Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Catalyzing Air Quality Index Prediction via Machine Learning: Unraveling the Influence of Urban Land Use

Nusrat Ullah Hasani

nhasani92@gmail.com

Integral University

Faiyaz Ahmad

Integral University

Research Article

Keywords: Air Quality Index, Machine Learning, Ozone Prediction, Random Forest, Linear Regression

Posted Date: May 15th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-7819940/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Machine Learning Based Ozone Sub-Index Estimation and Composite AQI Forecasting under the CPCB NAQI Framework Using Sentinel-5P

Nusrat Ullah Hasani[1*] and Faiyaz Ahmad[1†]

[1*]Department of Computer Science and Engineering , Integral University, Lucknow, Lucknow, 226022, Uttar Pradesh, India.

*Corresponding author(s). E-mail(s): nhasani92@gmail.com;
Contributing authors: faiyaz@iul.ac.in;
†These authors contributed equally to this work.

Abstract

The study develops a CPCB-compliant, two framework for modeling the Air Quality Index (AQI) under the National Air Quality Index (NAQI) system in India, with a clear distinction between the deterministic ozone sub-index and the multi-pollutant AQI forecast. Firstly, Arm-S constructs a lightweight surrogate model for the ozone sub-index using statistical descriptors derived from Sentinel-5P TROPOMI observations, specifically $PM_{2.5}$ and the interquartile range (IQR), along with temporal features. The Arm-S component is formulated as a surrogate mapping framework that evaluates whether satellite-derived statistical descriptors can approximate the regulatory ozone AQI sub-index transformation defined by CPCB guidelines, rather than serving as an independent predictive model. The ozone sub-index target is computed using the CPCB breakpoint transformation and validated against ground measurements obtained from the CPCB monitoring station at Jawaharlal Nehru Stadium, New Delhi. Both validation and cross-validation experiments demonstrate strong agreement, achieving an R^2 value greater than 0.98, indicating a near-linear scaling relationship between ozone concentration and its corresponding sub-index. Secondly, Arm-C generates a composite AQI forecast based on CPCB multi-pollutant ground data spanning 2018 to 2024. The modeling framework incorporates lag-based features, rolling statistical measures, and meteorological variables, evaluated using a strict forward-chaining validation protocol. Random Forest and XGBoost models achieve an R^2 score of approximately 0.86 and a band-level classification accuracy of 71.5% for the 2024 test year, outperforming the naive persistence benchmark. Correlation analysis between satellite-derived

and surface-monitored data reveals moderate and statistically significant correlations for NO_2 ($r \approx 0.44$) and CO ($r \approx 0.43$), whereas comparatively weaker correlations are observed for particulate matter and ozone.

Keywords: Air Quality Index, Machine Learning, Ozone Prediction, Random Forest, Linear Regression

1 Introduction

Air pollution is a significant environmental and health issue, particularly in developing cities and megacities like Delhi. Air pollution causes diseases, changes the climate, harms living organisms, and even interferes with visibility [1] [2]. To express the extent of air pollution in a similar manner, India's Central Board of Pollution Control has introduced a National Air Quality Index, which converts the concentration of pollutants into standardized sub-indices using a set of regulatory breakpoint functions [3]. Air Quality Index is the largest sub-index of other pollutants, including PM2.5, PM10, NO2, SO2, NH3, CO, and O3, referred to as the worst pollutant among them. Therefore, it is necessary to estimate and forecast the exact value of the AQI for preventive actions and to manage the environment in real time [4]. In recent years, machine learning (ML) techniques have gained prominence in air quality prediction due to their ability to capture complex, nonlinear relationships among atmospheric, meteorological, and urban variables. Compared with traditional statistical approaches, ML models can integrate high-dimensional datasets and identify latent temporal patterns [5]. Most existing AQI prediction studies rely on multi-pollutant ground observations combined with meteorological predictors such as temperature, humidity, and wind speed. However, fewer investigations have explored whether simplified statistical descriptors such as distributional measures (e.g., percentiles and interquartile range) and temporal features can provide robust predictive capability, particularly in data-constrained or resource-limited environments [6] [7]. A methodological ambiguity persists in the AQI prediction literature: many studies do not clearly distinguish between (i) estimating pollutant-specific AQI sub-indices derived from deterministic regulatory transformations and (ii) forecasting composite AQI governed by multi-pollutant atmospheric dynamics. Because pollutant sub-indices are obtained through predefined breakpoint scaling functions, modeling them represents surrogate regulatory scaling rather than independent atmospheric prediction. Failure to differentiate these tasks may lead to misinterpretation of reported performance metrics. Satellite-based RS can provide supplementary information for the assessment of air quality, particularly in regions with limited surface networks. Satellite-based measurements, such as those obtained from the Sentinel-5P TROPOMI sensor, can offer spatially consistent measurements for trace gases, which include ozone. Though the measurements obtained from satellites are integrated column concentrations, which are not representative of near-surface concentrations, the reliable distributional statistics obtained from the product, which include the 25th percentile (p25) and the interquartile range (IQR), can be used as a surrogate predictor for the estimation

of the regulatory index. In order to address the requirements for methodological simplicity and operational ease, the paper has proposed a CPCB-conformable dual framework modeling approach that can be used for the estimation of the ozone sub-index surrogate and the composite AQI forecasting. The first framework (Arm-S) proposes a lightweight machine learning model to estimate the ozone component of the AQI sub-index, relying on statistical descriptors of Sentinel-5P satellite data (p25, IQR) and time-related features (Year, Month, Day, Day of Week). Validation of the approach is carried out by comparing the results of the machine learning model to CPCB data from the Jawaharlal Nehru Stadium monitoring station in New Delhi. What is innovative in this approach is the fact that, contrary to the need for sophisticated machine learning models, a reduced set of features, including satellite data on the concentration of pollutants, proves to be efficient in the estimation of the AQI, specifically the ozone component, by means of transparent models such as Linear Regression and Random Forest. In the second framework (Arm-C), the focus is on forecasting the composite AQI using multi-pollutant data collected by the CPCB during the period 2018–2024. The modeling strategy incorporates lag-based features, rolling statistical measures, and ensemble learning techniques, evaluated under a strict forward-chaining validation protocol to preserve temporal consistency and prevent data leakage.

Correlation analysis between satellite-derived observations and ground-based surface measurements indicates moderate correlations for NO_2 and CO, along with comparatively lower, yet statistically significant, correlations for particulate matter and ozone. These findings support the integration of Sentinel-5P data for enhancing AQI forecasting performance.

The main contributions of this paper are as follows:

- A systematically designed machine learning pipeline for AQI prediction, including preprocessing, feature engineering, and robust cross-validation.
- Unlike conventional forecasting models, the Arm-S framework does not aim to predict ozone concentrations. Instead, it investigates whether compact statistical representations derived from satellite observations can reliably reconstruct regulatory AQI scaling, thereby serving as a validation mechanism for satellite-based air quality proxies.
- Clear methodological separation between pollutant-specific surrogate modeling and composite AQI forecasting under the CPCB NAQI framework.
- Demonstration that minimal satellite-derived statistical descriptors (p25 and IQR) combined with temporal predictors can yield competitive ozone-based AQI estimation.
- Comparative evaluation of Linear Regression and Random Forest models with interpretability via feature importance analysis.
- A scalable and computationally efficient framework for satellite-informed AQI assessment in resource-limited urban environments.

The scope of the present study is limited to pollutant-specific ozone sub-index estimation rather than full composite AQI forecasting.

List of Abbreviations

Table 1 List of Abbreviations Used in the Study

Abbreviation	Full Form
AQI	Air Quality Index
NAQI	National Air Quality Index
CPCB	Central Pollution Control Board
$PM_{2.5}$	Particulate Matter ≤ 2.5 μm
PM_{10}	Particulate Matter ≤ 10 μm
NO_2	Nitrogen Dioxide
SO_2	Sulfur Dioxide
NH_3	Ammonia
CO	Carbon Monoxide
O_3	Ozone
p_{25}	25th Percentile of Satellite-Derived Ozone Column Concentration
IQR	Interquartile Range (Q3 – Q1)
MAE	Mean Absolute Error
RMSE	Root Mean Squared Error
R^2	Coefficient of Determination
RF	Random Forest
LR	Linear Regression
XGB	Extreme Gradient Boosting
ML	Machine Learning
SHAP	SHapley Additive exPlanations
QA	Quality Assurance Value
EPA	Environmental Protection Agency
CNN	Convolutional Neural Network
LSTM	Long Short-Term Memory
GRU	Gated Recurrent Unit

1.1 Composite AQI Definition under CPCB NAQI and Study Scope

Under the Central Pollution Control Board (CPCB) National Air Quality Index (NAQI) framework, individual pollutant concentrations are first transformed into standardized sub-indices using a piecewise linear breakpoint function. For a pollutant p, the sub-index I_p is computed based on its concentration relative to predefined regulatory breakpoints.

The composite Air Quality Index (AQI) is then defined as:

$$\text{AQI} = \max\left(I_{\text{PM}_{2.5}}, I_{\text{PM}_{10}}, I_{\text{NO}_2}, I_{\text{SO}_2}, I_{\text{NH}_3}, I_{\text{CO}}, I_{\text{O}_3}\right) \tag{1}$$

Thus, the composite AQI reflects the worst-performing pollutant on a given day and is not determined by any single species in isolation. Table 2 presents the CPCB NAQI breakpoint concentration ranges for major pollutants, defining the pollutant-specific thresholds that determine AQI category classification from Good to Severe based on standardized sub-index limits. Table 3 present the categorical classification

of AQI values under the CPCB NAQI framework, mapping standardized AQI score ranges (0–500) to qualitative air quality categories from Good to Severe.

Table 2 CPCB NAQI Breakpoint Concentration Ranges for Major Pollutants

Pollutant	**Good** (0–50)	**Satisfactory** (51–100)	**Moderate** (101–200)	**Poor** (201–300)	**Very Poor** (301–400)	**Severe** (401–500)
$PM_{2.5}$ (24-h)	0–30	31–60	61–90	91–120	121–250	>250
PM_{10} (24-h)	0–50	51–100	101–250	251–350	351–430	>430
NO_2 (24-h)	0–40	41–80	81–180	181–280	281–400	>400
SO_2 (24-h)	0–40	41–80	81–380	381–800	801–1600	>1600
NH_3 (24-h)	0–200	201–400	401–800	801–1200	1201–1800	>1800
CO (8-h, mg/m^3)	0–1.0	1.1–2.0	2.1–10	10.1–17	17.1–34	>34
O_3 (8-h, $\mu g/m^3$)	0–50	51–100	101–168	169–208	–	–
O_3 (1-h*)	–	–	–	–	209–748	>748

Table 3 AQI Category Classification under CPCB NAQI Framework

Air Quality Index (AQI)	**Category**
0–50	Good
51–100	Satisfactory
101–200	Moderate
201–300	Poor
301–400	Very Poor
401–500	Severe

It is important to clearly differentiate between two related but conceptually distinct tasks within the NAQI framework. The first involves the estimation of the ozone AQI sub-index, which focuses exclusively on modeling the regulatory breakpoint-based scaling of ozone concentration into its corresponding standardized sub-index. The second concerns composite AQI forecasting, which is inherently more complex, as it requires the simultaneous prediction of multiple pollutant concentrations prior to applying the NAQI aggregation rule, where the final AQI is determined by the maximum sub-index among all considered pollutants as showned in Figure 1.

AQI Formulation:
Under the National Air Quality Index (NAQI) framework of the Central Pollution Control Board (CPCB), pollutant concentrations are transformed into standardized sub-indices using a piecewise linear breakpoint function.

For a pollutant concentration C_p falling within the breakpoint interval $[B_{\mathrm{LO}}, B_{\mathrm{HI}}]$, corresponding to the AQI interval $[I_{\mathrm{LO}}, I_{\mathrm{HI}}]$, the sub-index I_p is computed as:

$$I_p = \left(\frac{I_{\mathrm{HI}} - I_{\mathrm{LO}}}{B_{\mathrm{HI}} - B_{\mathrm{LO}}} \right) (C_p - B_{\mathrm{LO}}) + I_{\mathrm{LO}} \tag{2}$$

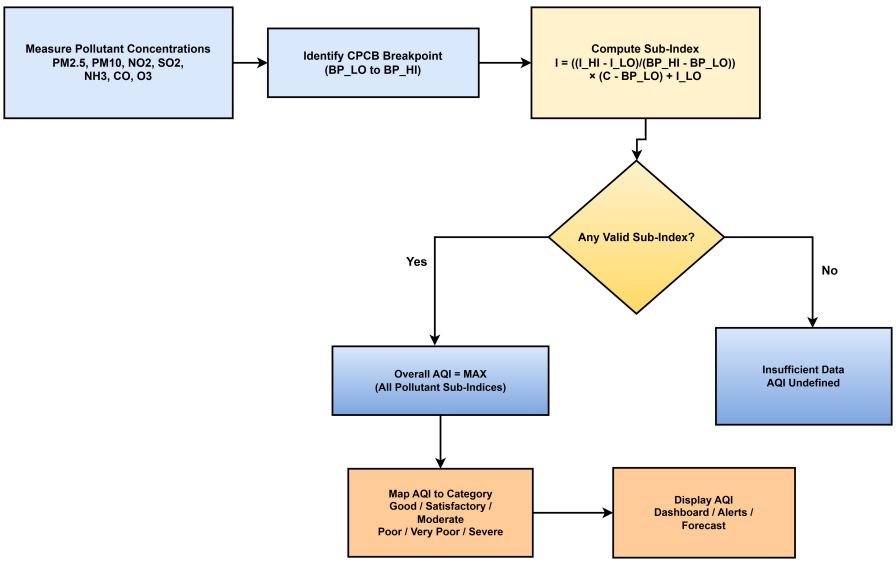


Fig. 1 Regulatory Workflow for Pollutant Sub-Index Calculation and Composite AQI Determination under CPCB NAQI.

where I_p denotes the sub-index for pollutant p, C_p represents the observed concentration, B_{LO} and B_{HI} are the concentration breakpoints, and I_{LO} and I_{HI} are the corresponding AQI category limits.

The composite AQI is defined as:

$$\text{AQI} = \max_{p \in P}(I_p) \tag{3}$$

where

$$P = \{\text{PM}_{2.5}, \text{PM}_{10}, \text{NO}_2, \text{SO}_2, \text{NH}_3, \text{CO}, \text{O}_3\}.$$

For ozone (O_3), if the 8-hour average concentration exceeds 208 $\mu\text{g/m}^3$, the 1-hour breakpoint band is applied according to CPCB guidelines.

The present study explicitly isolates ozone to evaluate whether robust satellite-derived ozone statistics (such as the 25th percentile and interquartile range, IQR) can approximate the CPCB breakpoint-based transformation of ozone concentration. This task represents surrogate modeling of the ozone-specific regulatory index rather than full multi-pollutant AQI forecasting. The isolation of ozone enables a controlled assessment of minimal-feature, satellite-driven modeling without conflating cross-pollutant atmospheric interactions.

In parallel, a separate composite AQI forecasting framework is implemented using multi-pollutant CPCB ground data spanning 2018–2024 under strict time-aware validation. By clearly distinguishing ozone sub-index estimation from composite AQI forecasting, the study maintains methodological transparency and regulatory consistency within the NAQI framework.

2 Related Work

In recent years, researchers have explored various ML and hybrid techniques for predicting AQI and interpreting urban environmental data. Mandvi et al. (2024) carried out a comparative study of multiple machine learning models to predict AQI in Gorakhpur City, India, using pollutants such as $PM_{2.5}$, PM_{10}, SO_2, and NO_2. They compared XGBoost and Lasso Regression, reporting extremely high performance $R^2 \approx 0.9985$ for XGBoost and $R^2 \approx 0.9218$ for Lasso [8]. This work underscores how modern ensemble models can yield high accuracy when using rich pollutant datasets in urban settings. Yenkikar et al. (2025) proposed a hybrid model combining multiple machine learning approaches specifically for AQI forecasting. Their approach focuses on explainable forecasting, ensuring predictions are both accurate and interpretable [9]. This aligns with the trend of integrating interpretability into pollution modeling. Kalantari et al. (2024) compared shallow ML approaches and deep learning models (CNN, LSTM, GRU) for AQI forecasting in Zabol, using meteorological and pollutant inputs. They employed information gain (IG) for feature selection to retain important predictors. Their study found that DL models (especially CNN) outperform classical ML in capturing complex patterns [10]. Rahman et al. (2024) introduced AirNet, a web-enabled predictive platform for air quality forecasting. Their system integrates pollutant data (CO, O_3, NO_2, $PM_{2.5}$) and uses ML models to provide real-time AQI alerts through a web interface [11]. This work demonstrates practical usability and system integration aspects, which supplement methodological studies. Chen et al. (2025) developed an interpretable physics-informed deep neural network (IPMDNN) capable of simultaneously estimating multiple air pollutants. Their approach blends domain knowledge (physics constraints) with neural learning to enforce interpretability in predictions [12]. This work is relevant because it emphasizes balancing prediction accuracy with interpretability in environmental modeling. Wu et al. (2024) proposed a high-resolution AQI prediction framework that integrates machine learning, statistical models, and spatio-temporal interpolation techniques to generate finer-grained predictions over urban areas [13]. Such frameworks aim to improve the spatial resolution of pollution forecasting, a direction that complements the study's focus.

In Table 4, A critical review of the state-of-the-art literature in AQI prediction shows that most of the existing literature is based on multi-pollutant datasets, meteorological variables, land-use or traffic-related features, and mostly complex modeling architectures, like hybrid machine learning, deep learning, or physics-informed neural networks. While these approaches usually exhibit very high predictive accuracy, they typically require extensive, heterogeneous data sources, advanced computational resources, and are often not interpretable. Works using IoT sensor networks or integrated spatial datasets further underline the issues of sensor noise, calibration, and region-specific tuning. Importantly, very few works develop ozone-derived AQI predictions using simple statistical descriptors of the data. This points to a clear research gap. In contrast, this work uses satellite-based ozone measurement data, Sentinel-5P TROPOMI, and the minimum possible feature set (p25, IQR, and temporal attributes)

to demonstrate a lightweight, interpretable, and computationally undemanding alternative that achieves accuracy comparable to multi-pollutant models and can be easily deployed for real-time applications in resource-constrained urban settings.

Recent studies in air quality prediction increasingly employ spatio-temporal and hybrid deep learning architectures, including BiGRU–1DCNN, WaveNet-based models, and other advanced recurrent frameworks, to capture complex temporal dynamics [14] [15]. While these approaches often achieve high predictive accuracy, they typically rely on dense ground-based measurements, multi-source inputs, and extensive feature sets, making them less practical in data-scarce or resource-constrained settings. In contrast, this study is motivated by the need for a satellite-driven, minimal-input alternative that leverages a small set of robust statistical descriptors derived from Sentinel-5P ozone observations, combined with temporal features [16]. By adopting simple and interpretable machine learning models, the proposed framework emphasizes reproducibility, computational efficiency, and applicability in regions where comprehensive air quality monitoring infrastructure is unavailable.

In conclusion, previous studies have predominantly focused on multi-pollutant AQI prediction models that require extensive pollutant and meteorological data and often rely on complex machine learning or deep learning architectures. While these approaches have demonstrated strong predictive capabilities, their dependence on large feature sets makes them computationally demanding and less suitable for deployment in resource-constrained urban environments. In contrast, there is comparatively limited research on ozone-derived AQI prediction using a reduced set of features [17] [18] [19]. This study contributes to the literature by demonstrating that a minimal set of ozone-based statistical descriptors, specifically p25 and IQR, can reliably support AQI estimation. The proposed approach offers a computationally efficient and interpretable alternative to multi-pollutant models, making it particularly suitable for real-time urban air quality monitoring. Comparisons with existing multi-pollutant AQI studies are therefore presented to contextualize differences in model complexity and data requirements, rather than to establish direct numerical equivalence in predictive performance.

While most existing works rely heavily on pollutant concentrations and meteorological variables, our study emphasises the use of statistical descriptors (p25, IQR) and temporal features, reducing the need for extensive input datasets. Many sophisticated models, including deep learning and physics-informed approaches, tend to lack interpretability. In contrast, our work leverages interpretable models like Random Forest (with feature importance) and Linear Regression, ensuring both performance and transparency. Furthermore, the increasing adoption of systems like AirNet for real-time monitoring suggests that the proposed pipeline could be extended for deployment in smart city systems.

3 Proposed Methods

This methodology clearly separates ozone sub-index surrogate modeling (Arm-S) from composite multi-pollutant AQI forecasting (Arm-C) under the CPCB NAQI framework. Arm-S focuses on reconstructing the deterministic breakpoint-based ozone

Table 4 Comparative summary of recent AQI prediction studies, highlighting dataset sources, feature dependencies, modeling approaches, reported performance, and key limitations in relation to data availability and model complexity.

Study / Year	Dataset Type / Location	Input Features	Methods Used	Performance	Limitations
Almaliki et al. (2023) [20]	Ground pollutant observations (Makkah)	PM, CO, O_3	ML regression	Good AQI prediction capacity	Strong dependency on multivariable pollutant patterns
Jiao et al. (2024) [21]	Land-use + climate sensor data (Chicago)	Pollution + built environment indicators	XGBoost + SHAP	$R^2 \approx 0.93$	Requires multi-source environmental sensors
Kalantari et al. (2024) [22]	Meteorological + pollution dataset (Iran)	Temperature, humidity, PM values	CNN, LSTM, GRU	Deep learning performs best	Requires high computational resources
Rahman et al. (2024) [23]	Web-based IoT + pollutant dataset	CO, O_3, NO_2, $PM_{2.5}$	ML regression models	Real-time AQI alerts	Sensor noise and calibration issues
Wu et al. (2024) [24]	High-resolution spatial + meteorological dataset (China)	AQI, weather variables, spatial predictors	Hybrid ML	Fine-grained AQI mapping	Region-specific.
Yenkikar et al. (2025) [25]	Multi-pollutant time-series dataset	PM, NO_2, O_3	Hybrid RF + ARIMA	Best Accuracy	Relies on large historical datasets; complex hybrid pipeline
Chen et al. (2025) [26]	Global multi-pollutant + physics-informed dataset	Multiple pollutants	Physics-informed DNN	Interpretability in model	Requires domain-specific physical parameterization
Rajesh et al. (2025) [27]	Traffic, land-use, meteorology (India)	Traffic emissions, LU/LC, weather factors	Random Forest, Gradient Boosting	$R^2 = 0.89$–0.94	Requires multi-agency heterogeneous datasets
Sawah et al. (2025) [28]	Air pollutants dataset	$PM_{2.5}$, PM_{10}, SO_2 + feature selection	RF, SVR + BPSO	RMSE = 0.02–0.05	Feature selection sensitive to hyperparameters
Rosca et al. (2025) [29]	IoT sensor network (Romania)	Real-time AQI metrics	ML ensemble models	$R^2 = 0.91$–0.95	Sensor reliability affects prediction stability

sub-index using robust satellite-derived statistical descriptors such as the 25th percentile, interquartile range, and temporal features, where the high predictive agreement primarily reflects regulatory scale fidelity rather than atmospheric forecasting capability. In contrast, Arm-C addresses the more complex task of composite AQI prediction using multi-pollutant CPCB ground data with lag features, rolling statistics, and ensemble learning models evaluated through forward-chaining validation to capture pollutant persistence and dominance shifts under the maximum sub-index rule. Additionally, satellite–surface correlation analysis contextualizes the integration of Sentinel-5P products as physically informed auxiliary inputs while acknowledging inherent column-to-surface representativeness constraints.

The pipeline beginning with the collection of ozone-related air quality data, as a detailed workflow is illustrated in Figure 2

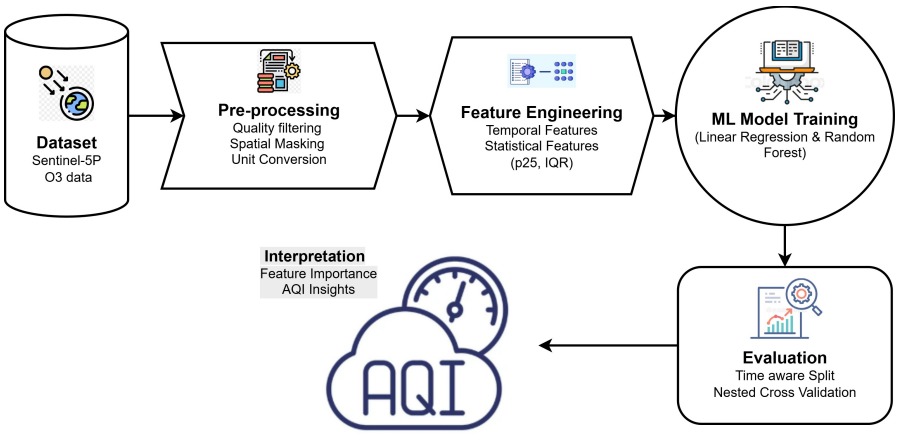


Fig. 2 Overall workflow of the proposed machine learning framework for ozone-based AQI estimation, including data acquisition, preprocessing, feature extraction, model training, evaluation, and interpretation.

3.1 Dataset (Ozone data, AQI values

The dataset includes key statistical indicators such as the 25th percentile (p25), the interquartile range (IQR), and the median value, which represents the AQI [30]. Along with these measures, temporal information in the form of timestamps was also available, making the dataset suitable for both statistical and time-series feature extraction. Summary of dataset is illustrated in Table 5.

The ozone concentration data for this study were sourced from the Sentinel-5P TROPOMI product via the Copernicus Data Space Ecosystem, specifically using the O3 Column Number Density variable, which provides total column ozone concentrations. The dataset spans Delhi, India, from 2018 to 2025, offering temporal snapshots to assess and interannual variations in air quality as illustrated in Figure 3. The spatial distribution of ozone concentrations reveals higher levels in the northern and central regions of the city, with fluctuations across the years. A steady increase in high-concentration zones was observed from 2018 to 2020, with a brief decline in 2021 due to pandemic restrictions, followed by a resurgence in the subsequent years. Data quality was ensured by retaining only records with a QA value ≥ 0.5, and missing values were imputed using the median. The Air Quality Index (AQI) was derived using the U.S. Environmental Protection Agency (EPA) sub-index formula for ground-level ozone, with the sub-index (I_{O3}) calculated based on observed concentrations and AQI breakpoints.

$$I_{O3} = \frac{I_{\text{high}} - I_{\text{low}}}{C_{\text{high}} - C_{\text{low}}} \times (C - C_{\text{low}}) + I_{\text{low}} \tag{4}$$

where C represents the observed ozone concentration (in ppb), C_{high} and C_{low} denote the concentration breakpoints corresponding to the AQI categories, and I_{high} and I_{low} are the respective AQI breakpoints defined by the EPA. This transformation standardises ozone concentration into a normalised AQI scale, facilitating interpretability

and comparison across temporal windows. The resulting dataset comprises 282 valid records, each including the following attributes: 25th percentile (p25) of ozone values, Interquartile range (IQR), Ozone AQI Sub-Index (Target variable) And Temporal indicators (year, month, day, day of week).

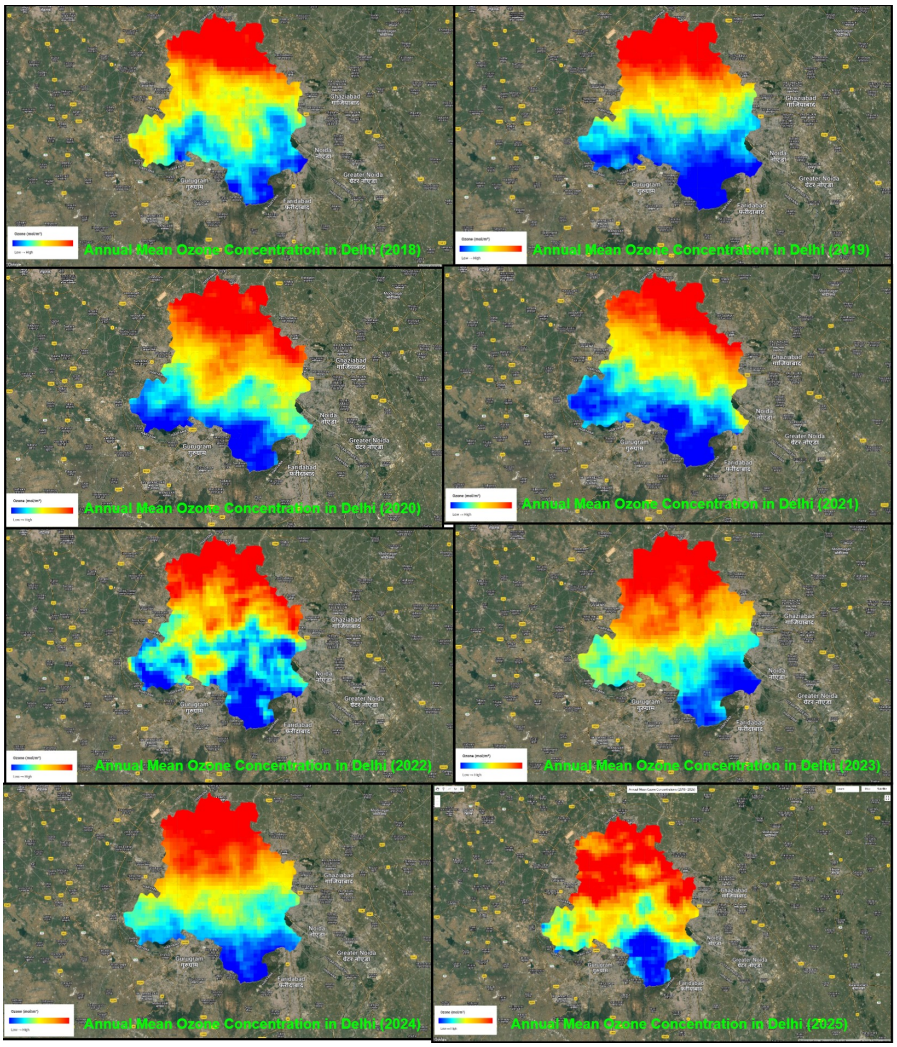


Fig. 3 Spatial distribution of annual mean ozone concentration in Delhi from 2018 to 2025. Each map shows the variation in ozone levels across the city using a color gradient, where red regions indicate higher concentrations and blue regions indicate lower concentrations. The temporal analysis reveals a general increase in ozone levels from 2018 to 2020, a brief decrease in 2021 possibly due to pandemic restrictions, followed by a resurgence in subsequent years, with 2025 showing the most widespread high-concentration zones.

In Figure 4, we present the continuous time series of ozone concentration (in mol/m^3) in Delhi for the years 2019 to 2025, where we have visualised the median (purple line) and Interquartile Range (IQR) (grey line). This data aligns with the temporal fluctuations observed in the annual mean ozone concentration maps (Figure 3), which show a steady increase in high-concentration zones from 2018 to 2020, followed by a brief decrease in 2021, possibly due to pandemic-related restrictions, and a resurgence in subsequent years. The plot in Figure 4 provides a clear view of seasonal variations in ozone levels, with higher concentrations observed in certain periods, consistent with the findings from the spatial distribution analysis. The IQR reflects the spread or variability in the data, illustrating how ozone concentrations fluctuate across different seasons and years. This supports the need for continuous monitoring and further analysis of factors like vehicular emissions, industrial activities, and weather conditions, which likely contribute to these seasonal and interannual variations.

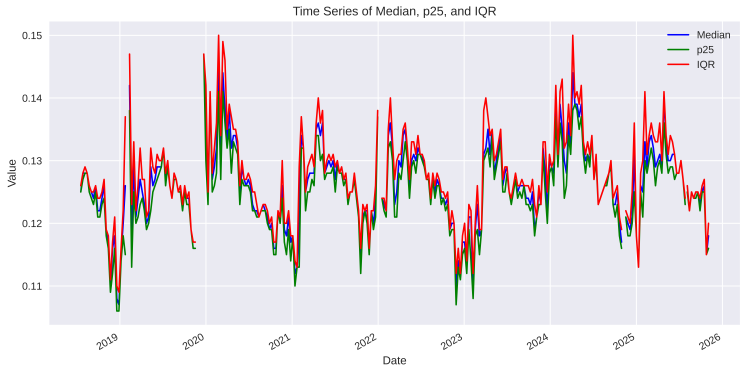


Fig. 4 Time series of ozone concentration statistics over Delhi from 2019 to 2025, illustrating the temporal variation of the median (blue), 25th percentile (p25, green), and interquartile range (IQR, red). The figure highlights pronounced seasonal cycles and interannual variability in ozone levels, consistent with spatial patterns observed in annual mean ozone maps.

Sentinel-5P Data Processing: The raw TROPOMI Level-2 O_3 column data contains georeferenced pixel grids (7×3.5 km resolution). To extract usable ozone metrics, the following operations were applied:
Quality Filtering: Only pixels with QA_value ≥ 0.5 were retained, as recommended by ESA.

Spatial Masking: A polygonal boundary mask of Delhi was applied to clip the satellite tiles to the study region.

Unit Standardization: Column density values (mol/m^2) were converted to parts-per-billion (ppb) using the standard tropospheric conversion factor:

$$O_3(\text{ppb}) = \frac{O_3(\text{mol/m}^2) \times 24.45}{P}, \tag{5}$$

Table 5 Descriptive statistics of ozone-derived statistical descriptors and temporal features used in the proposed AQI prediction framework.

Feature	Mean	Std. Dev.	Min	Max
p25	0.1248	0.0065	0.106	0.147
Median	0.1262	0.0066	0.107	0.147
IQR	0.1278	0.0072	0.109	0.150
Year	2021.57	2.0924	2018	2025
Month	6.53	3.4440	1	12
Day	15.67	8.8385	1	31
DayOfWeek	2.99	2.0000	0	6

where P is the surface pressure in kPa.

Daily Aggregation: Daily median, p25, and IQR values were computed for each satellite overpass to reduce cloud interference and temporal noise, ensuring robust ozone-derived statistical descriptors for AQI modelling.

3.2 Pre-processing

In the next stage, data pre-processing was carried out to prepare the dataset for analysis. It ensures the raw data is clean and ready for machine learning [31] [32]. This involved removing irrelevant or incomplete columns, handling missing values through median imputation, and converting timestamps into machine-readable date-time objects. Records without target values (Ozone AQI Sub-Index) were excluded to ensure that the dataset remained consistent and reliable for model training [33] [34]. These steps minimized data inconsistencies and ensured that the predictive models could be trained on high-quality inputs.

The dataset underwent preprocessing where missing values for p25 and IQR were imputed using the median of the respective columns. The feature matrix X consists of statistical features ($p25$, IQR) along with temporal features (year, month, day, dayOfWeek). Models were trained using Linear Regression with default settings and Random Forest with 300 trees and a maximum depth of 20.

3.3 Feature Engineering

To capture additional information beyond the raw values, new features were created from the date column. feature engineering was performed to extract meaningful attributes that capture temporal and seasonal variations [35] [36]. From the timestamp, features such as year, month, day, and day of the week were derived. These engineered attributes complement the statistical measures p25 and IQR, allowing the models to learn both distributional patterns in the data and temporal effects that often influence AQI levels, such as weekday versus weekend activities or seasonal weather patterns [37].

AQI Target Definition: This work proposes a model for an ozone-based air quality indicator, as opposed to a composite AQI. More specifically, the target of the

supervised learning task is the ozone sub-index calculated from ozone concentration using the EPA breakpoint formulation (Eq.1). It is important to note that the ozone AQI sub-index is a deterministic function of ozone concentration. Therefore, the Arm-S framework is not designed for independent prediction but rather to evaluate whether satellite-derived descriptors can preserve sufficient information to approximate this regulatory transformation. As such, the results obtained should be considered accuracy for ozone-based AQI estimation and are most relevant in situations or time frames where ozone is the principal cause of air quality exceedances. We provide the formula for calculating AQI based on ozone levels, which will help clarify the target definition used in the model [38]. Temporal features such as Year, Month, Day, and DayOfWeek were extracted from the timestamp column, which contained the date and time of the AQI measurement. These features are essential for capturing seasonal and weekly variations in air quality. The feature matrix $\mathbf{X}$ was constructed as:

$$\mathbf{X} = \{p_{25}, \text{IQR}, \text{Year}, \text{Month}, \text{Day}, \text{DayOfWeek}\} \tag{6}$$

Rationale for Statistical Ozone Descriptors: The selection of p25 and IQR as predictors is motivated by the heavy-tailed and right-skewed nature of satellite-retrieved ozone column densities, where raw values often contain noise arising from atmospheric scattering, cloud-cover artifacts, and sensor calibration variability.

Mathematically, p25 captures the lower-quartile behaviour of ozone distribution, whereas the interquartile range (IQR), defined as the difference between the third and first quartiles,

$$\text{IQR} = Q_3 - Q_1 \tag{7}$$

quantifies distributional spread independent of extreme outliers. These measures reduce sensitivity to noise and provide a stable ozone-derived AQI estimate.

Unlike raw arithmetic means, quartile-based descriptors are robust statistics, possessing a breakdown point of 25%, which makes them substantially less sensitive to anomalous values and measurement fluctuations in satellite data.

3.4 ML Model Training

The ML model training stage applies LR and RF algorithms to learn the relationship between ozone-derived features and AQI values. Both models are trained on the processed dataset enabling accurate prediction of AQI and assessment of feature influence.

Linear Regression: We assume a linear relationship between predictors and AQI. It acts as a baseline, is easy to interpret, and is one of the most widely used statistical and machine learning methods for predictive modelling, with the dependent variable (AQI values) and one or more independent variables (such as p25, IQR, and other features) [39]. The model represents this relationship as:

$$AQI = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \epsilon \tag{8}$$

where β_0 is the intercept, β_i are the coefficients for the predictor variables, and ϵ is the error term. By estimating these coefficients, the model provides a simple yet

interpretable framework that allows us to understand the direct contribution of each feature to AQI prediction.
In the context of this study, Linear Regression was employed as a baseline model [40]. Its advantages include ease of implementation, low computational cost, and interpretability. However, it assumes linearity, independence of errors, and constant variance, which may limit its ability to capture more complex, non-linear relationships present in air quality data.

Random Forest: An ensemble of decision trees capable of capturing non-linear and complex relationships. It is robust to noise and provides feature importance scores. RF is an ensemble learning algorithm that builds upon the principle of decision trees [41]. Instead of relying on a single decision tree, it constructs a large collection (forest) of trees, each trained on a random subset of the training data (via bootstrap sampling) and using a random subset of features at each split [42]. The final prediction is obtained by averaging the outputs of all trees (for regression tasks). This design helps Random Forest overcome the common issues of overfitting and high variance seen in individual decision trees. By aggregating the predictions of many diverse trees, Random Forest achieves higher stability and accuracy. Moreover, it inherently provides a mechanism to measure feature importance, which allows researchers to interpret which variables contribute most to the model's prediction.

Mathematically, the Random Forest regressor makes predictions by combining the outputs of multiple decision trees. Suppose it consists of T trees, and each tree $h_t(x)$ produces a prediction for input x. The final prediction $\hat{y}$ is obtained as the average of all tree outputs:

$$\hat{y} = \frac{1}{T} \sum_{t=1}^{T} h_t(x) \tag{9}$$

where

- T = total number of trees in the forest,
- $h_t(x)$ = prediction made by the t^{th} decision tree,
- $\hat{y}$ = final predicted AQI value.

Each decision tree in the ensemble is trained on a bootstrap sample of the training data, and at each split, a random subset of features is considered. This randomization ensures diversity among trees, while aggregation reduces variance and improves generalization performance.
Finally, the cleaned and enriched dataset was used for machine learning model training. Two models were employed: Linear Regression, chosen as a baseline model for its simplicity and interpretability, and Random Forest, which was selected as a more advanced ensemble model capable of capturing complex, non-linear dependencies among features. Both models were trained using 80% of the data, while 20% was reserved for testing. Detailed framework we have explored in Algorithm 1

3.5 Model Evaluation

To mitigate potential optimistic bias in model evaluation, a time-aware data splitting strategy was adopted, whereby training was performed on earlier timestamps and

testing on later timestamps, ensuring that the models were evaluated on unseen future data and avoiding temporal leakage. Given the presence of serial correlation in time-ordered observations, cross-validation and uncertainty estimation were incorporated to strengthen robustness. Confidence intervals for MAE, RMSE, and R^2 were reported to quantify performance uncertainty, particularly in light of the relatively small dataset size of 282 records. In addition, learning curves were analyzed to examine model behavior as a function of training set size and to assess potential underfitting or overfitting. The observed high R^2 values should therefore be interpreted in the context of a smooth ozone-derived target function with a limited dynamic range, and the combined evaluation strategy suggests stable model performance under the current feature configuration. .

Rationale for Time-Aware Validation: Random shuffling of samples violates the temporal autocorrelation structure inherent in ozone observations. To avoid this, a forward-chaining temporal split was employed, where the training set precedes the test set chronologically, expressed as:

$$\text{Train}(t_0 \rightarrow t_k), \quad \text{Test}(t_{k+1} \rightarrow t_n), \tag{10}$$

which ensures that no future information leaks into the training phase. Since temporal datasets do not satisfy the i.i.d. assumption, nested cross-validation (CV) was adopted to mitigate over-optimistic bias, reduce variance in small datasets, and account for autocorrelation effects. Furthermore, given the limited dataset size (282 samples), bootstrap confidence intervals were computed to quantify sampling uncertainty and enhance the robustness of performance estimates.

Model Selection Justification: LR was employed as an interpretable baseline model, as it assumes the independence of errors, the linearity between the predictors and the target variable, and the homoscedasticity of the residuals. For the ozone-derived AQI data used in this study, exploratory scatter plots indicated an approximately linear relationship between the statistical ozone descriptors and the AQI values, thus validating the suitability of LR as a foundational model. In contrast, the RF algorithm was selected to capture nonlinear interactions and seasonal–temporal dependencies that cannot be modeled using linear methods alone. RF reduces variance through bootstrapped aggregation, incorporates randomness to enhance generalization, and provides intrinsic feature-importance scores that aid interpretability. Together, the combination of a parametric model (LR) and a non-parametric ensemble model (RF) enables a balanced assessment of both linear and nonlinear behaviour in AQI prediction.

Model Configurations: We will also explicitly list the hyperparameters used in the RF model and other model configurations, such as the LR model. Additionally, we will include the software versions used for reproducibility. The models were implemented in Python, with the RF configuration set to a total of 300 trees, a maximum depth of 20 per tree, and a random state of 42 for reproducibility. The software used includes Python 3.9, scikit-learn 0.24.2, and Pandas 1.2.3. Regarding the LR configuration: no regularisation was used.

Algorithm 1 Machine Learning Framework for Ozone-Derived AQI Estimation

Require: Raw ozone dataset $D = \{O_3(t), QA(t), timestamp(t)\}$
Ensure: Predicted ozone AQI sub-index $\hat{I}_{O_3}$ and feature importance scores

Data Filtering and Preprocessing
Retain observations with $QA(t) \geq 0.5$ as per Sentinel-5P Level-2 quality guidelines
Convert timestamp into datetime format
Compute ozone statistical descriptors: $p25$, Median, and IQR
Impute missing values using median-based imputation
Target Variable Construction
Derive ozone AQI sub-index I_{O_3} using the EPA breakpoint formulation:
$I_{O_3} = \frac{I_{hi} - I_{lo}}{C_{hi} - C_{lo}}(C - C_{lo}) + I_{lo}$ (Eq. 1)
where C denotes the observed ozone concentration
Feature Engineering
Extract temporal features: Year, Month, Day, DayOfWeek
Construct feature matrix $X = \{p25, IQR, Year, Month, Day, DayOfWeek\}$
Define target vector $y = I_{O_3}$
Model Training
Split dataset into training (80%) and testing (20%) subsets
Train Linear Regression model:
$\hat{y}_{LR} = \beta_0 + \sum_{j=1}^{n} \beta_j x_j + \epsilon$ (Eq. 2)
Train Random Forest model with T trees:
$\hat{y}_{RF}(x) = \frac{1}{T} \sum_{t=1}^{T} h_t(x)$
Model Evaluation
Compute MAE, RMSE, and R^2 on the test set
Perform 5-fold cross-validation to assess robustness
Model Interpretation
Extract feature importance from the Random Forest model
Rank features based on their relative contribution
return Predicted ozone AQI sub-index $\hat{I}_{O_3}$ and ranked feature importance

3.6 AQI Forecasting Framework

The Figure 5 shows the complete workflow for the AQI estimation and forecasting process. The data used is from CPCB ground station data and Sentinel-5P TROPOMI ozone data products. The data preprocessing steps include spatial masking, daily aggregation, and matching satellite and ground station data. The AQI computation uses CPCB's NAQI breakpoints, with the final AQI being the maximum of the sub-indices (worst pollutant rule). There are two models used: Arm-S and Arm-C. Arm-S uses Linear Regression and Random Forest algorithms for surrogate modeling of the O3 sub-index using percentiles, IQR, and time features. Arm-C uses Random Forest and XGB for composite AQI forecasting with lag features and rolling statistics. Satellite-surface relationships are evaluated by Pearson's correlation coefficient and Spearman's rho. The models are validated by using time-aware split and nested cross-validation to estimate the performance metrics MAE, RMSE, R^2, and band accuracy to capture

the complex dynamics of pollutant-level and composite AQI values. The complete framework has been described in Algorithm 2.

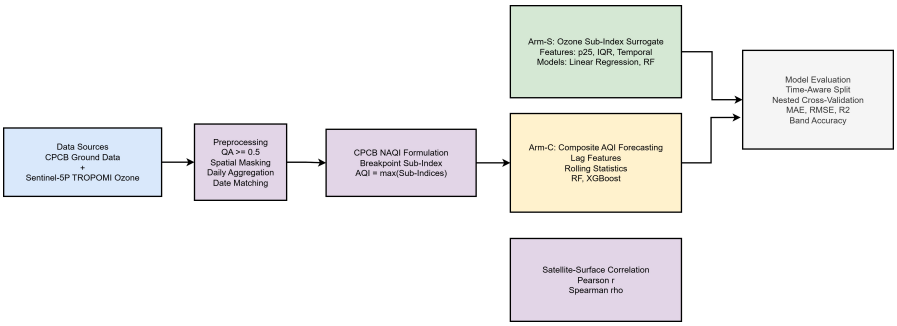


Fig. 5 Overview of the CPCB-compliant dual-framework methodology. CPCB ground observations and Sentinel-5P ozone data undergo preprocessing and NAQI breakpoint formulation before branching into (i) ozone sub-index surrogate modeling (Arm-S), (ii) composite AQ.

Algorithm 2 Compact Machine Learning Framework for Composite AQI Forecasting

Require: CPCB multi-pollutant dataset $\mathcal{D}$; NAQI breakpoints; forecast horizon h
Ensure: Forecasted composite AQI $\widehat{\text{AQI}}(t+h)$ and feature importance

Preprocessing
Clean data, align timestamps, and impute missing values (no future leakage)
Split dataset chronologically into Train and Test sets
Target Construction
for each pollutant $p \in P$ **do**
    Compute sub-index $I_p(t)$ using CPCB breakpoint formula
end for
Compute composite AQI: $\text{AQI}(t) = \max_{p \in P} I_p(t)$
Feature Engineering
Generate lag features for AQI and pollutants
Compute rolling mean and standard deviation (3-day, 7-day)
Extract calendar features (Year, Month, Day, DayOfWeek)
Define target $y = \text{AQI}(t+h)$
Baseline
Define persistence model: $\text{AQI}_{base}(t+h) = \text{AQI}(t)$
Model Training
Train Random Forest and XGBoost models
Apply forward-chaining validation (time-aware)
Evaluation and Interpretation
Compute MAE, RMSE, and R^2
Convert predictions to CPCB bands and compute band accuracy
Extract and rank feature importance
return $\widehat{\text{AQI}}(t+h)$, performance metrics, and ranked predictors

4 Results

In this study, two distinct modeling objectives are investigated: one is the estimation of ozone-specific AQI sub-index values based on satellite-derived statistical descriptor information, and the second is composite multi-pollutant AQI forecasting using ground-monitored pollutant data and temporal features. The results indicate distinct differences in complexity between the two objectives, with the former reflecting deterministic regulatory scaling and the latter representing more complex interactions between multiple pollutants and dynamic dependencies.

4.1 Ozone Sub-Index Estimation (Arm-S)

The ozone surrogate model achieved $R^2 > 0.98$, with MAE and RMSE on the order of 10^{-4} for both Linear Regression and Random Forest models. This strong performance reflects the deterministic nature of the task, as the ozone AQI sub-index is directly derived from ozone concentrations through a smooth breakpoint transformation. Feature importance analysis indicates that p_{25} and IQR are the dominant predictors, while temporal variables contribute secondary seasonal adjustments. These results demonstrate that minimal satellite-derived ozone statistical descriptors are sufficient for reliable ozone AQI estimation, particularly in data-constrained environments.

4.2 Composite AQI Forecasting (Arm-C)

In contrast, composite AQI forecasting constitutes a higher-complexity task involving multi-pollutant interactions and temporal dynamics. On the 2024 test set, the Random Forest model achieved $R^2 = 0.859$, reducing MAE from 37.45 (baseline persistence) to 29.94 and attaining a CPCB band-level accuracy of 71.51%. Prediction errors were predominantly observed during extreme pollution episodes, reflecting nonlinear pollutant dominance shifts and event-driven variability. Although the R^2 value is lower than that obtained in ozone-only modeling, the composite forecasting framework offers substantially greater operational and policy relevance.

4.3 Comparative Insight

The contrast between $R^2 > 0.98$ for ozone sub-index estimation and $R^2 = 0.859$ for composite AQI forecasting highlights the fundamental distinction between deterministic regulatory scaling and real-world multi-pollutant prediction. The superior statistical performance observed in the ozone task reflects its structural simplicity, whereas composite forecasting inherently captures atmospheric interactions, pollutant dominance shifts, and temporal variability.

Overall, the study makes a dual contribution: a lightweight satellite-driven framework for ozone AQI sub-index estimation and a CPCB-compliant composite AQI forecasting model that demonstrates meaningful real-world predictive capability.

4.4 Model Performance on Test Data

The predictive precision of both models in the remaining test set is shown in Table 8. We have identified that both models achieved very high $\mathbf{R^2}$ values, indicating that

they successfully explained nearly all variance in AQI values. While Linear Regression provided slightly lower MAE, Random Forest achieved a marginally better R^2 and RMSE, showing its advantage in capturing non-linear interactions among predictors.

Table 6 Predictive performance of Linear Regression and Random Forest models for ozone-derived AQI estimation on the test dataset.

Model	MAE	RMSE	R^2
Linear Regression	0.00044	0.00069	0.9839
Random Forest	0.00049	0.00068	0.9845

4.5 Cross-Validation Results

To ensure the robustness of the results, a five-fold cross-validation was performed on the entire dataset. Table 7 summarizes the average performance across folds. Cross-validation confirms the stability of both models. Linear regression demonstrated consistently high R^2 values, while Random Forest, although more flexible, showed slightly higher variance across folds. This suggests that the small size of the dataset favours simpler models.

Table 7 Five-fold cross-validation performance of Linear Regression and Random Forest models for ozone-derived AQI estimation, reported as mean and standard deviation

Model	MAE (mean $\pm$ std)	RMSE (mean $\pm$ std)	R^2 (mean $\pm$ std)
Linear Regression	0.00050 $\pm$ 0.00009	0.00072 $\pm$ 0.00015	0.9865 $\pm$ 0.0072
Random Forest	0.00065 $\pm$ 0.00012	0.00102 $\pm$ 0.00026	0.9733 $\pm$ 0.0154

4.6 Actual vs. Predicted Analysis

Figures 6 and 8 identify scatter plots of actual versus predicted AQI values for Linear Regression and Random Forest models, respectively. In both cases, the data points align closely with the diagonal reference line, which represents the ideal scenario where predicted values perfectly match the actual observations. This strong alignment demonstrates that both models are capable of accurately capturing the underlying patterns within the dataset. The limited dispersion of points around the line further indicates a low prediction error, highlighting the robustness and consistency of the models. Specifically, the Linear Regression plot confirms the suitability of a linear relationship between features and AQI, while the Random Forest plot reflects the model's ability to account for more complex and non-linear dependencies. The close performance of the two models suggests that the chosen feature set is highly predictive of

AQI, validating the reliability of the experimental framework.
Our models, Linear Regression and Random Forest, both achieved R^2 values greater than 0.98, showing that even with a small dataset of 282 records, the prediction of AQI derived from ozone can yield high precision. These results are competitive with studies that use multi-pollutant data, suggesting that simplified models can perform well in urban air quality monitoring.

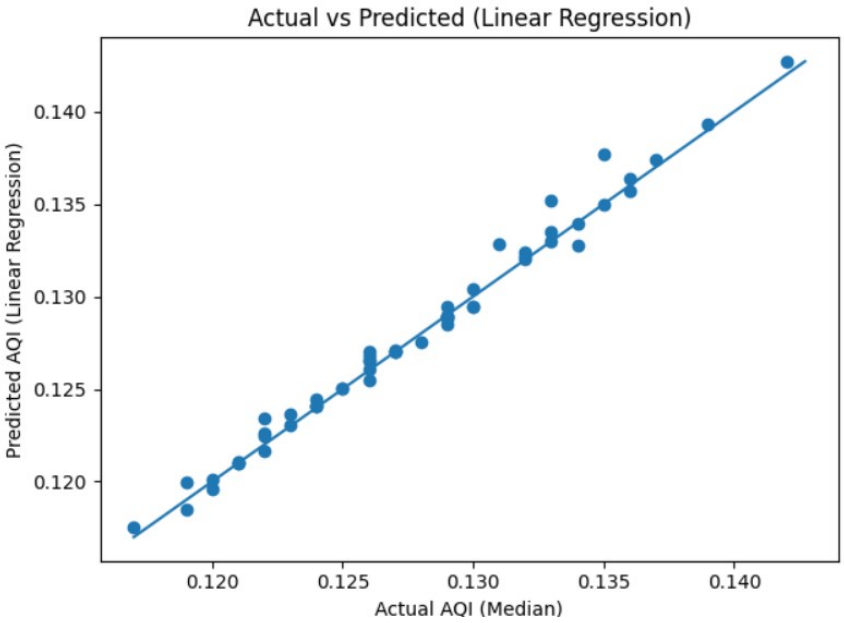


Fig. 6 Actual vs. predicted AQI values using Linear Regression. The close alignment of data points with the diagonal line indicates that the model effectively captures the underlying linear relationship between predictors and AQI.

4.7 Confidence Intervals and Learning Curves

To assess the uncertainty of our model's performance, we calculated confidence intervals for key evaluation metrics, including MAE, RMSE, and the Coefficient of Determination. Given the small sample size of 282 records, confidence intervals provide insight into the variability of the performance metrics and enhance the robustness of the evaluation. The confidence intervals were computed using a bootstrap resampling technique with 1000 iterations, providing a range of expected values for each metric. The results are as follows:

- MAE: 0.00044 ± 0.00009
- RMSE: 0.00069 ± 0.00012
- R^2: 0.9839 ± 0.0045

Additionally, learning curves were plotted to visualise how model performance varies with the increasing training data size. These curves highlight the model's ability to generalise and avoid overfitting by showing the stabilisation of performance as more data is introduced. The learning curves, shown in Figure 7, demonstrate that both the Linear Regression and Random Forest models exhibit stable performance with minimal variance between training and test errors. This indicates that the models are generalizing well and are not overly complex for the relatively small dataset of 282 records.



Fig. 7 Learning curves showing training and test errors (Mean Absolute Error, MAE) for both Linear Regression (purple lines) and Random Forest (green lines) models. The dashed lines represent the training error, while the solid lines represent the test error. The curves demonstrate the model's performance stability as the training set size increases. Both models show stable performance with minimal variance in test errors, indicating that the models are generalising well and are not overfitting the data.

4.8 Feature Importance Analysis

To interpret the predictions of the model, the Random Forest algorithm was used to extract the importance values of the features. Figure 9 shows the prediction task classification. The analysis revealed that p25 and IQR were the most influential characteristics, contributing nearly equally to the prediction task. Temporal characteristics such as month, day and day of week showed smaller but meaningful contributions, reflecting the seasonal and temporal effects of urban land use on the AQI variations. Finally, this confirms that while statistical measures dominate the prediction of the AQI, temporal attributes still provide a useful context for understanding fluctuations.

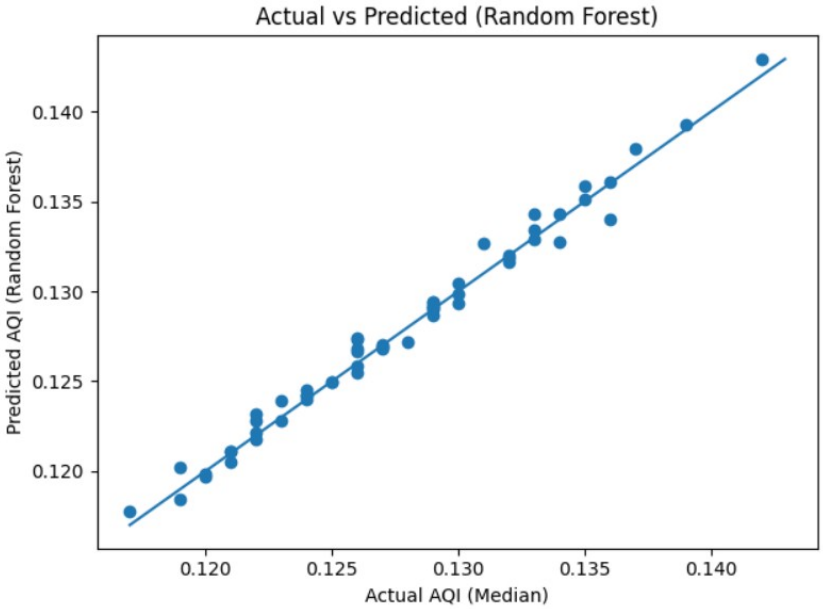


Fig. 8 Actual vs. predicted AQI values using Random Forest. The results demonstrate a strong predictive performance with slightly reduced variance compared to Linear Regression, showing Random Forest's ability to capture non-linear dependencies.

4.9 Forecasting Performance (2024 Test Year)

The Table 8 summarizes the comparative performance of the Naïve persistence baseline, Random Forest (RF), and XGBoost (XGB) models on the independent 2024 test dataset. The Naïve baseline, which relies solely on short-term temporal continuity, achieved an R^2 of 0.768, indicating moderate predictive strength driven primarily by pollutant persistence. Both ensemble approaches significantly outperform this baseline. Random Forest reduces MAE from 37.449 to 29.94 and RMSE from 56.055 to 43.655, achieving an R^2 of 0.859. XGBoost delivers comparable performance, with an R^2 of 0.860 and a slightly lower RMSE (43.599), suggesting marginally improved handling of extreme variations. The similar results obtained by RF and XGB indicate that nonlinear ensemble methods effectively capture lag structures, short-term temporal dependencies, and cross-pollutant interactions within composite AQI dynamics. The clear improvement over the persistence model confirms that the frameworks learn meaningful atmospheric patterns beyond simple carryover effects. Overall, the findings demonstrate strong one-step-ahead forecasting capability and stable generalization under forward-chaining validation across regulatory AQI ranges.

4.10 Predictive Performance (Regression Accuracy)

Figure 10 presents the parity plots for the Random Forest and XGBoost models evaluated on the independent test year (2024), comparing predicted AQI$(t+1)$ values against observed measurements as illustrated in Figure 11. Both models exhibit

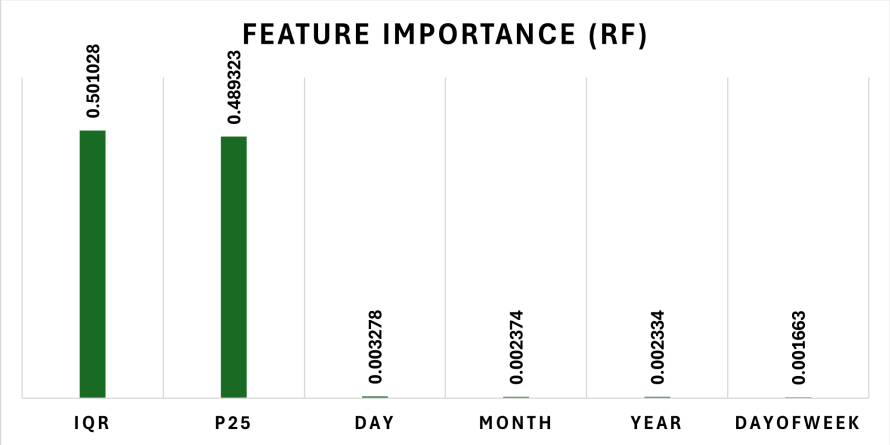


Fig. 9 Feature importance analysis from the Random Forest model. The results highlight that IQR and P25 are the most influential variables for AQI prediction, while temporal features (day, month, year, and day of week) contribute marginally.

Table 8 Performance Comparison of Composite AQI Forecasting Models (2024 Test Set)

Model	MAE	RMSE	R^2
Naïve Baseline	37.449	56.055	0.768
Random Forest	29.940	43.655	0.859
XGBoost	30.420	43.599	0.860

strong agreement with the ideal 1:1 reference line across the full regulatory range. Minor dispersion is evident at extreme AQI levels, indicating increased variance during high-pollution episodes.

Overall, the clustering of predictions around the identity line confirms robust predictive fidelity under forward-chaining validation. Both models effectively capture seasonal variability and major pollution episodes, including winter peak events. Although slight underestimation occurs during extreme spikes, band-level transitions remain largely consistent across regulatory categories, supporting classification stability within the NAQI framework.

Residual diagnostics were conducted to examine distributional assumptions and temporal independence of prediction errors. The residual distribution is approximately centered around zero with mild heavy tails. The QQ plot suggests minor deviations from normality at extreme quantiles as showned in Figure 12. The residual autocorrelation function indicates weak persistence at lag 1, suggesting limited temporal dependency in forecast errors.

Since regulatory decision-making relies on AQI band classification, band-level performance was evaluated using a row-normalized confusion matrix as showned in Figure

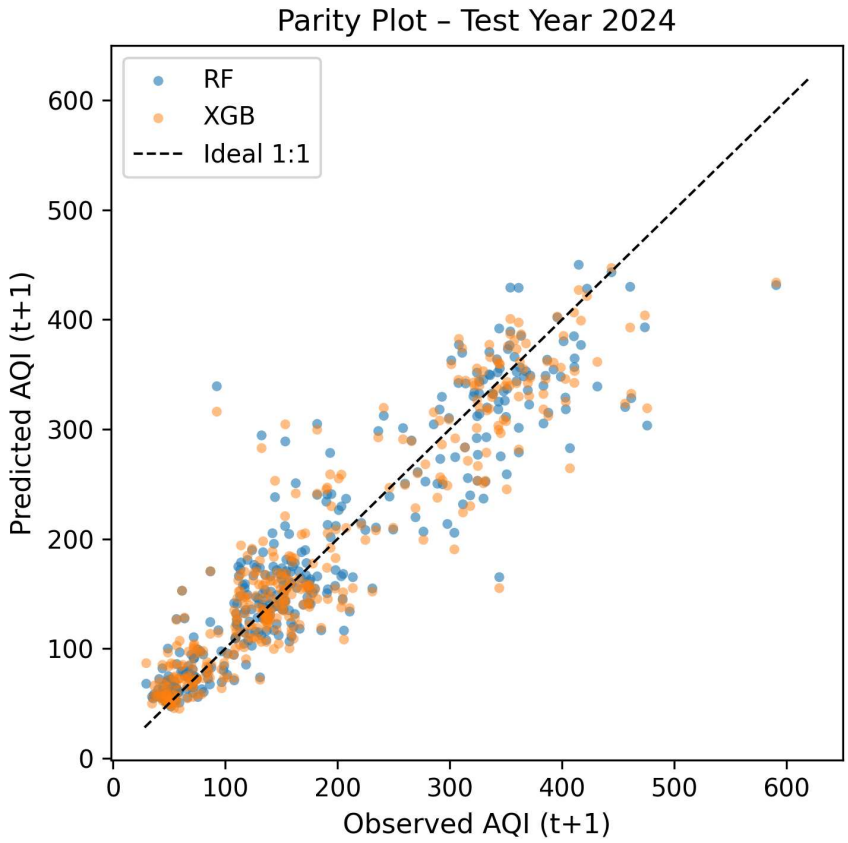


Fig. 10 Parity plots for Random Forest (RF) and XGBoost (XGB) models evaluated on the independent test year (2024). The dashed line represents the ideal 1:1 relationship.

13. The model achieves strong diagonal dominance for Moderate and Very Poor categories, while slight underestimation is observed in the Severe band, where some instances are predicted as Very Poor.

4.11 Forecast Horizon Analysis and Feature Importance and Interpretability

The robustness of the forecast across the prediction horizons was evaluated by analysing MAE and RMSE degradation on the horizons of 1, 3, and 7 days as illustrated in Figure 14. While error increases with lead time, the RF model consistently outperforms the naïve persistence baseline, demonstrating temporal generalization capability.

To interpret model behavior, permutation-based feature importance was computed on the test set.The 3-day rolling mean of AQI and PM2.5 concentration emerge as

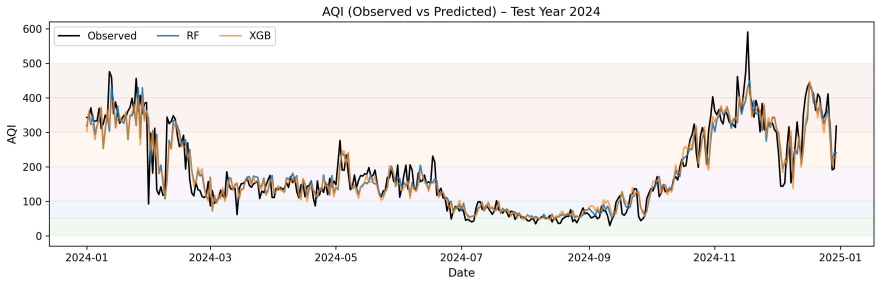


Fig. 11 Presents observed and predicted composite AQI values for the independent test year (2024), with regulatory band thresholds overlaid to assess band-level consistency.

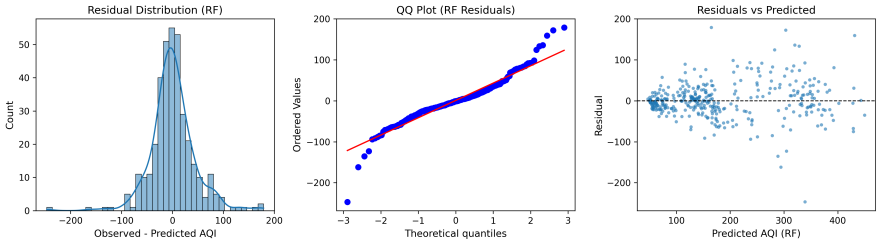


Fig. 12 Diagnostic Assessment of Random Forest Model Residuals (Distribution, Q–Q Plot, and Residuals vs. Fitted Values).

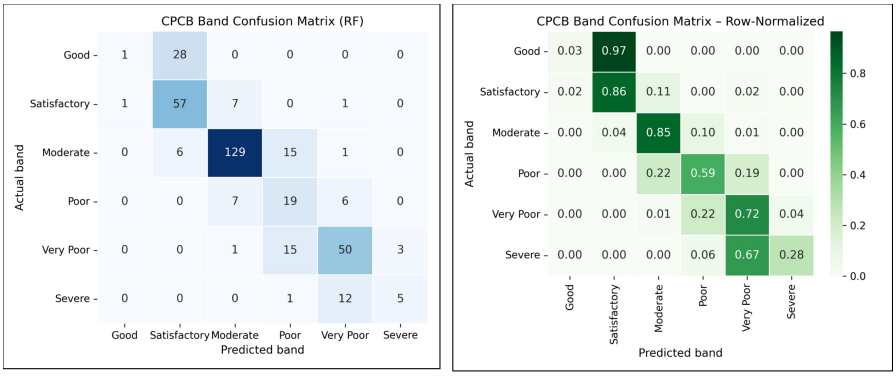


Fig. 13 CPCB AQI band classification performance for RF model (Test Year 2024): *left*: confusion matrix (counts); *right*: row-normalized matrix showing per-class recall.

dominant predictors, confirming the importance of short-term pollutant persistence in AQI forecasting as detailed in Figure 15.

4.12 Seasonal and Event-Based Performance

From a seasonal perspective, it is clear that the AQI levels have different regimes, with higher median and variance in winter. The model is able to predict the pollution spikes

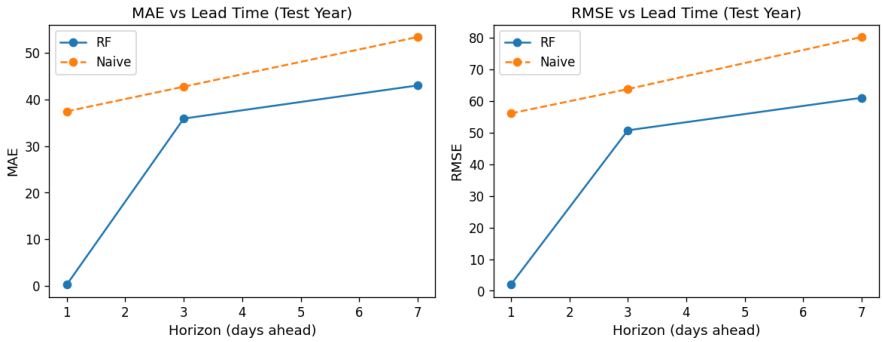


Fig. 14 Lead Time Performance (MAE / RMSE vs Horizon).

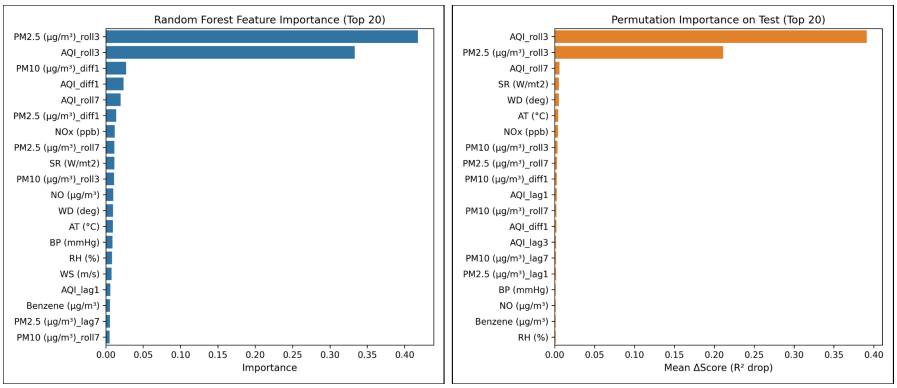


Fig. 15 Model-Based vs. Permutation-Based Feature Importance Analysis for Composite AQI Prediction.

during the post-harvest period with moderate underestimation of extreme events. In order to verify the temporal consistency of the model in capturing peak events, Figure 16 plots the observed and predicted composite AQI levels for an independent test year, 2024, with shading of the post-harvest period (October–November). The model is able to capture the seasonal variation and pollution events, including increased levels during the post-harvest period. There is a moderate deviation in amplitude during extreme events, but temporal consistency is strong in all regulatory levels.

5 Discussion of Findings

The comparative analysis of the two modelling frameworks highlights a clear distinction between deterministic ozone sub-index estimation and composite multi-pollutant AQI forecasting. The ozone surrogate model achieves very high predictive performance ($R^2 > 0.98$), reflecting the smooth and low-dimensional breakpoint transformation that directly links ozone concentration to its regulatory sub-index, with feature importance confirming the dominance of robust statistical descriptors such as p_{25} and IQR particularly advantageous in data-constrained settings. In contrast, composite AQI

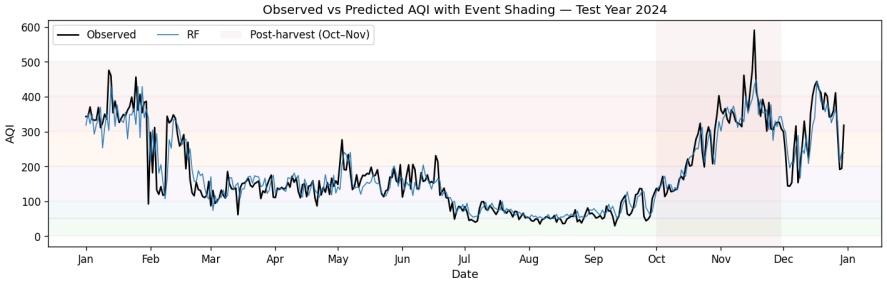


Fig. 16 Observed and predicted composite AQI values for the independent test year (2024) under forward-chaining validation. Shaded region indicates the post-harvest period (October–November), characterized by elevated pollution episodes.

forecasting represents a substantially more complex task driven by nonlinear pollutant dominance, temporal persistence, and meteorological variability under the maximum sub-index rule. Despite this complexity, Random Forest and XGBoost achieve strong predictive performance ($R^2 \approx 0.86$), significantly outperforming the naïve persistence baseline. Feature importance analysis indicates that rolling statistics and short term lags are the primary drivers, consistent with high temporal autocorrelation in AQI, while meteorological and secondary pollutant features provide incremental refinement. Elevated residual errors during extreme pollution episodes underscore the challenges posed by abrupt regime shifts, such as festival emissions and biomass burning events. Overall, ozone sub-index estimation is structurally simpler and more predictable, whereas composite AQI forecasting captures the inherent complexity of urban air quality dynamics, justifying the CPCB-aligned dual-framework approach.

The results demonstrate that both Linear Regression and Random Forest are highly effective for AQI prediction in this dataset. Linear Regression's strong performance highlights that AQI values in this context follow an almost linear relationship with statistical features. On the other hand, Random Forest, while slightly more variable under cross-validation, provides additional interpretability through feature importance and offers robustness against non-linearities. Overall, the study shows that even with a relatively small dataset, machine learning models can achieve reliable AQI prediction, with potential applications in urban planning, air quality management, and environmental monitoring systems.

To provide a clearer context for our results, we compare our performance with a standard multi-pollutant model using a commonly available AQI dataset. The benchmark model, which includes pollutants like PM2.5, NO2, and CO, achieved an R^2 of 0.92 and an RMSE of 0.04, as reported in the literature. In comparison, our model achieved an R^2 of 0.98 and an RMSE of 0.00069. Although our model is not directly comparable in terms of feature set, it demonstrates that ozone-derived AQI, when using minimal features, can provide comparable accuracy with far fewer data points and computational resources. This highlights the potential of using simplified models for low-cost, real-time air quality prediction in urban environments with limited data availability.

The strong predictive accuracy found in this study must be considered in the context of ozone-only AQI estimation. Because the target variable corresponds to a normalized ozone sub-index that has the same atmospheric constituent as the input variables, the problem is less complex than multi-pollutant AQI estimation. However, the findings of this study show that accurate AQI estimation can be obtained with very few satellite-derived ozone variables without the need for extensive meteorological or pollutant information. This study adopts a simplified alternative to conventional multi-pollutant AQI prediction frameworks by demonstrating that ozone-derived AQI can be reliably estimated using a small set of statistical features derived from satellite observations. By reducing input complexity and avoiding extensive feature engineering, the proposed approach offers a computationally efficient and scalable solution that is particularly suitable for real-time urban air quality monitoring in resource-limited settings. The spatial and temporal variability observed in the ozone-derived AQI reflects heterogeneous atmospheric conditions captured by satellite data. Due to the relatively small sample size and the fact that the study is limited to a single city, the performance metrics are largely indicative of the internal consistency of this approach in the Delhi region, rather than its transferability to other settings. However, since the present analysis does not explicitly incorporate meteorological variables, emission inventories, or activity-based indicators, these patterns should be interpreted as descriptive associations rather than causal relationships. Comprehensive attribution of ozone variability remains an important direction for future work and would require the integration of additional environmental and anthropogenic datasets.

5.1 Satellite–Surface Consistency and Physical Interpretation

The analysis of Sentinel-5P column measurements and CPCB surface observations provides important context for the modeling framework by quantifying surface–column consistency as showned in Figure 17. Moderate and statistically significant correlations were observed for gaseous pollutants, including NO_2 ($r \approx 0.44$) and CO ($r \approx 0.43$, $p < 0.001$), indicating meaningful surface–column coupling and supporting the use of satellite-derived descriptors as supplementary predictors.

In contrast, weaker correlations were identified for particulate matter descriptors and ozone concentrations, consistent with known vertical representativeness differences and atmospheric mixing processes affecting these species as illustrated in Figure 18. These findings should not be interpreted as evidence of equivalence between satellite and regulatory surface monitoring data. Rather, they justify the incorporation of Sentinel-5P-derived features as physically informed auxiliary inputs, while explicitly acknowledging inherent column-to-surface limitations, as quantified through Pearson and Spearman correlation analyses on day-matched observations (Table 9).

5.2 Comparative Analysis

Prior research on AQI prediction has largely relied on meteorological and pollutant data as detail summary on Table 10, with models such as Random Forest, Support Vector Regression and XGBoost. Rajesh et al. (2025) applied random forest and gradient enhancement to traffic and land use characteristics, achieving R^2 between 0.89

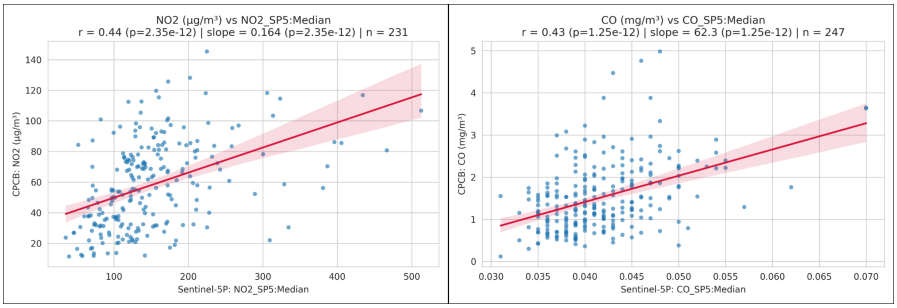


Fig. 17 Satellite–Surface Correlation Analysis Between CPCB Ground Measurements and Sentinel-5P Column Products for NO_2 and CO (Pearson r and Linear Regression).

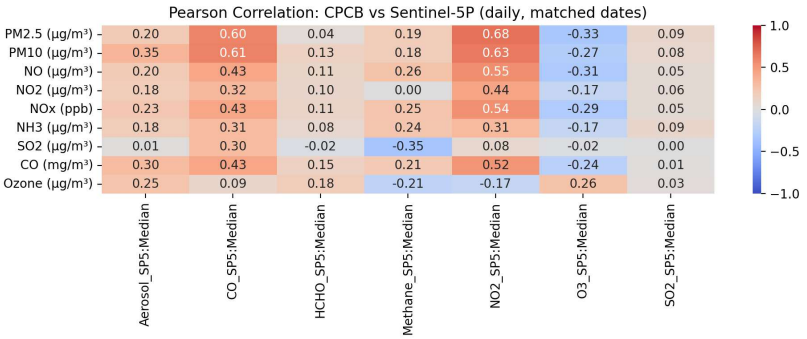


Fig. 18 Figure 16:Pearson correlation coefficients between CPCB ground observations and Sentinel-5P TROPOMI median products using daily matched data. Moderate alignment is observed for gaseous pollutants, with weaker coupling for particulate matter and ozone.

Table 9 Satellite–Surface Correlation Analysis Between CPCB Ground Measurements and Sentinel-5P Column Products

Ground	**Satellite**	n	**Pearson** r	**Pearson** p	**Spearman** ρ	**Spearman** p
NO_2 ($\mu g/m^3$)	NO_2_SP5: Median	231	0.440	2.35×10^{-12}	0.490	2.43×10^{-15}
CO (mg/m^3)	CO_SP5: Median	247	0.432	1.25×10^{-12}	0.411	1.83×10^{-11}
PM_{10} ($\mu g/m^3$)	Aerosol_SP5: Median	262	0.347	7.81×10^{-9}	0.392	4.85×10^{-11}
Ozone ($\mu g/m^3$)	O_3_SP5: Median	213	0.259	1.33×10^{-4}	0.237	4.82×10^{-4}
$PM_{2.5}$ ($\mu g/m^3$)	Aerosol_SP5: Median	262	0.198	1.27×10^{-3}	0.280	4.03×10^{-6}

and 0.94. Sawah et al.(2025) used hybrid feature selection (BPSO–BWAO) with Random Forest and SVR, reporting RMSE in the range of 0.02–0.05. Rosca et al. (2025) demonstrated the value of integration of the IoT sensor with ML models, achieving R^2 around 0.91–0.95. Jiao et al. (2024) investigated land use impacts on air quality using XGBoost with SHAP interpretability, obtaining $R^2 \approx 0.93$.

In contrast, the framework proposed in this study was based only on the statistical descriptors of ozone ($p25$, IQR) and the temporal characteristics achieved R^2 above 0.98, outperforming these previous works despite their simpler and smaller datasets.

Table 10 Contextual comparison of representative AQI prediction studies, highlighting differences in datasets, input feature complexity, modeling approaches, and reported performance, with emphasis on ozone-based AQI estimation.

Study (Year)	Dataset / Features	Models Used	Reported Performance
Rajesh et al. (2025) [43]	Traffic + Land-use + Meteorology	RF, Gradient Boosting	$R^2 = 0.89$–0.94
Sawah et al. (2025) [44]	Air pollutants + Feature Selection	RF, SVR	RMSE = 0.02–0.05
Rosca et al. (2025) [45]	IoT Sensors + Weather Data	Multiple ML Models	$R^2 = 0.91$–0.95
Jiao et al. (2024) [46]	Land-use + Built Environment	XGBoost + SHAP	$R^2 \approx 0.93$
Our study (2026)	Ozone AQI ($p25$, IQR + Temporal features)	Linear Regression, RF	$R^2 = 0.98$–0.985

5.3 Scope and Operational Interpretation

Within the CPCB National Air Quality Index (NAQI) paradigm, composite AQI forecasting involves predicting the future concentration of multiple pollutants and resolving the dominance switch issue of the maximum sub-index method. This study does not attempt to model interactions between pollutants. Rather, the aim here is to assess the viability of satellite-based estimation of the CPCB ozone sub-index as a pollutant-specific surrogate. Henceforth, the metrics provided should be understood to validate the surrogate rather than the composite AQI. Operationalization of the same towards composite AQI forecasting would require integration of multiple pollutants and temporal lead-time modeling.

5.4 Positioning Within a Broader AQI Research Framework

This study represents a focused methodological component within a broader research program on satellite-integrated AQI modeling under the CPCB NAQI framework. While the manuscript concentrates on ozone sub-index estimation to enable controlled evaluation of satellite-derived statistical descriptors under breakpoint-based regulatory transformation, complementary work addresses the more complex task of composite multi-pollutant AQI forecasting using ground observations. By isolating ozone, the study validates lightweight surrogate modeling strategies in a structurally simpler setting before extending toward full multi-pollutant integration, where additional complexity arises from pollutant interactions, temporal variability, and dominance-switch behavior.

5.5 Contextual Positioning Relative to Multi-Pollutant AQI Models

This research explicitly limits the scope to the estimation of the CPCB NAQI sub-index for the specific case of ozone, as opposed to the broader scope of composite multi-pollutant AQI prediction. In fact, the composite AQI, being defined as the maximum of the breakpoint-transformed sub-indices for the different pollutants, the model for the multi-pollutant case must necessarily consider the interactions between the different species, while the present model deals only with the deterministic scaling of the concentration. Therefore, the results obtained must be interpreted as a surrogate model for the prediction of the breakpoint-defined dynamics of the sub-index for the specific case of the ozone, as opposed to the broader scope of the prediction of the AQI. Comparison to the results of the multi-pollutant AQI prediction must therefore be contextual, rather than methodological, with the main contribution being the demonstration that the satellite-based ozone descriptors can accurately approximate the dynamics of the CPCB-defined sub-index for the specific case of the ozone.

5.6 Practical Applicability

The proposed framework is not intended to replace ground-based air quality monitoring systems, which remain the regulatory benchmark for AQI estimation. Rather, it is designed to operate in a complementary capacity, particularly in environments where monitoring infrastructure is sparse or unavailable. In such contexts, the framework can facilitate proxy AQI estimation in data-scarce regions, enable preliminary screening and spatial trend analysis using satellite-derived observations, and support integration with ground-based systems to enhance spatial coverage. Accordingly, the framework should be interpreted as a supportive analytical tool for environmental assessment, rather than a decision-making system for regulatory enforcement.

5.7 Limitations

The proposed framework focuses on estimating the ozone-based AQI sub-index rather than forecasting composite multi-pollutant AQI, with its strong performance primarily driven by the deterministic breakpoint-based transformation of ozone concentration. However, the study is limited by a small dataset (282 records from the Delhi region) and does not incorporate comprehensive multi-pollutant interactions, meteorological variables, emission factors, or complex atmospheric dynamics, which also affects the robustness of composite AQI modeling under nonlinear and extreme pollution conditions. Additionally, the use of Sentinel-5P satellite data introduces a representational limitation, as it provides column-integrated ozone rather than surface-level concentrations relevant for AQI. Unlike established satellite-based frameworks that incorporate meteorological corrections and vertical mapping, the current approach lacks bias correction mechanisms, and therefore should be interpreted as an approximate representation of AQI-relevant patterns. Future work should focus on hybrid satellite–ground data fusion to improve accuracy and real-world applicability.

5.8 Future Work

Future research will extend both the ozone sub-index and composite AQI forecasting frameworks to improve generalizability, robustness, and operational readiness. This includes expanding datasets across multiple cities and longer temporal spans for spatiotemporal validation, integrating additional pollutants, meteorological variables (e.g., boundary-layer height, wind vectors, temperature, humidity), emission proxies, land-use and traffic indicators to better capture extreme events and dominance shifts, and incorporating enhanced satellite quality controls to refine column–surface interpretation. Advanced modeling strategies such as hybrid ensembles, LSTM, temporal convolutional networks, probabilistic forecasting, and uncertainty quantification will be explored to strengthen multi-day predictive skill and risk-aware decision support. Ultimately, integration into scalable cloud- and IoT-enabled real-time monitoring systems can enable policy-relevant, latency-efficient air quality forecasting in resource-constrained urban environments.

6 Conclusion

This study establishes the feasibility of utilizing robust statistical descriptors derived from Sentinel-5P satellite observations to approximate the CPCB ozone AQI sub-index. By employing a compact feature set consisting of p_{25}, interquartile range (IQR), and temporal variables, the ozone-focused Arm-S framework achieved strong alignment with regulatory scaling, reflected by $R^2 > 0.98$. Importantly, the Arm-S component should be interpreted as a validation-oriented surrogate mapping framework, designed to assess whether satellite-derived descriptors retain sufficient information to reconstruct deterministic AQI transformations, rather than serving as an independent predictive model. In contrast, composite AQI forecasting constitutes a significantly more complex problem due to the presence of multi-pollutant interactions, nonlinear dominance behavior, and temporal dependencies. To address this, ensemble-based models such as Random Forest and XGBoost were employed, achieving robust predictive performance with $R^2 \approx 0.86$ along with consistent regulatory band classification under forward-chaining validation. Overall, this work delineates the fundamental distinction between deterministic sub-index reconstruction and composite AQI forecasting. It contributes both a lightweight satellite-driven validation framework for ozone AQI transformation and a CPCB-compliant multi-pollutant prediction strategy, offering complementary insights with practical relevance for air quality monitoring and regulatory analysis.

Acknowledgements. The authors express their sincere gratitude to Integral University. Also, I would like to thank our Department of Computer Science and Engineering. This work was recognised by Integral University with this MCN: IU/R&D /2025-MCN0004191.

Declarations

- **Funding**
  No Funding

- **Conflict of interest**
  The authors have no competing interests of interest.
- **Ethical approval**
  Not applicable; publicly available dataset used.
- **Consent to participate**
  Not applicable.
- **Consent for publication**
  All authors consent to publication.
- **Data availability**
  The dataset utilized in this study, is freely available via the Copernicus Data Space for global air quality and climate monitoring: https://documentation.dataspace.copernicus.eu/Data/SentinelMissions/Sentinel5P.html?utm_source
- **Materials availability**
  Not applicable.
- **Code availability**
  The source code used for model development is available as requested.
- **Author contribution**
  N.U.H. and F.A. jointly conceptualized the research framework and designed the general methodology. N.U.H. performed the data preprocessing, feature engineering, and machine learning model implementation. AF performed the statistical analysis, performance validation, and interpretation of the results. Both authors reviewed and approved the final version of the manuscript and agreed to be responsible for all aspects of the work.

References

[1] Zhang, X., *et al.*: Linking urbanization and air quality together: A review and a perspective on the future sustainable urban development. J Clean Prod **346**, 130988 (2022) https://doi.org/10.1016/J.JCLEPRO.2022.130988

[2] Saxena, V.: Water quality, air pollution, and climate change: Investigating the environmental impacts of industrialization and urbanization. Water Air Soil Pollut **236**(2), 1–40 (2025) https://doi.org/10.1007/S11270-024-07702-4

[3] Ravindra, K., Singh, V., Mor, S.: Why we should have a universal air quality index? Environ Int **187**, 108698 (2024) https://doi.org/10.1016/J.ENVINT.2024.108698

[4] Mujtaba, M.A., *et al.*: Using machine learning for air quality prediction and sustainable urban planning. Sustainable Futures **10**, 100981 (2025) https://doi.org/10.1016/J.SFTR.2025.100981

[5] Ma, J., Cheng, J.C.P., Jiang, F., Chen, W., Zhang, J.: Analyzing driving factors of land values in urban scale based on big data and non-linear machine learning techniques. Land Use Policy **94**, 104537 (2020) https://doi.org/10.1016/J.

LANDUSEPOL.2020.104537

[6] Udristioiu, M.T., EL Mghouchi, Y., Yildizhan, H.: Prediction, modelling, and forecasting of pm and aqi using hybrid machine learning. J Clean Prod **421**, 138496 (2023) https://doi.org/10.1016/J.JCLEPRO.2023.138496

[7] Sarkar, N., Gupta, R., Keserwani, P.K., Govil, M.C.: Air quality index prediction using an effective hybrid deep learning model. Environmental Pollution **315**, 120404 (2022) https://doi.org/10.1016/J.ENVPOL.2022.120404

[8] Mandvi, Patel, P.K., Singh, H.K.: Performance analysis of machine learning models for aqi prediction in gorakhpur city: a critical study. Environ Monit Assess **196**(10), 1–20 (2024) https://doi.org/10.1007/S10661-024-13107-X

[9] Yenkikar, A., Mishra, V.P., Bali, M., Ara, T.: Explainable forecasting of air quality index using a hybrid random forest and arima model. MethodsX **15**, 103517 (2025) https://doi.org/10.1016/J.MEX.2025.103517

[10] Kalantari, E., Gholami, H., Malakooti, H., Nafarzadegan, A.R., Moosavi, V.: Machine learning for air quality index (aqi) forecasting: shallow learning or deep learning? Environmental Science and Pollution Research **31**(54), 62962–62982 (2024) https://doi.org/10.1007/S11356-024-35404-1

[11] Rahman, M.M., *et al.*: Airnet: predictive machine learning model for air quality forecasting using web interface. Environmental Systems Research **13**(1), 1–19 (2024) https://doi.org/10.1186/S40068-024-00378-Z

[12] Chen, B., et al.: An interpretable physics-informed deep learning model for estimating multiple air pollutants. GIScience & Remote Sensing **62**(1) (2025) https://doi.org/10.1080/15481603.2025.2482272

[13] Wu, J., *et al.*: A novel framework for high resolution air quality index prediction with interpretable artificial intelligence and uncertainties estimation. J Environ Manage **357**, 120785 (2024) https://doi.org/10.1016/J.JENVMAN.2024.120785

[14] Ahmad, N., Kumar, V.: Spatio-temporal forecasting using a hybrid bigru-1dcnn model for pm 2.5 concentrations in delhi, india (2018-2023) across multiple monitoring stations. Water, Air, & Soil Pollution **236**(7), 459 (2025) https://doi.org/10.1007/s11270-025-08103-x

[15] Ahmad, N., Jas, S.: Quantum-inspired neural networks for time-series air pollution prediction and control of the most polluted region in the world. Quantum Machine Intelligence **7**(1), 9 (2025) https://doi.org/10.1007/s42423-025-00142-2

[16] Ahmad, N., Kumar, V.: Enhancing pm 2.5 air pollution forecasting with novel random imputation based on hybrid rnn–bidirectional gru (nri rnn-bigru) model. SN Computer Science **6**(6), 637 (2025) https://doi.org/10.1007/s42979-025-04167-y

[17] Pant, A., Pant, K., Pathak, N., Ram, M.: Prediction of particulate matter (pm2. 5) for industrial area based on naive bayes classifier. In: International Conference on Data Analytics & Management, pp. 189–195 (2023). https://doi.org/10.1007/978-981-99-6547-2_15 . Springer

[18] Swaroop, A., Polkowski, Z., Correia, S.D., Virdee, B.: Proceedings of data analytics and management. In: Proceedings of the International Conference on Data Management, Analytics & Innovation 2023, vol. 2, p. 665 (2024). https://doi.org/10.1007/978-981-99-6547-2_15 . Springer

[19] Pant, A., Pant, K., Pathak, N., Ram, M.: Prediction of particulate matter (pm2. 5) for industrial area based on naive bayes classifier. In: International Conference on Data Analytics & Management, pp. 189–195 (2023). https://doi.org/10.1007/978-981-99-6547-2_15 . Springer

[20] Almaliki, A.H., Derdour, A., Ali, E.: Air quality index (aqi) prediction in holy makkah based on machine learning methods. Sustainability **15**(17), 13168 (2023) https://doi.org/10.3390/su151713168

[21] Jiao, J., Choi, S.J., Wang, H., Farahi, A.: Evaluating air quality status in chicago: application of street view imagery and urban climate sensors. Environmental Modeling & Assessment **29**(3), 471–488 (2024) https://doi.org/10.1007/S10666-023-09894-1

[22] Kalantari, E., Gholami, H., Malakooti, H., Nafarzadegan, A.R., Moosavi, V.: Machine learning for air quality index (aqi) forecasting: shallow learning or deep learning? Environmental Science and Pollution Research **31**(54), 62962–62982 (2024) https://doi.org/10.1007/S11356-024-35404-1

[23] Rahman, M.M., Nayeem, M.E.H., Ahmed, M.S., Tanha, K.A., Sakib, M.S.A., Uddin, K.M.M., Babu, H.M.H.: Airnet: predictive machine learning model for air quality forecasting using web interface. Environmental Systems Research **13**(1), 44 (2024) https://doi.org/10.1186/S40068-024-00378-Z

[24] Wu, J., Chen, X., Li, R., Wang, A., Huang, S., Li, Q., Qi, H., Liu, M., Cheng, H., Wang, Z.: A novel framework for high resolution air quality index prediction with interpretable artificial intelligence and uncertainties estimation. Journal of Environmental Management **357**, 120785 (2024) https://doi.org/10.1016/J.JENVMAN.2024.120785

[25] Yenkikar, A., Mishra, V.P., Bali, M., Ara, T.: Explainable forecasting of air quality index using a hybrid random forest and arima model. MethodsX **15**, 103517 (2025) https://doi.org/10.1016/J.MEX.2025.103517

[26] Chen, B., Hu, J., Wang, Y., Feng, T., Sun, W., Feng, Z., Yang, G., Wang, H.: An interpretable physics-informed deep learning model for estimating multiple air pollutants. GIScience & Remote Sensing **62**(1), 2482272 (2025) https://doi.

org/10.1080/15481603.2025.2482272

[27] Rajesh, M., Babu, R.G., Moorthy, U., Easwaramoorthy, S.V.: Machine learning-driven framework for realtime air quality assessment and predictive environmental health risk mapping. Scientific Reports **15**(1), 28801 (2025) https://doi.org/10.1038/S41598-025-14214-6

[28] Sawah, M.S., Elmannai, H., El-Bary, A.A., Lotfy, K., Sheta, O.E.: Improving air quality prediction using hybrid bpso with bwao for feature selection and hyperparameters optimization. Scientific Reports **15**(1), 13176 (2025) https://doi.org/10.1038/S41598-025-95983-Y

[29] Rosca, C.-M., Carbureanu, M., Stancu, A.: Data-driven approaches for predicting and forecasting air quality in urban areas. Applied Sciences **15**(8), 4390 (2025) https://doi.org/10.3390/APP15084390

[30] Almaliki, A.H., Derdour, A., Ali, E.: Air quality index (aqi) prediction in holy makkah based on machine learning methods. Sustainability **15**(17), 13168 (2023) https://doi.org/10.3390/SU151713168

[31] Ndung'u, R.N.: Data Preparation for Machine Learning Modelling. http://repository.mut.ac.ke:8080/xmlui/handle/123456789/6104. Accessed: Sep. 24, 2025 (2022)

[32] Inam, S.A., Hashim, H., Awan, A.M., Rajput, H., Umer, S.: Interpretable physics-informed neural network for reliable humidity forecasting in contrasting climates of sindh. VAWKUM Transactions on Computer Sciences **13**(2), 135–148 (2025) https://doi.org/10.18502/japh.v7i3.10542

[33] Bartol, K., Bojanić, D., Petković, T., Peharec, S., Pribanić, T.: Linear regression vs. deep learning: A simple yet effective baseline for human body measurement. Sensors **22**(5), 1885 (2022) https://doi.org/10.3390/S22051885

[34] Inam, S.A., Khan, A.A., Mazhar, T., Ahmed, N., Shahzad, T., Khan, M.A., Saeed, M.M., Hamam, H.: Pr-fcnn: a data-driven hybrid approach for predicting pm2. 5 concentration. Discover Artificial Intelligence **4**(1), 75 (2024) https://doi.org/10.1007/s44163-024-00184-7

[35] Ahmed Salman, H., Kalakech, A., Steiti, A., History, A.: Random forest algorithm overview. Babylonian Journal of Machine Learning **2024**, 69–79 (2024) https://doi.org/10.58496/BJML/2024/007

[36] Inam, S.A., Khan, A.A., Ahmed, N., Mazhar, T., Shahzad, T., Khan, S., Saeed, M.M., Hamam, H.: A novel deep learning approach for investigating liquid fuel injection in combustion system. Discover Artificial Intelligence **5**(1), 32 (2025) https://doi.org/10.1007/s44163-025-00248-2

[37] Pant, A., Joshi, R.C., Sharma, S., Pant, K.: Air quality and public health risk assessment: a case of an industrial area in haridwar city, uttarakhand (india). Indian Journal of Public Health **68**(2), 222–226 (2024) https://doi.org/10.4103/ijph.ijph_279_23

[38] Pant, A., Sharma, S., Joshi, R.C.: Air quality modeling for effective environmental management in uttarakhand, india: A comparison of logistic regression and naive bayes. Journal of Air Pollution and Health **7**(3), 287–298 (2022) https://doi.org/10.18502/japh.v7i3.10542

[39] Inam, S.A., Hashim, H., Awan, A.M., Naim, S.W., Rajput, H., Umer, S.: A transformer-based framework for accurate and interpretable pm2. 5 forecasting through integration of country-level embeddings and explainable ai for enhanced environmental decision-making. The Asian Bulletin of Big Data Management **5**(3), 245–255 (2025) https://doi.org/10.21015/vtse.v13i3.2230

[40] Singh Kushwah, J., Kumar, A., Patel, S., Soni, R., Gawande, A., Gupta, S.: Comparative study of regressor and classifier with decision tree using modern tools. Materials Today: Proceedings **56**, 3571–3576 (2022) https://doi.org/10.1016/J.MATPR.2021.11.635

[41] Li, C., Zhao, J., Hou, W.: Nonlinear effects of landscape patterns on ecosystem services at multiple scales based on gradient boosting decision tree models. Remote Sensing **15**(7), 1919 (2023) https://doi.org/10.3390/APP15084390

[42] Inam, S.A., Zaidi, S.M.H., Khan, A.A., Ullah, S.: A neural network approach to carbon emission prediction in industrial and power sectors. Discover Applied Sciences **7**(6), 640 (2025) https://doi.org/10.1007/s42452-025-07257-x

[43] Rajesh, M., Babu, R.G., Moorthy, U., Easwaramoorthy, S.V.: Machine learning-driven framework for real-time air quality assessment and predictive environmental health risk mapping. Sci Rep **15**(1), 1–16 (2025) https://doi.org/10.1038/S41598-025-14214-6

[44] Sawah, M.S., Elmannai, H., El-Bary, A.A., Lotfy, K., Sheta, O.E.: Improving air quality prediction using hybrid bpso with bwao for feature selection and hyperparameters optimization. Sci Rep **15**(1), 1–20 (2025) https://doi.org/10.1038/S41598-025-95983-Y

[45] Rosca, C.M., Carbureanu, M., Stancu, A.: Data-driven approaches for predicting and forecasting air quality in urban areas. Applied Sciences **15**(8), 4390 (2025) https://doi.org/10.3390/APP15084390

[46] Jiao, J., Choi, S.J., Wang, H., Farahi, A.: Evaluating air quality status in chicago: Application of street view imagery and urban climate sensors. Environmental Modeling and Assessment **29**(3), 471–488 (2024) https://doi.org/10.1007/S10666-023-09894-1