] [11]

The importance of PM_{2.5} is also consistent with the broader air-quality literature, where PM_{2.5} is one of the most frequently studied and predicted pollutants. The presence of rolling and lagged PM_{2.5} variables among the important predictors further indicates that both current pollutant state and recent historical behavior contributed to the model representations.

The results also demonstrate the usefulness of comparing alternative feature configurations. The Pollution-History Model provides a representation based primarily on pollutant dynamics, while the Pollution + Temporal Model provides additional information about recurring temporal structure. This distinction allows the incremental contribution of explicit temporal variables to be studied without introducing meteorological variables into the predictor space.

17. COMPARISON WITH PREVIOUS STUDIES

The present study differs from several existing AQI prediction studies in terms of predictor composition, model selection, and experimental objective.

Maltare and Vahora investigated Ahmedabad AQI prediction using SARIMA, SVM, and LSTM, focusing on model selection for an Indian urban setting. [5]

Natarajan et al. investigated AQI prediction across major Indian cities using Grey Wolf Optimization and decision-tree regression, comparing the resulting approach with KNN, Random Forest, and support-vector regression. [7]

Udristoiu et al. used hybrid methodologies for particulate matter and AQI prediction and forecasting in Romania, demonstrating the potential of combining feature-selection and machine-learning approaches. [8]

Wang et al. proposed an interval-valued ensemble framework specifically for AQI time-series forecasting, using temporal decomposition and ensemble learning to capture changing AQI behavior. [9]

Gao et al. compared several machine-learning algorithms for PM_{2.5} prediction and found that predictive behavior varies according to model architecture and prediction task, reinforcing the importance of comparative evaluation rather than relying on a single algorithm. [10]

Compared with these approaches, the present work deliberately concentrates on classical machine-learning models and historical pollutant information. Its principal purpose is comparative evaluation rather than development of a highly optimized or deep-learning architecture.

The study therefore provides a relatively interpretable framework in which the contribution of pollutant history, AQI

history, pollutant trends, rolling statistics, and basic temporal variables can be examined independently.

18. LIMITATIONS

Several limitations should be considered when interpreting the results.

First, although the observations are predominantly hourly, the time series is not perfectly continuous. Of the 8,232 expected hourly intervals between the earliest and latest observations, 1,970 intervals were missing. This means that temporal representation and forecasting horizon should be interpreted in terms of successive observations where appropriate.

Second, the study uses only five pollutant variables. Other pollutants available in broader AQI frameworks, including NO₂ and NH₃, were excluded because their observed data were substantially incomplete.

Third, meteorological variables were not incorporated. Previous air-quality studies have frequently used variables such as temperature, humidity, wind speed, and atmospheric pressure in addition to pollutant concentrations. The present analysis therefore specifically investigates the predictive information contained in pollutant and temporal variables rather than attempting to construct a complete atmospheric forecasting system. [1]

Fourth, some AQI classes are much less frequent than others. The final distribution consisted of 4,101 Good observations, 1,201 Satisfactory observations, 760 Moderate observations, 133 Poor observations, 25 Very Poor observations, and 42 Severe observations. This class imbalance should be considered when interpreting aggregate accuracy and weighted metrics.

Finally, the reported study evaluates a relatively limited temporal period. Longer multi-year datasets would allow more robust assessment of seasonal variability, extreme-pollution events, and year-to-year changes.

19. FUTURE WORK

Several directions can extend the proposed framework.

First, future research can investigate additional forecasting horizons, such as 3-hour, 6-hour, 12-hour, and 24-hour AQI prediction. Such experiments would provide a more comprehensive characterization of how predictive performance decreases as the forecast horizon increases.

Second, future studies can incorporate meteorological information, including temperature, relative humidity, wind speed, atmospheric pressure, and precipitation. This would permit comparison between pollution-only, pollution-plus-temporal, and integrated environmental models.

Third, more advanced machine-learning and deep-learning methods can be investigated. Candidate approaches include XGBoost, LightGBM, Support Vector Regression, LSTM, GRU, temporal convolutional networks, and attention-based architectures.

Fourth, future work can investigate explainable artificial intelligence methods such as SHAP to provide both global and observation-specific explanations of AQI predictions.

Fifth, future research should investigate longer continuous datasets and multiple monitoring stations to determine whether the learned relationships generalize across locations.

Finally, future studies can develop a real-time prediction framework in which new pollutant observations are automatically incorporated into the model and near-future AQI predictions are generated continuously.

20. CONCLUSION

This study presented a comparative machine-learning framework for Air Quality Index prediction based on historical pollutant and temporal information. An initial dataset of 7,420 observations was subjected to pollutant selection, missing-value treatment, AQI calculation, and invalid-record removal, resulting in 6,262 observations for the primary analysis.

Five pollutants, PM_{2.5}, PM₁₀, SO₂, CO, and O₃, were used as the core environmental predictors. AQI was calculated using pollutant-specific sub-indices and a maximum-sub-index framework, followed by classification into six AQI categories. Two feature configurations were evaluated. The Pollution-History Model contained 86 predictors representing current pollutant conditions, pollutant history, historical AQI, pollutant changes, and rolling statistics. The Pollution + Temporal Model expanded this representation to 95 predictors by incorporating calendar and cyclical temporal variables.

Six classification algorithms and five regression algorithms were evaluated using chronological time-series validation. The classification results demonstrated particularly strong performance from tree-based and ensemble methods. The Decision Tree produced a reported accuracy of 96.61% and weighted F1-score of 95.61% in Experiment B, while Bagging produced 96.55% accuracy and 95.55% weighted F1. The persistence classification baseline achieved 89.50% accuracy. For direct numerical AQI prediction, the persistence baseline achieved MAE = 14.6852, RMSE = 51.1176, and $R^2 = 0.4581$. The machine-learning regression models produced lower R^2 values, ranging from approximately 0.3070 to 0.3388. These results demonstrate the importance of including a simple persistence benchmark when evaluating short-term AQI prediction systems.

Feature-importance analysis consistently showed that PM_{2.5} was the most influential predictor in the reported tree-based regression models, followed by PM₁₀ and features representing recent pollutant changes, historical concentrations, rolling statistics, and historical AQI.

The study demonstrates that historical pollutant information provides substantial predictive information for AQI-category modeling and that classical machine-learning approaches can provide strong categorical predictions. At the same time, the results show that numerical AQI prediction is more difficult and that temporal persistence represents a strong benchmark. The findings provide a foundation for subsequent research involving longer observation periods, additional environmental variables, advanced time-series models, and explainable artificial intelligence.

ACKNOWLEDGEMENT

The author gratefully acknowledges the guidance and support received throughout this research. Special thanks are extended to the University of Kashmir for providing the academic environment and resources necessary to conduct this study. The author also acknowledges the Central Pollution Control Board (CPCB) for its air-quality standards and AQI framework, which formed the basis of the analysis.

REFERENCES

1. 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.
2. H. Liu, G. Yan, Z. Duan, and C. Chen, "Intelligent modeling strategies for forecasting air quality time series: A review," *Applied Soft Computing*, vol. 102, p. 106957, 2021. doi: 10.1016/j.asoc.2020.106957.
3. D. Tang, Y. Zhan, and F. Yang, "A review of machine learning for modeling air quality: Overlooked but important issues," *Atmospheric Research*, vol. 300, p. 107261, 2024. doi: 10.1016/j.atmosres.2024.107261.
4. Central Pollution Control Board, Government of India, *National Air Quality Index*, CPCB, New Delhi
5. N. N. Maltare and S. Vahora, "Air Quality Index prediction using machine learning for Ahmedabad city," *Digital Chemical Engineering*, vol. 7, p. 100093, 2023. doi: 10.1016/j.dche.2023.100093.
6. K. Kumar and B. P. Pande, "Air pollution prediction with machine learning: a case study of Indian cities," *International Journal of Environmental Science and Technology*, vol. 20, no. 5, pp. 5333–5348, 2023. doi: 10.1007/s13762-022-04241-5.
7. S. K. Natarajan, P. Shanmurthy, D. Arockiam, B. Balusamy, and S. Selvarajan, "Optimized machine learning model for air quality index prediction in major cities in India," *Scientific Reports*, vol. 14, p. 6795, 2024. doi: 10.1038/s41598-024-54807-1.
8. M. T. Udristoiu, Y. El Mghouchi, and H. Yildizhan, "Prediction, modelling, and forecasting of PM and AQI using hybrid machine learning," *Journal of Cleaner Production*, vol. 421, p. 138496, 2023. doi: 10.1016/j.jclepro.2023.138496.
9. Z. Wang, R. Gao, P. Wang, and H. Chen, "A new perspective on air quality index time series forecasting: A ternary interval decomposition ensemble learning paradigm," *Technological Forecasting and Social Change*, vol. 191, p. 122504, 2023. doi: 10.1016/j.techfore.2023.122504.
10. Z. Gao, K. Do, Z. Li, X. Jiang, K. J. Maji, C. E. Ivey, and A. G. Russell, "Predicting PM_{2.5} levels and exceedance days using machine learning methods," *Atmospheric Environment*, vol. 323, p. 120396, 2024. doi: 10.1016/j.atmosenv.2024.120396.
11. S. Zhou, W. Wang, L. Zhu, Q. Qiao, and Y. Kang, "Deep-learning architecture for PM_{2.5} concentration prediction: A review," *Environmental Science and Ecotechnology*, vol. 21, p. 100400, 2024. doi: 10.1016/j.eso.2024.100400.
12. R. Mohd, M. A. Butt, and M. Z. Baba, "Comparative Study of Rainfall Prediction Modeling Techniques (A Case Study on Srinagar, J&K, India)," *Asian Journal of Computer Science and Technology*, vol. 7, no. 3, pp. 13–19, 2018. doi: 10.51983/ajcst-2018.7.3.1901.