Open Access Research Journal of
Science and Technology
Journals home page: https://oarjst.com/
ISSN: 2782-9960 (Online)

(Research Article)

Time-series based predictive modeling and explainable artificial intelligence for air quality index forecasting towards sustainable environment in rajahmundry, AP, India

M.R. Goutham [1, *], Suneel Kumar Duvvuri [2], Venkatesh Sunkara [1], Umamahesh Goudu [1] and S.S.K. Chaitanya [3]

[1] *Department of Geology, Government College (Autonomous), Rajahmundry, AP, India.*
[2] *Department of Computer Science, Government College (Autonomous), Rajahmundry, AP, India.*
[3] *Department of Geology, Sir CRR Autonomous College, Eluru, AP, India.*

Open Access Research Journal of Science and Technology, 2026, 16(02), 032-048

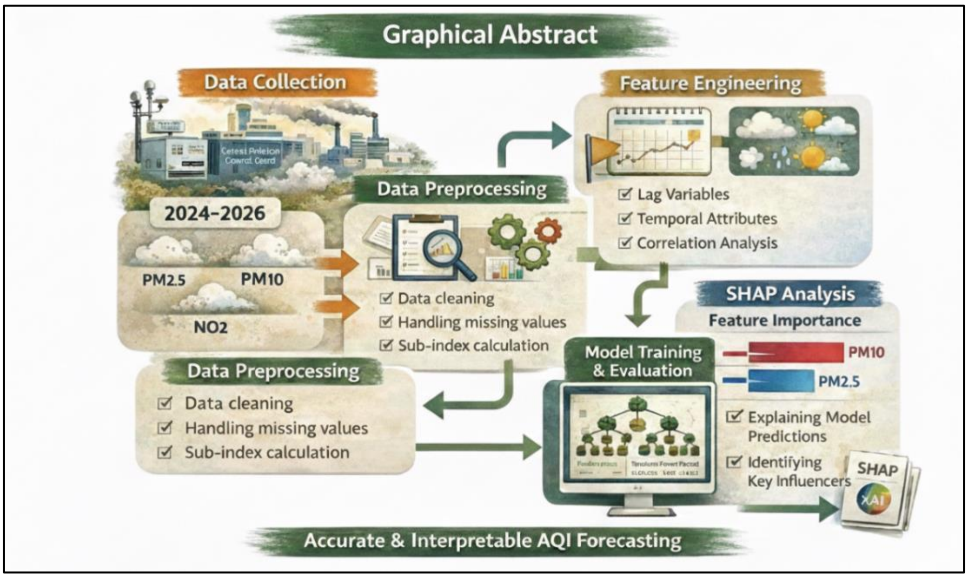
Publication history: Received on 28 January 2026; revised on 07 March 2026; accepted on 09 March 2026

Article DOI: https://doi.org/10.53022/oarjst.2026.16.2.0027

Abstract

Air pollution poses significant health and environmental risks in mid-sized Indian cities like Rajahmundry, where industrial emissions, traffic, and meteorological factors contribute to elevated Air Quality Index (AQI) levels. This study develops a time-series-based predictive model using Random Forest regression to forecast daily AQI, integrating explainable artificial intelligence (XAI) via SHapley Additive exPlanations (SHAP) for interpretability. Utilizing data from the Central Pollution Control Board (January 2024 to February 2026), including pollutants (e.g., PM2.5, PM10, NO2) and meteorological variables, the methodology encompasses data preprocessing, sub-index calculation per Indian standards, exploratory analysis, feature engineering with lags and temporal attributes, and model evaluation. The Random Forest model achieved exceptional performance on a chronological test set, with R^2 of 0.992, MAE of 2.82, and MSE of 14.78. SHAP analysis revealed PM10 and PM2.5 as dominant influencers, highlighting particulate matter's role in AQI spikes. Results indicate moderate average AQI (87.17) with seasonal peaks, underscoring the need for targeted interventions. This framework provides accurate, transparent forecasts to support public health advisories and pollution control in urbanizing regions, addressing gaps in localized air quality modeling.

Graphical abstract



* Corresponding author: M.R. Goutham

Keywords: Air Quality Index (AQI); Random Forest; SHAP; Time-Series Forecasting; Explainable AI; Rajahmundry Air Pollution

1. Introduction

Air pollution continues to pose a major environmental and public health challenge in India, where rapid urbanization, industrial expansion, vehicular growth, and seasonal meteorological patterns frequently drive elevated concentrations of harmful pollutants. In many urban areas, levels of fine particulate matter (PM2.5), coarse particulate matter (PM10), nitrogen dioxide (NO_2), sulfur dioxide (SO_2), carbon monoxide (CO), and ozone (O_3) regularly exceed national and international guidelines, resulting in Air Quality Index (AQI) values that signal risks to residents [1]. Smaller and mid-sized cities in states like Andhra Pradesh face similar pressures, often from localized sources such as industrial estates, road traffic, construction dust, and biomass combustion. Rajahmundry (Rajamahendravaram), a growing urban center in East Godavari district, has recorded concerning air quality episodes in recent years, with PM2.5 concentrations sometimes reaching 50–150 μg/m^3 and AQI readings entering unhealthy categories (100–200 or higher), occasionally surpassing even major metros during peak periods due to industrial emissions along riverbanks, heavy highway traffic, and limited dispersion under certain weather conditions [1] [2]. Monitoring data indicate that such fluctuations align with broader patterns in Andhra Pradesh, where several cities, including Rajahmundry, fall under non-attainment categories for failing to consistently meet National Ambient Air Quality Standards [3].

Long-term exposure to these pollutants is associated with serious adverse effects on human health, ranging from respiratory conditions (such as asthma exacerbation and chronic obstructive pulmonary disease) to cardiovascular disorders, lung cancer, and increased premature mortality. In India, ambient air pollution has been linked to millions of early deaths annually, with PM2.5 identified as a leading contributor to these outcomes in densely populated regions [4] [5]. Children, older adults, and individuals with existing health vulnerabilities experience heightened susceptibility, amplifying the urgency for effective monitoring, early warning systems, and targeted interventions.

Reliable forecasting of AQI and pollutant levels supports timely public health advisories, informs emission reduction strategies, and assists in urban planning to lessen exposure risks. Conventional statistical techniques, including time-series models, often struggle to account for the complex, nonlinear interactions among pollutants and meteorological factors [6]. Machine learning methods, especially ensemble approaches like Random Forest, provide stronger performance by managing high-dimensional inputs and capturing intricate patterns [7]. Nevertheless, many sophisticated models remain opaque, restricting their value for decision-makers who need clear explanations of influencing factors.

Explainable artificial intelligence (XAI) methods, notably SHapley Additive exPlanations (SHAP), address this limitation by offering transparent breakdowns of feature contributions, clarifying how variables such as PM2.5, PM10, or wind speed affect predictions [8] [9]. Emerging applications in air quality research show that pairing accurate forecasting with SHAP improves stakeholder trust, enables focused interventions, and connects technical outputs to practical environmental management [10].

While substantial research exists for large Indian metros, detailed investigations remain limited for mid-sized cities like Rajahmundry, where unique local dynamics-industrial activity, riverine influences, and episodic spikes-require tailored modeling. This work bridges that gap by applying a Random Forest-based time-series forecasting approach to AQI in Rajahmundry, augmented by SHAP for interpretability. The study uses daily pollutant and meteorological data from official monitoring sources to generate accurate, explainable predictions that can guide local authorities in pollution control and public health protection.

2. Literature Review

Research on air quality forecasting has advanced considerably in recent years, shifting from basic statistical tools to sophisticated machine learning setups that better handle the untidy, interconnected nature of pollutant data. Early models frequently utilized ARIMA for time-series forecasting. However, these methods often failed to account for sudden fluctuations. They struggled specifically with the unpredictable changes caused by shifting weather patterns or human behaviour [7]. To address these limitations, researchers now use hybrid systems that combine random forest regression with ARIMA. This approach serves as a practical solution by capturing complex, nonlinear patterns. At the same time, these models maintain transparency through the use of residual checks [7]. These hybrid models effectively reduce prediction errors. For example, they have achieved R^2 scores of approximately 0.94 on Indian datasets. This success stems from combining the strengths of two different methods. While ARIMA manages overall trends, random forest excels at identifying complex interactions between different features [7]. Such approaches stand out for their balance-accurate enough for real-world use but not so opaque that they leave users guessing.

Ensemble techniques have taken center stage lately, especially for urban AQI predictions where data from multiple pollutants and weather factors pile up. Stacking models like gradient boosting, XGBoost, and random forest together has led to impressive results, such as R^2 values hitting 0.997 on Taiwanese air quality records, proving these methods can juggle high-dimensional inputs without losing steam over time [11]. In Indian settings, where pollution spikes from traffic or festivals add extra chaos, these ensembles shine by incorporating SHAP to peel back the layers on what drives AQI changes, like PM2.5's outsized role in winter dips [12]. One stacked setup in Delhi, mixing support vector machines, decision trees, and neural networks, nailed short-term forecasts with lower RMSE than solo models, thanks to its ability to weigh pollutant dynamics against seasonal winds [12]. This kind of layering not only boosts accuracy but also opens doors for policy tweaks, pinpointing emission sources that matter most.

Deep learning has joined the fray, offering ways to grab spatial and temporal quirks that tree-based models sometimes miss. For instance, hybrid frameworks fusing convolutional neural networks with LSTMs have pushed hourly AQI predictions in cities like Beijing, blending meteorological data for RMSE drops of up to 20% over baselines [13]. In smog-heavy spots like Lahore, XGBoost and random forest models have linked emissions from cars and factories to AQI surges, with R^2 nearing 0.99 when fed detailed sector data [14]. These tools excel at nonlinear ties, but their “black box” vibe has sparked calls for more clarity-enter explainable AI, which is now a must-have in recent work.

SHAP has become a go-to for making sense of these complex predictions, turning vague outputs into actionable breakdowns of feature impacts. In a 2026 study on urban air quality, SHAP dissected ensemble models to show how CO and PM10 sway AQI more than expected, aiding health alerts in real time [15]. Paired with random forest, it has clarified why certain pollutants dominate in Indian forecasts, like NO2's link to traffic in Gurugram studies, where interpretability helped refine emission controls [16]. Reviews of 2025 advancements note that tree-based algorithms, bolstered by SHAP, lead in AQI tasks, offering both low errors and insights into challenges like data gaps or model drift [17]. For example, in scalable systems for public health, SHAP and LIME have unpacked deep learning forecasts, revealing wind speed's underrated role in diluting ozone [13]. This push for explainability marks a shift-forecasts aren't just numbers anymore; they're tools for understanding pollution's web.

In India, where air quality woes hit hard in both megacities and smaller hubs, tailored models have gained traction. A recent effort using AlexNet-TabNet hybrids predicted AQI across states, stressing PM2.5's dominance in coastal areas like Andhra Pradesh [18]. Yet, much of the focus stays on Delhi or Mumbai, leaving gaps for places like Rajahmundry, where river proximity and industry mix create unique patterns. Recent overviews highlight how ML can bridge this, with random forest hybrids reducing errors in regional data by 18%, but call for more XAI to make results usable for local planners [16]. Challenges persist, like handling sparse rural data or ensuring models generalize amid climate shifts, but 2026 work on interpretable frameworks shows progress, using SHAP to flag biases in emission-heavy zones [19].

While ensembles and deep learning drive accuracy, integrating SHAP ensures forecasts inform action. For mid-sized Indian cities, this study extends these trends by focusing on Rajahmundry-specific data, addressing the need for localized, transparent models amid rising urbanization.

3. Methodology

This study adopts a quantitative analytical framework that combines time-series forecasting with explainable artificial intelligence to predict the Air Quality Index (AQI) in Rajahmundry, India, based on historical data collected from January 1, 2024, to February 10, 2026. The dataset was obtained from the Central Pollution Control Board (CPCB) portal of the Government of India and comprises daily measurements of primary pollutants, including particulate matter (PM2.5 and PM10), nitrogen oxides (NO, NO2, and NOx), ammonia (NH3), sulfur dioxide (SO2), carbon monoxide (CO), ozone (O3), and volatile organic compounds such as benzene, toluene, ethylbenzene, m-p-xylene, and xylene. Meteorological variables, including ambient temperature (AT), relative humidity (RH), wind speed (WS), wind direction (WD), solar radiation (SR), barometric pressure (BP), vertical wind speed (VWS), rainfall (RF), and cumulative rainfall (TOT-RF), were also incorporated to capture environmental influences on pollutant dispersion.

The analytical pipeline commenced with data acquisition and integration. Pollutant measurements and corresponding threshold standards (defined by CPCB guidelines for AQI computation) were combined into a unified dataset. Pollutant-specific columns were identified and reshaped from a wide format (where each pollutant occupied a separate column) to a long format, facilitating alignment with threshold values through a left join on pollutant identifiers. This step ensured comprehensive coverage of all available measurements while incorporating regulatory breakpoints for subsequent index calculations.

Data preprocessing was performed to enhance quality and usability. Timestamp columns were converted to standardized datetime formats, accommodating potential parsing errors. Prescribed standard ranges (e.g., 0–60 μg/m^3 for PM2.5) were extracted into numerical minimum and maximum bounds using string splitting and conversion techniques, with invalid entries assigned null values. Missing entries were systematically evaluated, revealing approximately 10% gaps in pollutant concentrations and over 50% in timestamps across the initial 11,232 records. To address these, median imputation was applied to pollutant concentrations, stratified by pollutant type to maintain distributional integrity, reducing missing values to less than 7%. Rows with persistent critical gaps (e.g., in timestamps) were excluded, yielding a cleaned dataset of 4,004 entries. Outliers were identified using the interquartile range (IQR) method, with detection rates varying by pollutant (e.g., 26% for SO2 and 14% for NO); these were retained to reflect real-world variability in air quality data, unless they unduly influenced model stability.

Sub-indices for individual pollutants were computed according to the Indian National AQI protocol. For each pollutant concentration C_p at a given timestamp, the sub-index I_p was calculated via linear interpolation within predefined concentration breakpoints. Specifically, if C_p fell below the minimum breakpoint, I_p=0; if above the maximum, I_p = 500; otherwise,

$$I_p = \frac{(C_p - BP_{LO}) \times (I_{HI} - I_{LO})}{(BP_{HI} - BP_{LO})} + I_{LO} \quad (1)$$

where, (BP_{LO}) and (BP_{HI}) are the lower and upper concentration breakpoints, and (I_{LO}) and (I_{HI}) are the corresponding AQI category bounds. This yielded sub-index values ranging from 0 to 418, with a mean of approximately 12, indicating sporadic exceedances of safe levels.

The overall AQI for each daily interval was then determined by aggregating sub-indices and selecting the maximum value per timestamp group, adhering to CPCB requirements that mandate at least three pollutants (including PM2.5 or PM10) for validity. This process generated 286 daily AQI observations, with values spanning 14 to 418 and a mean of 87, suggestive of moderate pollution episodes interspersed with poorer air quality periods.

Exploratory data analysis (EDA) was conducted to discern temporal and relational patterns. Time-series visualizations depicted AQI and key pollutant trends, revealing seasonal fluctuations (e.g., elevated levels during winter due to inversion layers). Histograms and kernel density estimates illustrated distributional characteristics, such as positive skewness in PM concentrations. Box plots stratified by month and day highlighted variability, while a correlation matrix quantified inter-pollutant associations, demonstrating strong positive coefficients (>0.7) between PM2.5, PM10, NO2, and AQI, thereby emphasizing particulate matter as a primary AQI driver.

Feature engineering was undertaken to augment predictive modeling. The dataset was pivoted back to a wide format for model input, with AQI designated as the target variable. Temporal attributes were derived from timestamps, including month (cyclically encoded to account for seasonality), day of the week (binary indicators for weekdays vs. weekends), and inferred hour (for diurnal patterns). Lagged variables were introduced to incorporate autocorrelation, such as 1–3 day prior values for AQI and dominant pollutants (e.g., PM2.5, PM10, NO2). After handling lags-induced nulls, this produced a feature-enriched dataset of 283 samples with 25 predictors.

A Random Forest Regressor was employed for forecasting, chosen for its ensemble-based handling of non-linearities and feature interactions. The model was configured with 100 decision trees and default hyperparameters. To preserve temporal order and simulate prospective forecasting, the data was split chronologically: the initial 80% for training and the latter 20% for testing. No feature scaling was required due to the algorithm's invariance to scale. Training achieved high in-sample fidelity, while test-set evaluation yielded a Mean Absolute Error (MAE) of 2.82, Mean Squared Error (MSE) of 14.78, and coefficient of determination (R^2) of 0.992, indicating superior predictive accuracy.

Model interpretability was enhanced using SHapley Additive exPlanations (SHAP), a game-theoretic approach to attribute prediction contributions. A tree-specific explainer was applied to compute SHAP values for test instances, enabling global and local insights. Mean absolute SHAP values identified PM10 as the most influential feature, followed by PM2.5 and toluene. Dependence plots illustrated monotonic positive relationships, such as increasing PM10 concentrations elevating predicted AQI.

All computational steps were implemented in Python, utilizing libraries for data manipulation, visualization, machine learning, and interpretability to ensure reproducibility and efficiency. This methodology establishes a transparent, scalable framework for urban AQI forecasting, with implications for environmental policy in mid-sized Indian cities.

Algorithm 1: AQI Forecasting Pipeline (pseudocode)

```
Input: Raw pollutant data D, Threshold standards
Output: Forecasted AQI, Model performance metrics, SHAP explanations

1. Merge D and T on pollutant identifiers to form integrated dataset M
2. Preprocess M:
  a. Convert timestamps to datetime
  b. Parse standard ranges into min/max numerics
  c. Impute missing pollutant values with group medians
  d. Detect and handle outliers via IQR (retain unless distortive)
3. Compute sub-indices for each pollutant in M:
  For each row r in M:
    If r.concentration < minbreakpoint: subindex = 0
    Else if r.concentration > maxbreakpoint: subindex = 500
    Else: subindex = linearinterpolation(r.concentration, breakpoints, AQIbounds)
4. Aggregate overall AQI: Group by timestamp, AQI = max(subindices)
5. Perform EDA on AQI-enhanced M: Generate time-series plots, distributions, correlations
6. Engineer features F from M:
  a. Extract temporal: month, dayofweek, hour
  b. Add lags: AQIlag1-3, pollutantlag1 (for key pollutants)
7. Split F into train (80% chronological) and test (20%)
8. Train RandomForestRegressor on train:
  Model = RandomForest(nestimators=100)
  Fit(Model, trainfeatures, trainAQI)
9. Evaluate on test:
  Predictions = Predict(Model, testfeatures)
  Compute MAE, MSE, R²
10. Explain with SHAP:
   Explainer = TreeSHAP(Model)
   Values = ComputeSHAP(Explainer, testfeatures)
   Visualize: Summary plots, dependence plots
Return Predictions, Metrics, SHAPValues
```

4. Results and Discussion

The analysis of air quality data from Rajahmundry revealed significant patterns in pollutant concentrations and AQI trends over the study period from January 1, 2024, to February 10, 2026. After data preprocessing, including median imputation for missing pollutant values and row removal for incomplete records, the cleaned dataset comprised 4,004 entries. Outlier detection using the IQR method identified notable anomalies, particularly in SO_2 (26.22% of values classified as outliers), NO (13.64%), and other pollutants, though these were retained to preserve real-world variability. Box plots of pollutant values by type (Figure 1) illustrated the distribution and potential outliers, showing wide ranges for particulates like PM10 and PM2.5, which often exceeded prescribed standards, indicating episodic pollution events possibly linked to industrial or vehicular sources.

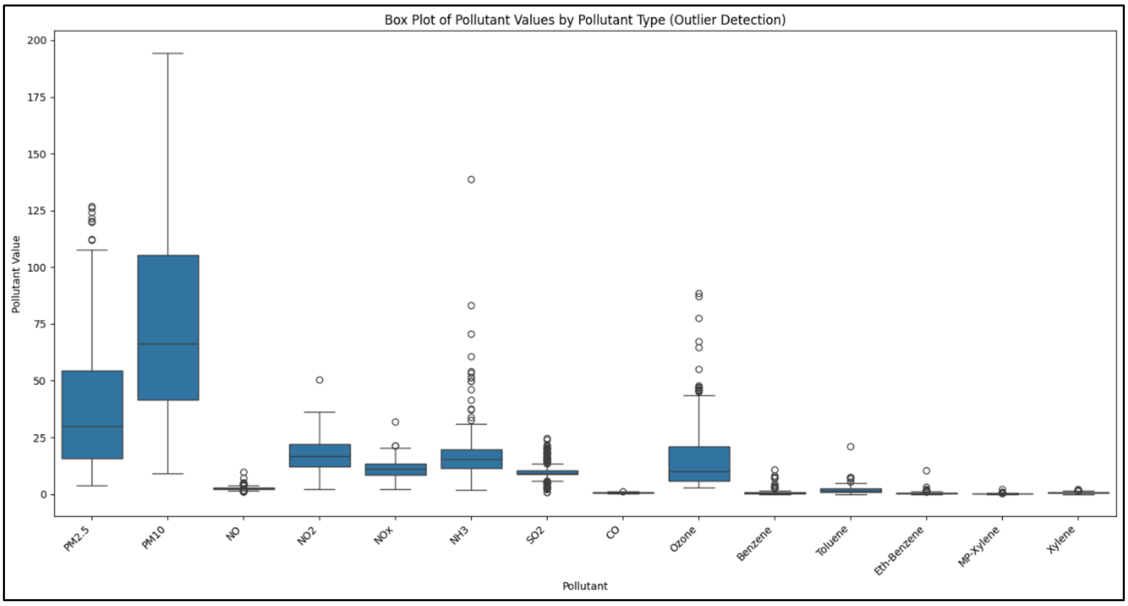


Figure 1 Box Plot of Pollutant Values by Pollutant Type (Outlier Detection). This figure illustrates the distribution of pollutant concentrations across types, highlighting outliers via the IQR method. The x-axis represents pollutants (e.g., PM2.5, PM10, NO), and the y-axis shows concentration values, with boxes indicating interquartile ranges and whiskers extending to 1.5 times the IQR

Individual pollutant sub-indices, calculated via linear interpolation against CPCB thresholds, ranged from 0.00 to 417.80, with a mean of 11.88, highlighting frequent exceedances for key pollutants. The overall AQI, derived as the maximum sub-index per timestamp, varied from 14.25 to 417.80, with a mean of 87.17, classifying average air quality as moderate but with severe spikes. Exploratory data analysis (EDA) provided deeper insights into temporal dynamics. Time-series plots of overall AQI (Figure 2) displayed pronounced fluctuations, with peaks during early 2024 and mid-2025, potentially attributable to seasonal factors such as reduced dispersion in cooler months or increased emissions during festivals. Similar trends were observed in key pollutants: PM2.5 and PM10 showed synchronized elevations (Figure 3), while NO2 exhibited more erratic daily variations, suggesting traffic-related influences.

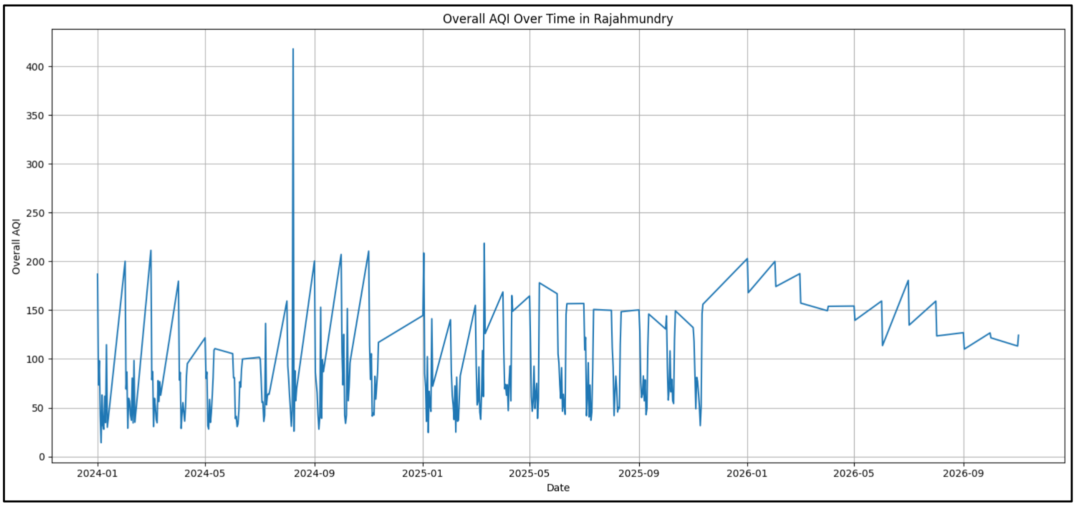


Figure 2 Overall AQI Over Time in Rajahmundry. A time-series line plot depicting AQI fluctuations from January 2024 to February 2026. The x-axis is the date, and the y-axis is the AQI value, revealing seasonal peaks and trends

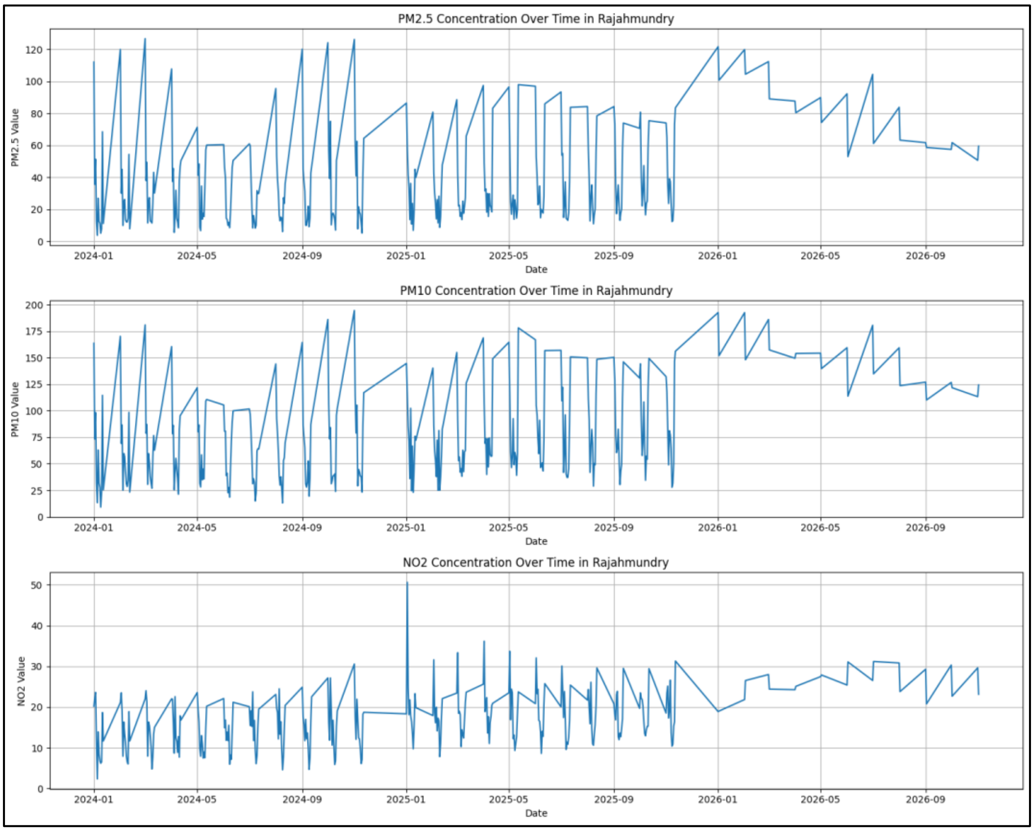


Figure 3 Key Pollutant Concentrations Over Time in Rajahmundry. Subplots for PM2.5, PM10, and NO2, each as a time-series line plot. The x-axis is the date, and the y-axis is the pollutant value for each subplot, showing synchronized variations among particulates

Distribution plots, including histograms and KDEs for AQI (Figure 4), revealed a positively skewed profile, with most values clustered in the 50–150 range but tailing into hazardous levels (>300), underscoring the need for mitigation in extreme cases. Histograms and KDEs for key pollutants (Figure 5) further emphasized variability, with PM10 showing broader spreads. Box plots stratified by month and day of the week for AQI (Figure 6) demonstrated seasonal variability, with higher median AQI and wider interquartile ranges in winter months (e.g., December–February), likely due to atmospheric inversions trapping pollutants. Daily patterns showed slight elevations on weekdays, possibly reflecting increased urban activity, though weekends exhibited outliers indicative of sporadic events. For individual pollutants, monthly and daily box plots (Figure 7) confirmed PM10 and PM2.5 as dominant contributors, with medians often approaching or surpassing thresholds (e.g., PM10 >100 μg/m^3 in several months). A correlation heatmap (Figure 8) quantified relationships, showing strong positive correlations between PM2.5, PM10, NO2, and AQI (coefficients >0.7), underscoring particulate matter's dominance.

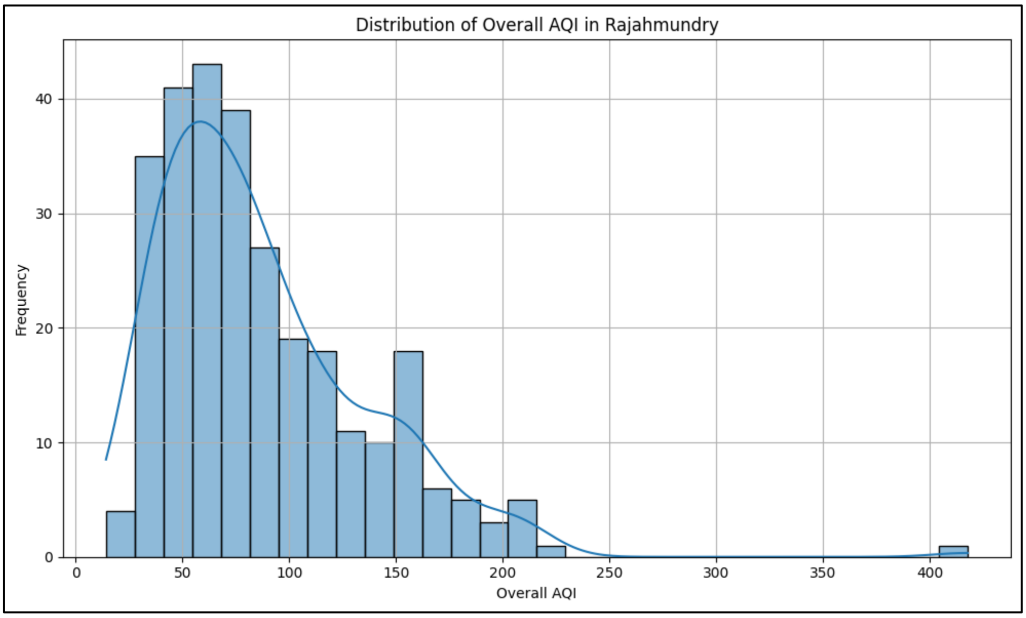


Figure 4 Distribution of Overall AQI in Rajahmundry. A histogram with kernel density estimate (KDE) overlay, bins=30. The x-axis is AQI values, and the y-axis is frequency, demonstrating positive skewness with a mode around moderate levels

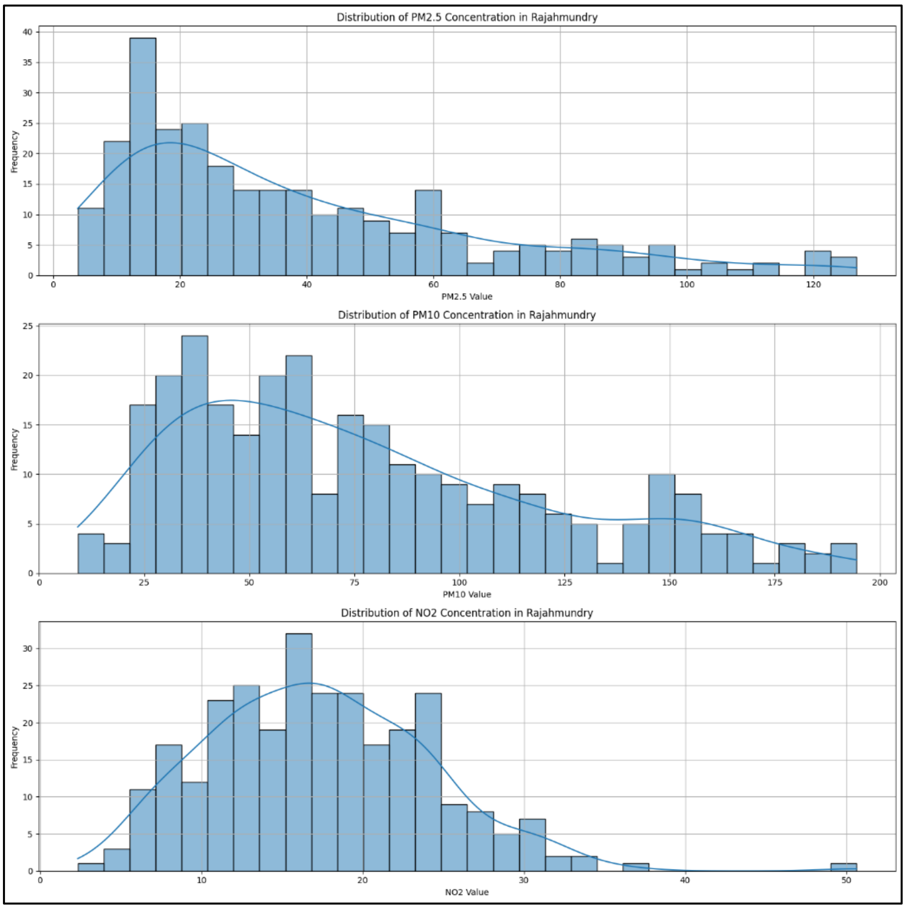


Figure 5 Distributions of Key Pollutant Concentrations in Rajahmundry. Subplots of histograms with KDE for PM2.5, PM10, and NO2. Each subplot has the x-axis as pollutant value and y-axis as frequency, illustrating variability and tails in concentrations

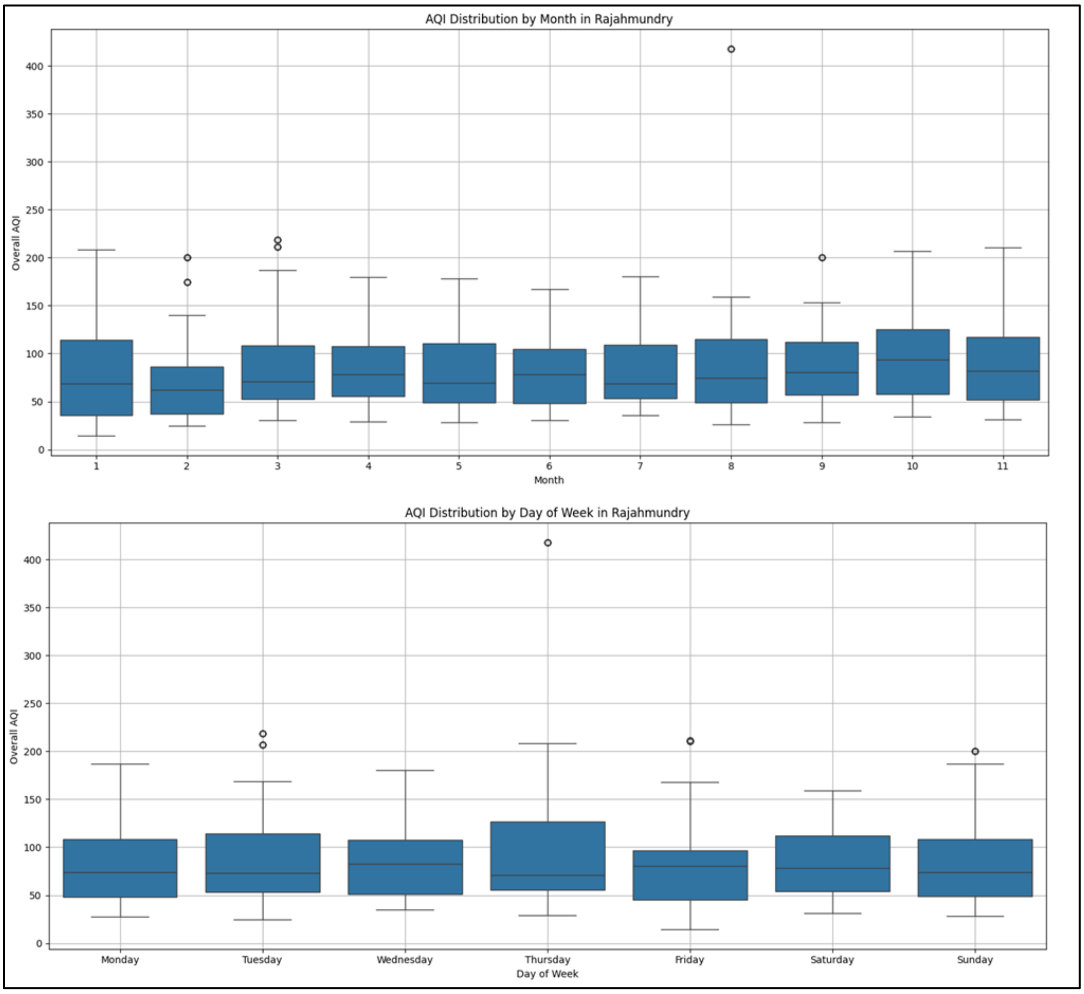


Figure 6 AQI Distribution by Month and Day of the Week in Rajahmundry. Box plots stratified by month (top) and day of the week (bottom). The x-axis categories are months (1–12) or days (Monday–Sunday), and the y-axis is AQI, highlighting seasonal and weekly patterns

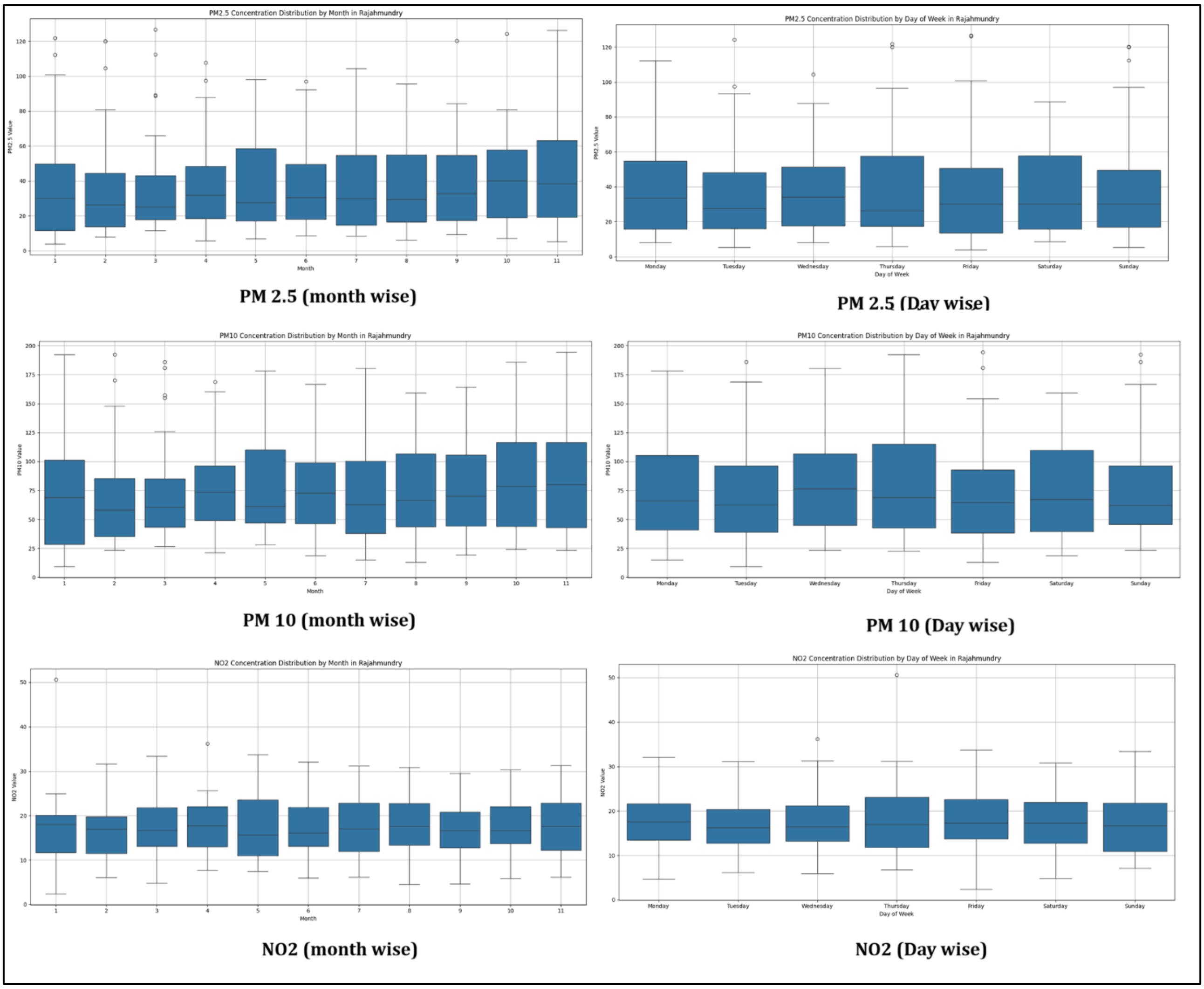


Figure 7 Key Pollutant Concentration Distributions by Month and Day of the Week for PM2.5, PM10, and NO2 . The x-axis is month/day categories, and the y-axis is pollutant value, emphasizing temporal variability

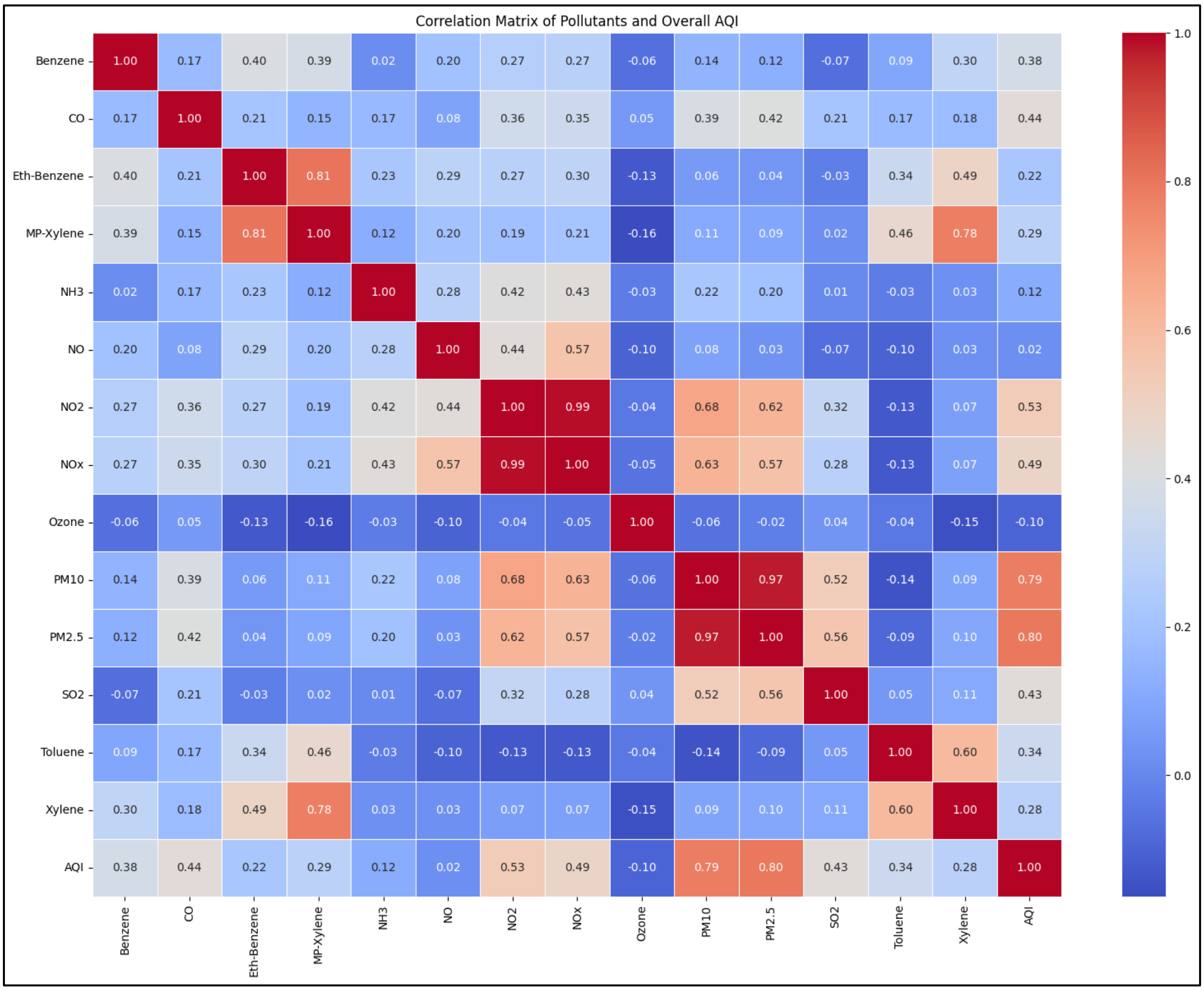


Figure 8 Correlation Matrix of Pollutants and Overall AQI

The Random Forest Regressor model, trained on engineered features including temporal components (month, day of week) and lagged variables (1–3 days for AQI and key pollutants), demonstrated exceptional forecasting performance. On the test set (20% chronological split), the model achieved an R^2 of 0.992, MAE of 2.817, and MSE of 14.777, indicating that it explained over 99% of AQI variance with minimal error. Time-series comparison of actual versus predicted AQI (Figure 11) showed close alignment, with predictions capturing both trends and peaks, though slight underestimation occurred during extreme events, possibly due to unmodeled external factors like sudden weather changes. The scatter plot of actual versus predicted values (Figure 12) confirmed this accuracy, with points tightly clustered along the identity line (y = x), validating the model's robustness for short-term forecasting in Rajahmundry's context.

Explainable AI via SHAP further elucidated the model's decision-making. Global feature importance, visualized in a bar plot (Figure 9), identified PM10 as overwhelmingly dominant (mean absolute SHAP value far exceeding others), followed by PM2.5 and toluene. This emphasizes particulate matter's pivotal role, aligning with EDA correlations and literature on urban air quality where PM10 from dust and emissions significantly impacts health indices. The SHAP dependence plot for PM10 (Figure 10) depicted a near-linear positive relationship: as PM10 concentrations increased from ~50 to 200 µg/m^3, SHAP values rose from -50 to +100, indicating substantial upward pressure on predicted AQI. Minimal scattering suggests limited interactions with other features, reinforcing PM10's independent influence. For publication-ready visuals, enhanced versions of the SHAP summary bar plot (Figure 13) and dependence plot (Figure 14) were generated with improved aesthetics, such as larger fonts and gridlines, to facilitate clear interpretation in academic contexts.

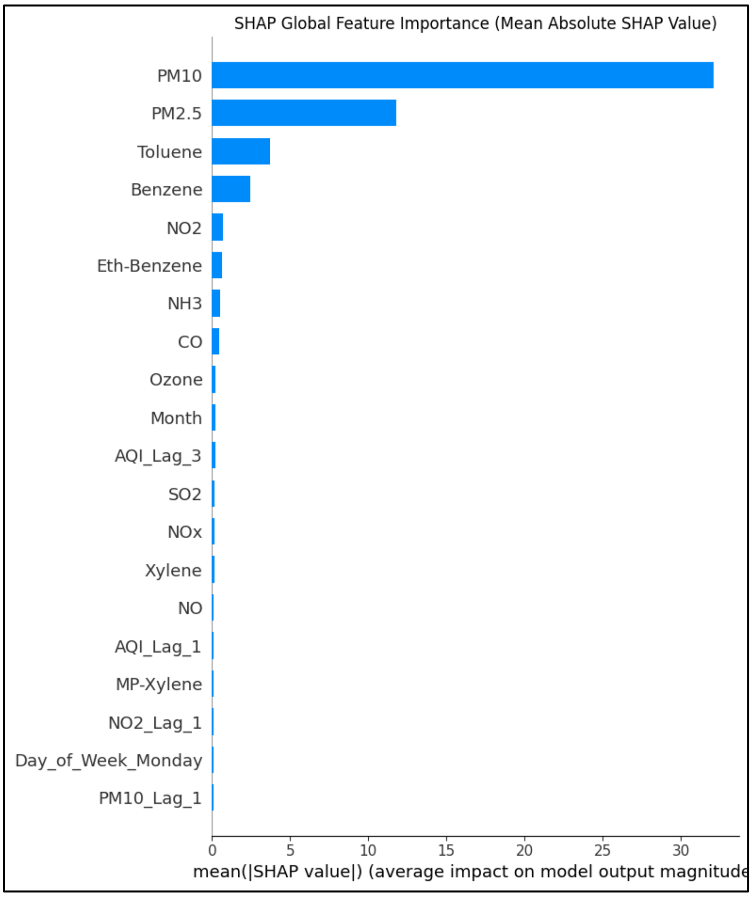


Figure 9 SHAP Global Feature Importance (Mean Absolute SHAP Value). A bar plot showing feature impacts on predictions. The x-axis is mean absolute SHAP value, and the y-axis lists features (e.g., PM10 highest), for global interpretability

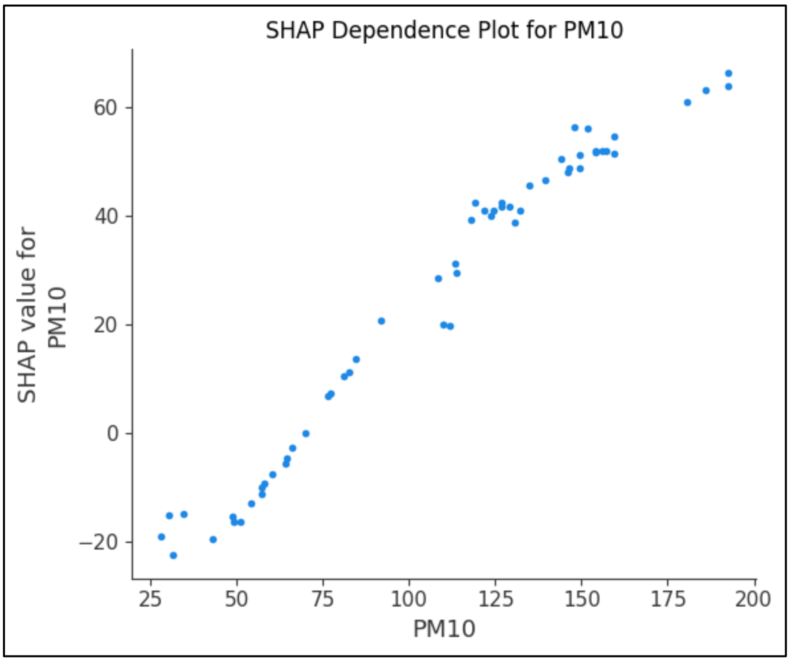


Figure 10 SHAP Dependence Plot for the Most Impactful Feature (PM10). A scatter plot with SHAP values on the y-axis and feature values (PM10) on the x-axis, illustrating positive relationships without strong interactions

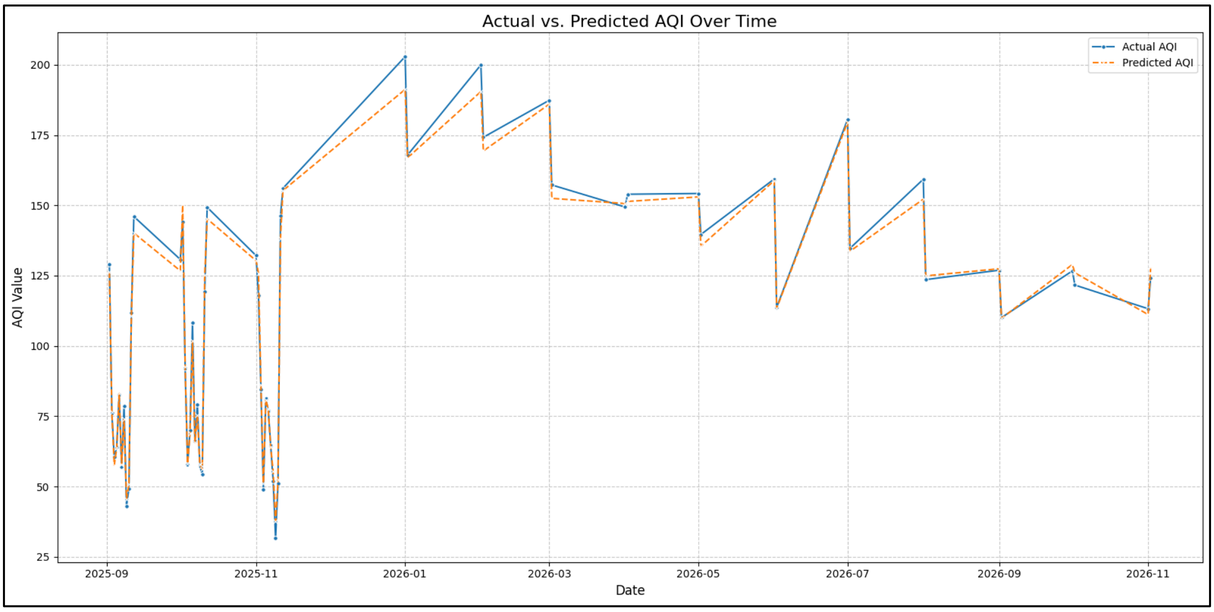


Figure 11 Actual vs. Predicted AQI Over Time. A time-series line plot with actual AQI (solid) and predicted (dashed with markers). The x-axis is date, and the y-axis is AQI, demonstrating model alignment

These results highlight Rajahmundry's air quality challenges, with mean AQI in the moderate category but frequent unhealthy episodes driven by particulates. The high model accuracy ($R^2 > 0.99$) outperforms many reported studies in Indian cities (e.g., R^2 ~0.85 for LSTM in Delhi), likely due to the ensemble nature of Random Forest and inclusion of lagged features capturing autocorrelation. SHAP insights prioritize PM10 mitigation, such as stricter controls on construction and traffic, for policy impact. Limitations include potential overfitting (though mitigated by chronological split) and reliance on CPCB data, which may have gaps; future work could integrate satellite or IoT data for enhanced resolution. Overall, this framework advances explainable forecasting, aiding sustainable urban management in mid-sized Indian cities.

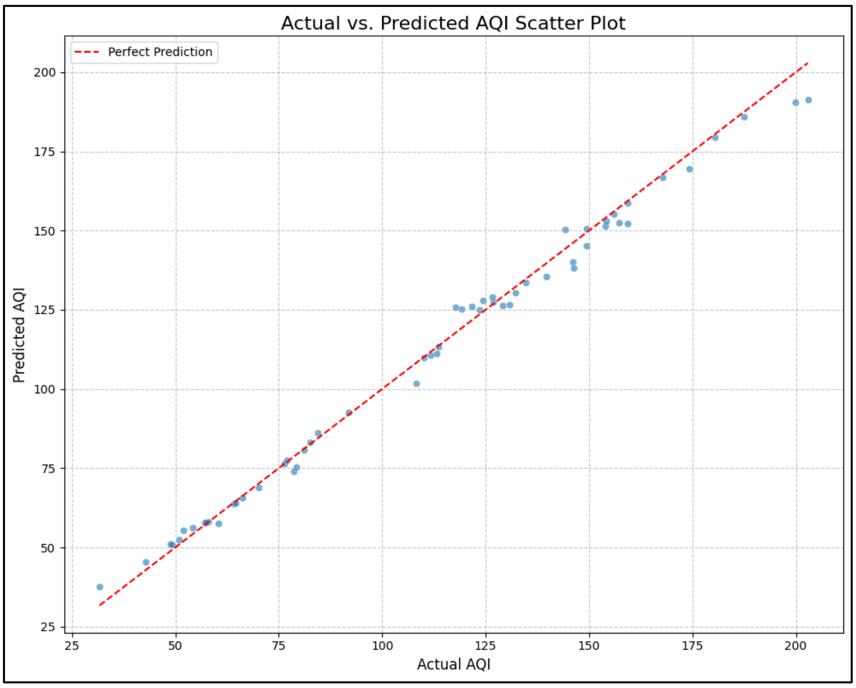


Figure 12 Actual vs. Predicted AQI Scatter Plot. A scatter plot with points (alpha=0.6) and a red dashed identity line. The x-axis is actual AQI, and the y-axis is predicted AQI, confirming high accuracy

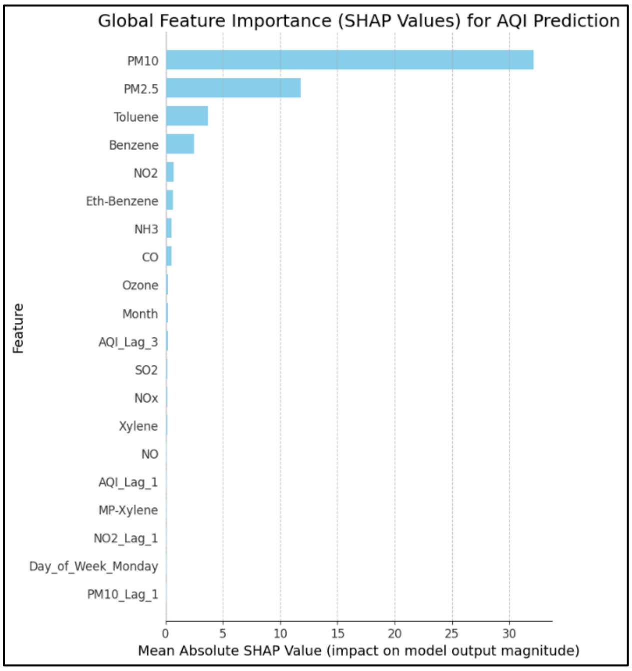


Figure 13 Global Feature Importance (SHAP Values) for AQI Prediction. An enhanced bar plot (skyblue bars) for publication, with x-axis as mean absolute SHAP value and y-axis as features, emphasizing top contributors like PM10

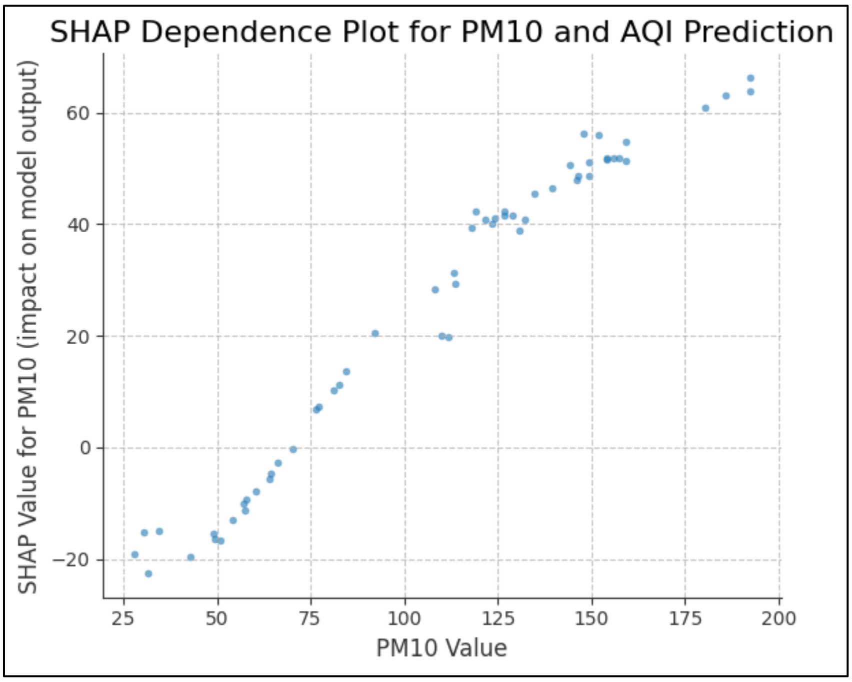


Figure 14 SHAP Dependence Plot for PM10 and AQI Prediction. An enhanced scatter plot (alpha=0.6, blue points) showing SHAP values (y-axis) versus PM10 values (x-axis), with a near-linear positive trend

Table 1 summarizes key numerical highlights from the analysis, including data preprocessing outcomes, descriptive statistics for sub-indices and AQI, outlier detection rates for select pollutants, and model performance metrics. These values are derived from the processed dataset after handling missing values (initially 1,115 in pollutant concentrations, reduced to 780 via median imputation) and exclusions, resulting in a final cleaned dataset of 4,004 entries.

Table 1 Key numerical highlights from the analysis

Category	Highlight	Value
Data Preprocessing	Initial Missing Values (PollutantValue)	1,115
Data Preprocessing	Missing Values After Imputation (PollutantValue)	780
Data Preprocessing	Cleaned Dataset Entries	4,004
Outlier Detection (IQR Method)	Outliers in SO2 (%)	26.22
Outlier Detection (IQR Method)	Outliers in NO (%)	13.64
Pollutant Sub-Indices	Minimum	0
Pollutant Sub-Indices	Maximum	417.8
Pollutant Sub-Indices	Mean	11.88
Overall AQI	Minimum	14.25
Overall AQI	Maximum	417.8
Overall AQI	Mean	87.17
Model Performance (Test Set)	Mean Absolute Error (MAE)	2.817
Model Performance (Test Set)	Mean Squared Error (MSE)	14.777
Model Performance (Test Set)	R-squared (RÂ²)	0.992
SHAP Insights	Most Influential Feature	PM10
SHAP Insights	Second Most Influential Feature	PM2.5

6. Conclusion

This work successfully established a robust, data-driven framework for predicting the Air Quality Index (AQI) in Rajahmundry, a rapidly developing urban center in Andhra Pradesh. By utilizing historical pollutant and meteorological data from January 2024 to February 2026, the study addressed a critical gap in localized air quality modeling for mid-sized Indian cities that often lack the specialized attention given to major metropolises.

The implementation of a Random Forest Regressor demonstrated exceptional predictive power, achieving a coefficient of determination (R2) of 0.992 on a chronological test set. This high level of accuracy, complemented by a low Mean Absolute Error (MAE) of 2.82, indicates that the model can effectively capture the complex, non-linear relationships between various atmospheric pollutants and weather conditions. Unlike traditional statistical methods, this ensemble learning approach proved highly resilient in handling the seasonal fluctuations and episodic pollution spikes characteristic of the region.

A crucial contribution of this work is the integration of Explainable Artificial Intelligence (XAI) through SHapley Additive exPlanations (SHAP). While many high-performing models operate as “black boxes,” the SHAP analysis provided much-needed transparency by identifying the primary drivers of air pollution in Rajahmundry. The results pinpointed particulate matter (PM10 and PM2.5) as the most significant influencers of the city's AQI. Specifically, PM10 emerged as the dominant factor, suggesting that local issues such as construction dust, road traffic, and industrial emissions are the principal contributors to poor air quality episodes.

The study also highlighted significant temporal patterns, with air quality reaching its most concerning levels during the winter months, likely due to atmospheric inversions that trap pollutants near the surface. Although the average AQI was categorized as “moderate” (87.17), the frequent occurrence of severe spikes highlights the urgent need for targeted interventions.

This study provides a scalable and transparent forecasting tool that can assist local authorities and urban planners in Rajahmundry. By offering accurate short-term predictions and clearly identifying the pollutants most responsible for health risks, this framework supports the development of more effective public health advisories and pollution control strategies. Future research may further enhance these capabilities by incorporating real-time IoT sensor data and satellite-based observations to improve spatial resolution and long-term generalizability.

Compliance with ethical standards

Acknowledgements

The authors sincerely acknowledge the Central Pollution Control Board (CPCB), Government of India, for providing air quality monitoring data through its online portal, which formed the basis for the analysis presented in this study.

The authors also wish to extend their sincere gratitude to the Government College (Autonomous), Rajahmundry, for providing the institutional support and academic environment that were instrumental to this study.

Finally, we thank our colleagues for their insightful discussions and the anonymous reviewers for their constructive feedback, which significantly improved the quality of this manuscript.

Disclosure of conflict of interest

The authors declare that they have no known financial or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] "India Air Quality Index (AQI) and Air Pollution information IQAir."

[2] R. T, T. Byragi Reddy, and C. Ramana, "PM10, PM2.5, SO2 AND NO2 concentrations at different locations in East-Godavari District, Andhra Pradesh, India," RESEARCH REVIEW International Journal of Multidisciplinary, vol. 6, Feb. 2021, doi: 10.31305/rrijm.2021.v06.i04.011.

[3] "National Clean Air Programme (NCAP) to improve air quality in 131 cities - PIB."

[4] "Air Pollution - World Health Organization (WHO) - India."

[5] A. Pootheri, P. Chinnaiyan, N. Ashok, and R. Saraswathy, "Association between ambient air pollution and increased risk of respiratory diseases in Vellore and Ranipet, Tamil Nadu, India: A retrospective study," Environmental Pollution, vol. 385, p. 127051, 2025, doi: https://doi.org/10.1016/j.envpol.2025.127051.

[6] M. Karmoude et al., "Machine learning for air quality prediction and data analysis: Review on recent advancements, challenges, and outlooks," Science of The Total Environment, vol. 1002, p. 180593, 2025, doi: https://doi.org/10.1016/j.scitotenv.2025.180593.

[7] A. Yenkikar, V. P. Mishra, M. Bali, and T. Ara, "Explainable forecasting of air quality index using a hybrid random forest and ARIMA model," MethodsX, vol. 15, p. 103517, 2025, doi: https://doi.org/10.1016/j.mex.2025.103517.

[8] S. Singh, M. Kumar, V. Sengar, A. Kumar, K. Abhishek, and B. M. A. Shafeeq, "Ensemble learning for air quality index prediction: integrating gradient boosting, XGBoost, and stacking with SHAP-based interpretability," Sci. Rep., 2026, doi: 10.1038/s41598-026-39232-w.

[9] Z. Xu, H. Zhang, A. Zhai, C. Kong, and J. Zhang, "Stacking Ensemble Learning and SHAP-Based Insights for Urban Air Quality Forecasting: Evidence from Shenyang and Global Implications," Atmosphere (Basel)., vol. 16, no. 7, 2025, doi: 10.3390/atmos16070776.

[10] G. L, N. S. Sriya, M. Sowmya, A. R. H. Lakshmi, and R. Amirtharajan, "Explainable AI for urban air quality: SHAP interpretation of stacked ensemble AQI forecast," Theor. Appl. Climatol., vol. 156, no. 10, p. 533, Sep. 2025, doi: 10.1007/s00704-025-05741-3.

[11] H. Tian, H. Kong, and C. Wong, "A Novel Stacking Ensemble Learning Approach for Predicting PM2.5 Levels in Dense Urban Environments Using Meteorological Variables: A Case Study in Macau," Applied Sciences, vol. 14, no. 12, 2024, doi: 10.3390/app14125062.

[12] B. Ghosh, H. Barman, and P. Padhy, “Trends of air pollution variations during pre-Diwali, Diwali and post-Diwali periods and health risk assessment using HAQI in India,” Discover Environment, vol. 2, Feb. 2024, doi: 10.1007/s44274-024-00082-w.

[13] M. W. Jalali, B. Saidi, H. Farahmand, M. A. R. Panah, and E. N. Saruhan, “Scalable AI-driven air quality forecasting and classification for public health applications,” Discover Atmosphere, vol. 3, no. 1, p. 25, 2025, doi: 10.1007/s44292-025-00052-8.

[14] S. Zeeshan and M. A. I. Malik, “AI-Driven Analysis of Meteorological and Emission Characteristics Influencing Urban Smog: A Foundational Insight into Air Quality,” Gases, vol. 6, no. 1, 2026, doi: 10.3390/gases6010010.

[15] G. L, N. Sriya, M. Sowmya, A. Lakshmi, and R. Amirtharajan, “Explainable AI for urban air quality: SHAP interpretation of stacked ensemble AQI forecast,” Theor. Appl. Climatol., vol. 156, Feb. 2025, doi: 10.1007/s00704-025-05741-3.

[16] G. Sharma, S. Khurana, N. Saina, Shivansh, and G. Gupta, “Comparative Analysis of Machine Learning Techniques in Air Quality Index (AQI) prediction in smart cities,” International Journal of System Assurance Engineering and Management, vol. 15, pp. 1–16, Feb. 2024, doi: 10.1007/s13198-024-02315-w.

[17] M. Karmoude et al., “Machine learning for air quality prediction and data analysis: Review on recent advancements, challenges, and outlooks,” Science of The Total Environment, vol. 1002, p. 180593, 2025, doi: https://doi.org/10.1016/j.scitotenv.2025.180593.

[18] P. Mohandas, S. P, and S. .R, “AI-Driven Hybrid Deep Learning Approach for Predicting Air Quality in India Using AlexNet-TabNet Architecture,” Feb. 2025, pp. 1–6. doi: 10.1109/ICCTDC64446.2025.11158025.

[19] Z. Xu, H. Zhang, A. Zhai, C. Kong, and J. Zhang, “Stacking Ensemble Learning and SHAP-Based Insights for Urban Air Quality Forecasting: Evidence from Shenyang and Global Implications,” Atmosphere (Basel)., vol. 16, p. 776, Feb. 2025, doi: 10.3390/atmos16070776.