Journal of Information Systems Engineering and Management
2025, 10(54s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

RegBoost: A Novel Hybrid Methodology for Enhanced Spatiotemporal Forecasting of PM2.5 in Gurugram (2019-2023) Using Multi-Source Ground-Based, Meteorological, and Sentinel-5P Data

Anuradha Dhull, Duiena Rai, Tripti Sharma, Avi Aneja, Maanya Johri
The NorthCap University, Gurugram, Haryana, India

ARTICLE INFO

Received: 28 Dec 2024

Revised: 18 Feb 2025

Accepted: 26 Feb 2025

ABSTRACT

Introduction: This research addresses Particulate Matter (PM2.5) pollution dynamics in Gurugram, an integral part of the National Capital Region (NCR) of India, it holds infamy for possibly the most notorious pollution levels for the past decade, thus rendering it a priority research and intervention area. This research mainly focuses on the Gurugram air quality data during 2019 to 2023.

Objectives: In this research study, a novel hybrid methodology, referred to in this research paper as 'RegBoost', for enhanced spatiotemporal forecasting of PM2.5 has been developed and further evaluated against the recent research within this domain. The existing models face challenges related to data accuracy, missing values, scarcity of data and unreliable data sources, which limit their predictive performance.

Methods: In this research, through a multi-faceted approach, we employ the Central Pollution Control Board (CPCB), National Aeronautics and Space Agency (NASA) and Sentinel 5-Precursor (Sentinel 5P) data to bolster our accuracy. The novel frameworks employed in this research study includes hybrid imputation technique that fuses Matrix Factorisation and K-Nearest Neighbors (Hybrid MF+KNN), hybrid predictive model fusing Ridge Regression and XGBoost (RR+XGBoost). RegBoost apart from the incorporation of the hybrid imputation and preprocessing steps, it uses the combined methodology of the aforementioned Ridge Regression and XGBoost along with the data from the CPCB, NASA and Sentinel 5P. The performance of this methodology is benchmarked against current literature and previous studies to assess its comparative efficacy. This research further provides the most comprehensive and up-to-date analysis within Gurugram which is still a major problem area due to rising pollution levels and limited solutions. This research seeks to fill the gap of spatiotemporal forecasting as most of the models are limited to temporal forecasting or spatial forecasting for a single location. The methodology is specifically designed to address the challenges in spatiotemporal air quality forecasting, including data inconsistencies and the complex interplay of various pollution factors.

Results: This research was able to achieve a remarkable improvement in the accuracy of air quality forecasting due to the incorporation of the novel hybrid methods. The hybrid imputation method significantly reduced the data gaps, which previously affected the predictive performance. The RR+XGBoost model demonstrated superior performance in capturing complex patterns and relationships within the spatiotemporal data, leading to more precise and reliable PM2.5 predictions. Our models also show consistency across different seasonal variations and during stubble burning periods, proving their robustness.

Journal of Information Systems Engineering and Management
2025, 10(54s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Conclusions: The RegBoost methodology offers a robust and effective solution for enhanced spatiotemporal PM2.5 forecasting in urban environments. Its ability to integrate diverse data sources and handle missing values effectively positions it as a valuable tool for environmental monitoring, policy-making, and public health initiatives. The insights gained from this study contribute significantly to the understanding of air quality dynamics in Gurugram and provide a scalable framework for similar pollution challenges globally. Future work will focus on integrating real-time data streams and exploring the applicability of RegBoost in other geographic regions with varying pollution characteristics.

Keywords: Air Quality Prediction, Spatiotemporal forecasting, time series analysis, machine learning, PM2.5, Central Pollution Control Board, SENTINEL 5P, Pollutants

INTRODUCTION

The unrelenting growth of particulate matter 2.5 (PM2.5) in urban settings emerges as a prime determinant of public health and ecological balance, a phenomenon starkly visible in the burgeoning metropolitan of India. Gurugram stands as a prime example of a rapidly developing city within the National Capital Region of Delhi. This area combines various adverse influences, including a rapid rate of industrialization, increasing road traffic, ongoing construction work, changing climatic conditions, and influence from pollution sources across borders [1]. Gurugram, although unfortunately, holds infamy for possibly the most notorious pollution levels, thus rendering it a priority research and intervention area (Figure 1). It is first vital to appreciate the deep-set effects of extremely high PM2.5 concentrations. These effects include a striking rise in cardiovascular and respiratory diseases and direct correlation with rising mortality figures. Those vulnerable groups-duration: children, the aged, and diseased people-likely sustains the brunt. Likewise, PM2.5 inflicts effects beyond the human health domain, causing ecological disruption, reduced visibility into the atmosphere, and aggravating climate change. The atmospheric meteorological conditions evoke a rich interplay acting subsequently on the distribution and concentration of PM2.5 [2]. Therefore, it is necessary to develop accurate PM2.5 prediction models and effective interpretive frameworks to formulate useful mitigation strategies.

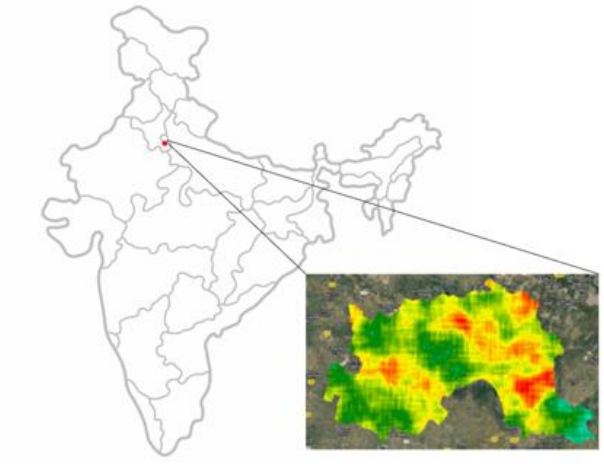

Figure 1: Research Study Area - Gurugram

OBJECTIVES

The following research questions are addressed in this paper:

- RQ1: What is the multivariate meteorological forcing of PM2.5 spatiotemporal heterogeneity?

Journal of Information Systems Engineering and Management
2025, 10(54s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

- RQ2: How do merged and remotely sensed datasets differ in primary pollutant identification?
- RQ3: What is the coupled impact of atmospheric stratification and aerosol loading on PM2.5 dynamics?
- RQ4: Do novel techniques optimize data precision and modeling accuracy?

Inspired form the research question, The following contribution has been presented in the research study:

1. We have integrated data from three distinct sources (CPCB, NASA, Sentinel-5P) in Phase 1 of proposed methodology, addressing data sparsity and providing a richer dataset for improved PM2.5 prediction.
2. We applied a novel MF+KNN imputation technique in Phase 2 of the proposed pipeline to handle missing values, ensuring data completeness and enhancing model accuracy.
3. We utilized a combined RR+XGBoost prediction model in Phase 3 of the proposed model to capture both linear and non-linear relationships, resulting in more accurate PM2.5 predictions.
4. We also formulated meteorological-based indices to extract meaningful features from weather data, providing targeted inputs for improved prediction precision.

The rest of the paper is organized as follows:

- Section 2 contains the systematic literature review of the previous studies pursued in this domain.
- Section 3 is divided into two parts, Section 3.1 deals with data collection and data description and Section 3.2 deals with the different phases of our proposed methodology "RegBoost".
- Section 4 is divided into six sub sections - 4.1 deals with evaluation of Hybrid MF+KNN accuracy, section 4.2 deals with evaluation of Hybrid RR+XGBoost accuracy, section 4.3 deals with seasonality of the pollutants, 4.4 deals with stubble burning period analysis, section 4.5 deals with comparison of temporal and spatial granularity of the data and section 4.6 deals with the comparison of the novel models and current research studies.
- Section 5 contains the conclusion and future scope of our study.

This research of ours proposes a prediction and analytical methodology (RegBoost) of PM2.5 concentrations in Gurugram for the period 2019-2023. To fulfill the requirement of accurate location specific air quality forecasts, we have used meteorological data from NASA POWER [31], air quality observations from Central Pollution Control Board (CPCB) [32] and early Sentinel-5P [33] retrievals. In this paper, we have proposed a hybrid machine learning model named "RegBoost" which uses two powerful algorithms, Ridge Regression (RR) and XGBoost (RR+ XGBoost). Hybrid imputation strategies (Hybrid MF+KNN) are employed to deal with missing data, and feature engineering in conjunction with machine learning and deep learning algorithms.

METHODS

3.1 Data Source and Data Description

Data from the Central Pollution Control Board (CPCB) and National Aeronautics and Space Administration (NASA) Power were used in our study. The CPCB data, which consists of 7,304 rows, 26 columns, and 189,904 data points, includes pollutants as well as meteorological factors including air temperature, wind speed, and relative humidity. The NASA Power dataset, which has 1,826 rows, 15 columns, and 27,390 data points, offers meteorological information such as temperature, wind speed, and pressure. It was merged on the following attributes, year, month, and day. The merged dataset has 7,304 rows and 41 columns. Sentinel-5P satellite data from Google Earth Engine (NO2, CO, O3, SO2) will be added later for enhanced spatial coverage. This merged dataset containing 3,00,778 data points is the basis for analyzing air pollution behaviors and meteorological relationships.The detailed data description is provided in Phase 1 (Refer section 3.2.1) of our proposed methodology.

The Sentinel-5P satellite image (Figure 2) illustrates nitrogen dioxide (NO_2) concentrations over Gurugram processed using EO Browser. The color gradient reveals high NO_2 accumulation in certain areas, likely attributed to vehicular emissions, industrial activities, and urban pollution. In contrast, lower concentration regions indicate better air quality.

Journal of Information Systems Engineering and Management
2025, 10(54s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

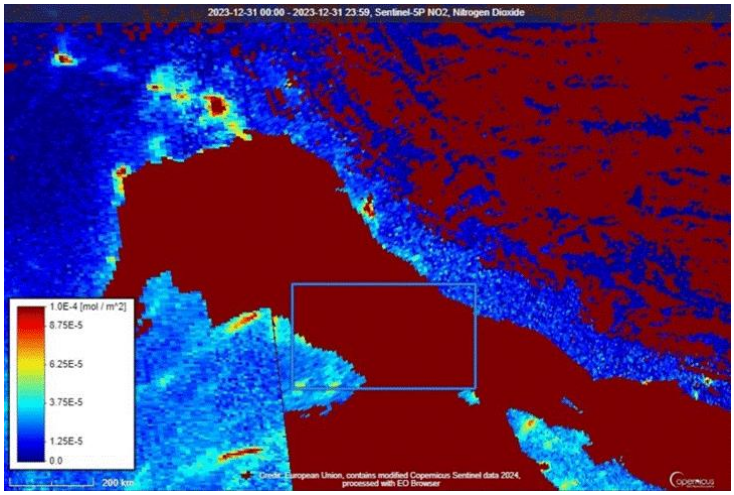

Figure 2: Sentinel-5P Derived NO_2 Concentration Over Gurugram

3.2 Proposed Methodology

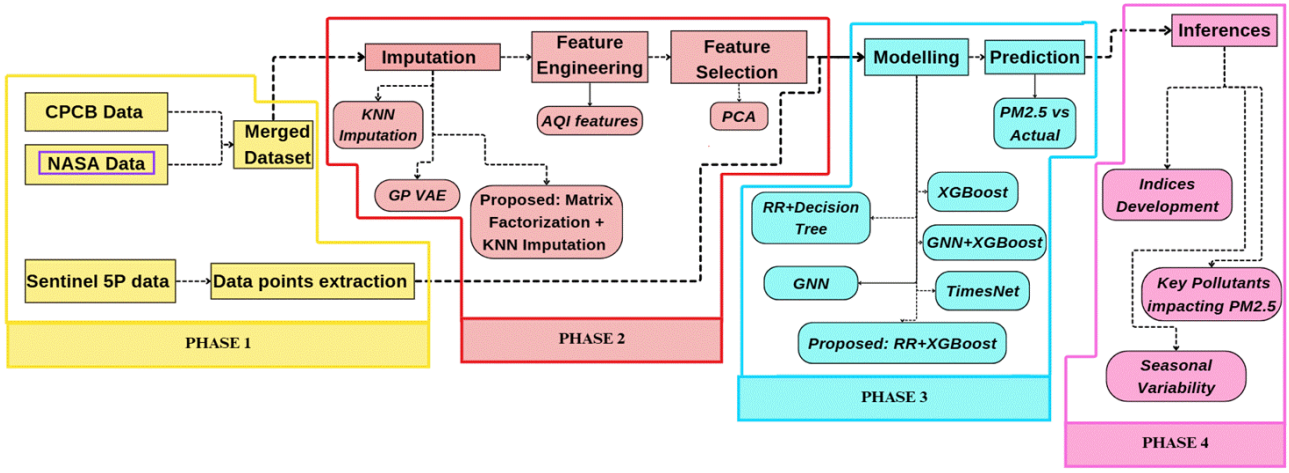

Figure 3: Operational layout of the proposed model “RegBoost”

3.2.1 Phase 1: Data Collection

This first phase of our research combines NASA Power meteorological data with Central Pollution Control Board (CPCB) air quality data to analyze pollution trends and their correlation with weather as shown in Figure 4. Timestamps were standardized, CPCB data from several stations was aggregated, and a "Station ID" feature was created. Any missing or incorrect dates were removed. The NASA Power dataset was transformed to match the CPCB data and then integrated using Year, Month, and Day. Sentinel-5P satellite data (NO_2, CO, SO_2 and O_3) will be added later for better geographical coverage. This integrated dataset serves as the foundation for the further phases of our study.

Journal of Information Systems Engineering and Management
2025, 10(54s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

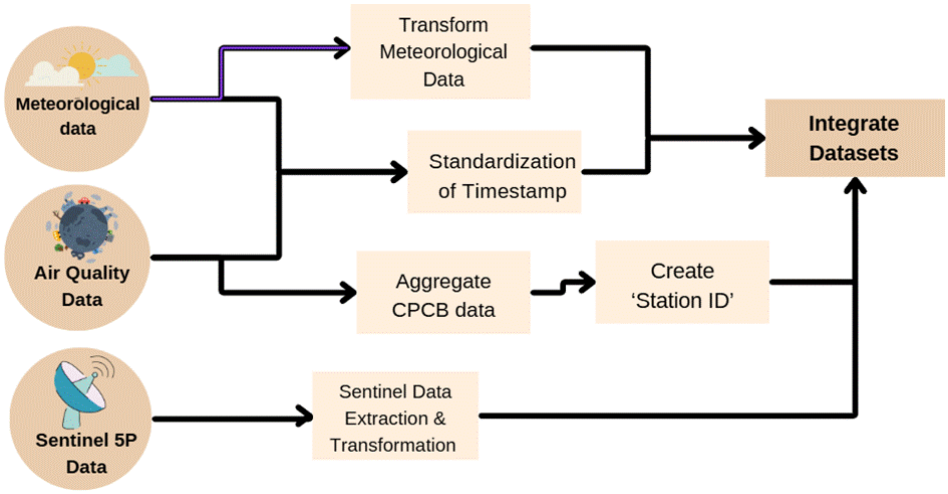

Figure 4: RegBoost Data Collection Pipeline: Meteorological, Air Quality, and Sentinel 5P

3.2.2 Phase 2: Data Preprocessing

One of the key challenges that we faced after Phase 1 was the presence of missing values in our dataset. A combined strategy leveraging KNN and matrix factorization was adopted to combat this challenge. The Matrix Factorization technique identified hidden linkages in the aggregate information, through the extraction of the latent patterns.[14][15] This research further employed KNN imputation to improve the imputed values produced by the Matrix Factorization procedure by using local neighbourhood knowledge. The mathematical representation of KNN imputation is as follows:

$$d(x, y) = \sqrt{((x_i - y_i)^2)} \quad (1)$$

where x and y are data points, and x_i and y_i are their respective features. As we compared single-stage KNN with the hybrid imputation approach, which addressed both global and local dependencies, we found that hybrid imputation approach greatly enhanced imputation quality as shown by higher R-squared and lower MSE. Feature engineering enhanced predictive capability by adding a 'Station ID' column for AQI categorization and inter-agency variances (Figure 5). Next in our research study, we used PCA, which reduced the multicollinearity and dimensionality of the 30 original features.[16] The basic mathematical concept underlying PCA is the solution of the eigenvalue problem for the data's covariance matrix:

$$Cov(X)v = \lambda v \quad (2)$$

where λ is the matching eigenvalue that indicates the variance explained by that principal component, v is an eigenvector that is the principal component, and Cov(X) is the covariance matrix of the data X.

Journal of Information Systems Engineering and Management
2025, 10(54s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

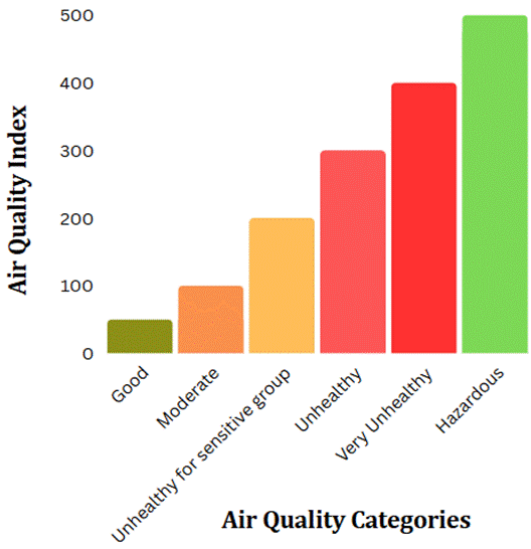

Figure 5: AQI scale developed through feature engineering, with custom-defined ranges

3.2.3 Phase 3: Modelling & Prediction

This research leveraged preprocessed data collected during Phase 1 and Phase 2 to put together predictive models for PM2.5 levels as discussed in this, Phase 3 of our RegBoost. Numerous approaches to modeling are implemented and assessed, leading to the determination of the most efficient functioning model (Figure 6). Alongside predicting PM2.5 concentrations, we intend to determine each of the four prominent contaminants which influence PM2.5 pollution levels, establishing the base for our RegBoost.

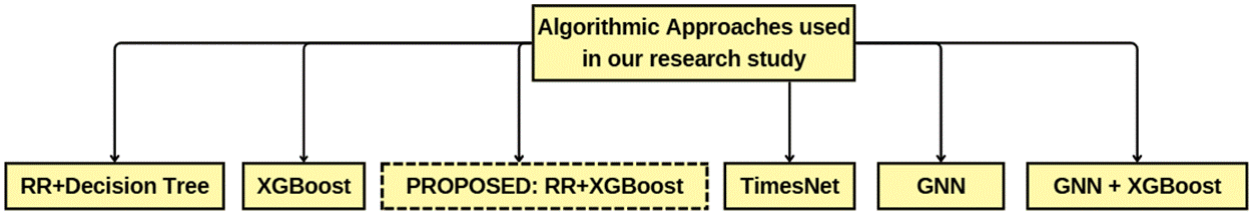

Figure 6: Algorithmic Approaches Employed in the RegBoost Framework

Three modeling techniques were employed in this study to predict PM2.5 levels: the integration of Ridge Regression and Decision Tree, XGBoost, and integration of Ridge Regression and XGBoost. Ridge Regression was employed to address multicollinearity within the data, while non-linear impacts were represented using the Decision Tree (max_depth = 9) and the XGBoost (n_estimators=100 and learning rate of 0.1). The Ridge Regression + XGBoost model, illustrated in Figure 7, performed Ridge Regression for feature engineering and prediction using XGBoost. After the model step, model-predicted PM2.5 was used to compare actual CPCB. Hybrid model Ridge Regression + XGBoost was effective, where there was very high positive correlation and close predicted and actual values clusters, even though outliers highlighted unusual pollution amounts. This mixed model performed better than the other two models. This research also extended our RegBoost framework with state-of-the-art models: TimesNet, GNNs, and GNN+XGBoost.[17] TimesNet was used to capture intricate temporal relationships in PM2.5 data, detecting possible cyclical patterns. GNNs were used to embed spatial correlations between monitoring stations in the form of a graph, allowing the network structure to be exploited. A GNN+XGBoost hybrid model was created to leverage GNNs for spatial feature extraction, combined with XGBoost for improved prediction, in an attempt to beat our current Ridge Regression + XGBoost performance.[18] These additions were designed to refine PM2.5 prediction and pollutant

Journal of Information Systems Engineering and Management
2025, 10(54s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

identification, advancing the RegBoost system. Besides PM2.5 predictions, the study also found the top four major sources of PM2.5 among the pollutants and selected them for further detailed analysis as described in section 4.

```
1.alpha = 1.0 // Regularization parameter for Ridge Regression
2. model_RR = RidgeRegression(alpha)
3.model_RR.fit(X, y) // Train Ridge Regression on all features (X) and PM2.5 (y)
4. X_rr_predictions = model_RR.predict(X) //Use Ridge Regression predictions as
input
5. n_estimators = 100
6. learning_rate = 0.1
7. model_XGB = XGBRegressor(n_estimators, learning_rate)
8. model_XGB.fit(X_rr_predictions, y) // Train XGBoost on the RR predictions
9. X_test_rr_predictions = model_RR.predict(X_test)
10. y_pred = model_XGB.predict(X_test_rr_predictions) // Predict PM2.5 using
the trained XGBoost model and the RR predictions of the test set
11. return y_pred // Return the predicted PM2.5 values
```

Figure 7: Novel RR+XGBoost Pseudocode within the RegBoost Framework

3.2.4 Phase 4: Inferences of RegBoost

Leveraging Indices: To gain a comprehensive understanding of air pollution dynamics, we have created multiple indices into our PM2.5 forecasting model, as shown in Figure 8. While traditional approaches focus solely on PM2.5 concentrations, this study recognizes that meteorological conditions and pollutant interactions play a crucial role in air quality variations.[19] This research have developed the Atmospheric Stability Index (ASI) to assess whether or not pollutants disperse or settle, as stable weather leads to higher pollution levels and unstable weather promotes dispersion. The ASI levels, their atmospheric conditions and PM2.5 Impact is shown in Table 2. The relationship between ASI and PM2.5 concentrations is shown in Table 3. Although PM2.5 alone is not a full indicator of overall air pollution, we have developed the Combined Pollution Index (CPI), which provides a more general assessment based on multiple pollutants. This helps in the determination of primary sources of pollution and improving air quality management.[20] Table 3 classifies CPI levels with air quality measurements and associated health hazards. This research has also developed the Meteorological Stress Index (MSI) to study the effect of weather conditions such as temperature, humidity, and wind speed on PM2.5 concentrations. By integrating these indices, we have a better analysis of pollution patterns, leading to more efficient forecasting and control measures, as shown in Table 4.

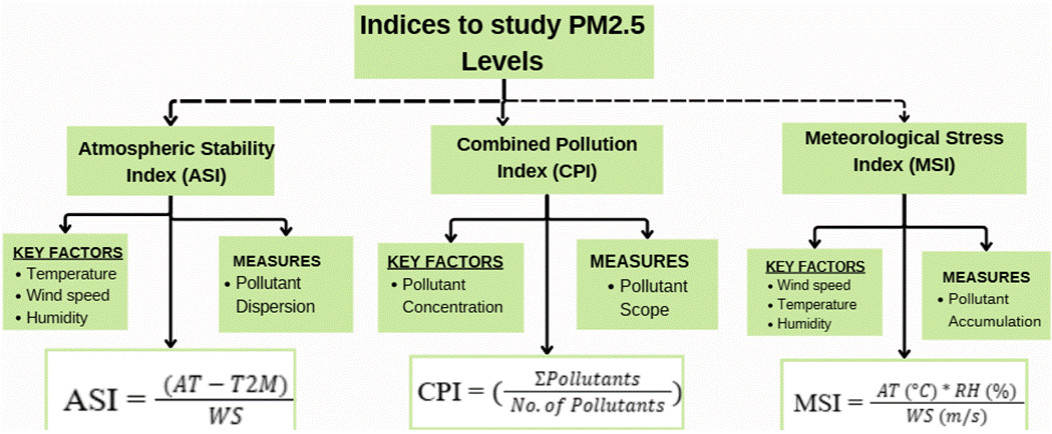

Figure 8: Creation of integrated indices for multifactor influence on PM2.5 Concentration

Journal of Information Systems Engineering and Management
2025, 10(54s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

ASI Levels	Atmospheric Conditions	PM2.5 Impact
High	Stable (Temperature Inversion)	Accumulation (Inhibited Vertical Mixing)
Low	Unstable (Enhanced Vertical Mixing)	Dispersion (Improved Air Quality)
Moderate	Neutral	Variable (Susceptible to other factors)

Table 2: ASI levels, their atmospheric conditions and PM2.5 Impact

CPI Level	Value	Air Quality	PM2.5 Impact	Health Risk
Low	<1.0	Good	Minimal	Few
Moderate	1.0 to 2.5	Somewhat Elevated	Moderate	Potential Respiratory Issues
High	>2.5	Extreme Pollution	Significant	Major Health Hazards

Table 3: CPI Levels and their range along with the PM2.5 Impact

MSI Level	AT (°C)	RH (%)	WS (m/s)	PM2.5 Impact	Meteorological Conditions
High	High	High	Low	Increased	Hot, Humid, Calm
Moderate	Moderate	Moderate	Moderate	Variable	Warm, moderately humid, light to moderate wind
Low	Low	Low	High	Decreased	Cool, dry, windy

Table 4: MSI levels and other parameters impacting PM2.5

Seasonal Influence on Pollution: Weather variability influences the interaction between ASI and MSI, leading to high seasonal dependency in Gurugram's air quality. In the spring (March–April) season, PM2.5 concentrations are generally moderate with intermittent fluctuations due to moderate ASI and low to moderate MSI. High ASI and moderate to high MSI create moderate PM2.5 with dust event increases throughout the pre-monsoon (May–June) period. Because of the high MSI and rapid ASI reduction caused by rainfall, PM2.5 levels are lowest during the monsoon season (July–September). PM2.5 rises during the post-monsoon (October–November) when ASI rises and MSI falls.[21] The fall season (late November to early December) has low MSI and high ASI, with PM2.5 levels often beyond dangerous thresholds (Figure 9). Winter (December–February) exhibits peak ASI and persistently low MSI,

Journal of Information Systems Engineering and Management
2025, 10(54s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

intensifying pollution episodes and resulting in the highest PM2.5 and worst air quality. These seasonal dynamics are crucial for developing effective mitigation strategies.

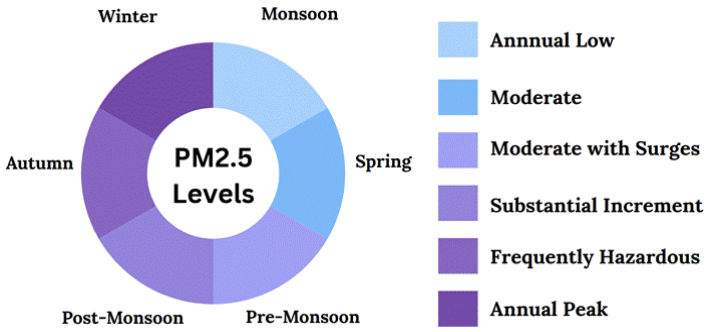

Figure 9: Seasonal PM2.5 Concentration Trends

Key Pollutants and Performance Evaluation: RegBoost shows that CO, O_3, NO_2, and SO_2 are the major pollutants that cause PM2.5 in Gurugram and the correlations between the variables are very high. In order to improve the analysis, this research compared temporal data from CPCB and spatial data from Sentinel, using the strengths of each. As can be seen from Table 2, CPCB data gave real time and precise information at the local level while Sentinel data was useful in identifying long term trends and regional pollution hotspots.[22] This underscores the importance of combining the two approaches for air quality monitoring. To check the accuracy of the models, key statistical metrics were used. Root Mean Square Error (RMSE) was used to measure prediction errors and lower values are better. Mean Absolute Error (MAE) gave a less sensitive and stable assessment, than using mean error. Finally, R-squared (R^2) was used to determine the ability of the model to explain PM2.5 variability.

RESULTS

This section deals with the results of our analysis of air pollution, compares data sources and seasons, evaluates the performance of data refinement and modeling tools, and concludes with a comparison of our proposed technique to other research.

4.1 The Efficiency of Imputation and Data Transformation

We used a combination of K-Nearest Neighbours (KNN) and Matrix Factorization (MF) for PM2.5 data refinement, which resulted in enhanced data sufficiency and precision (R^2 = 0.838, MSE = 767.709) when juxtaposed with single-step KNN imputation (R^2 = 0.674, MSE = 1199.759). Figure 10 provides a brief overview of the correlation of PM2.5 with the top 10 transformed features. Moreover, model performance was boosted due to the feature engineering method of adding station ID and AQI grade. Also, to decrease the multicollinearity, PCA was added to enhance the analysis' strength.

Journal of Information Systems Engineering and Management
2025, 10(54s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

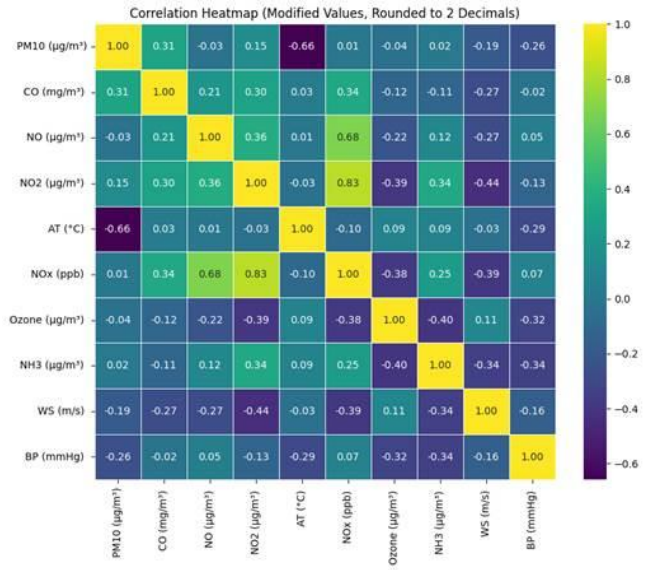

Figure 10: Heatmap of the Top 10 features impacting PM2.5

4.2 Predictive modeling and performance

This research compared three modeling methods for PM2.5 prediction: XGBoost, Ridge Regression combined with Decision Tree, and a new hybrid of Ridge Regression with XGBoost. The hybrid RR+XGBoost model showed outstanding performance, so we chose it for further investigation (Table 5). This research believes this improvement stems from Ridge Regression's effective feature selection, which helps prevent XGBoost from overfitting high-dimensional data.

Model	Mean Squared Error	Mean Absolute Error	R-squared	Adjusted R-squared
RR + Decision Tree	1212.218	21.781	0.672	0.663
XGBoost	1217.595	24.108	0.670	0.668
Proposed Methodology (RegBoost)	**564.907**	**15.246**	**0.947**	**0.943**
TimesNet	2188.757	32.819	0.407	0.402
GNN	1354.061	26.217	0.633	0.631
GNN+XGBOOST	1776.231	28.463	0.519	0.514

Table 5: Comparisons of baseline models with the proposed work

This research found that the RR+XGBoost model generalized better and was more accurate in prediction (Table 5). In addition, the scatter plot we created (Figure 11) shows how actual and predicted PM2.5 levels relate, confirming the model's validity as a pollution trend tracker. This research concluded that deep learning-powered models like TimesNet, GNN, and GNN+XGBoost did not produce good results, possibly because the dataset is very small in size.

Journal of Information Systems Engineering and Management
2025, 10(54s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Deep learning models prefer vast amounts of data to learn efficiently and generalize accordingly, and thus the restricted amount of data may have resulted in overfitting or incomplete learning of features.[23][24] On the contrary, our introduced methodology gave the best performance and proved that it is compatible enough to handle the provided size of data with higher predictive efficiency.

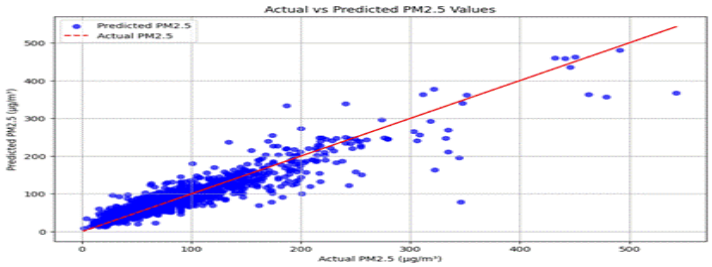

Figure 11: Actual vs. Predicted PM2.5 Values – RegBoost

4.3 Examining How Seasons Affect the Main Pollutants

This research identified CO, O3, NO2, and SO2 as the main pollutants affecting PM2.5 levels, in addition to PM2.5 itself. This research observed that CPCB data provides more accurate, real-time measurements of key pollutants, while Sentinel data excels at illustrating regional pollution trends, as shown in Table 6. To gain a more comprehensive understanding of air pollution dynamics, we incorporated the Atmospheric Stability Index (ASI), Combined Pollution Index (CPI), and Meteorological Stress Index (MSI) into our research. This research found, and as Table 7 demonstrates, that the monsoon season has the lowest PM2.5 concentrations of the year, likely due to a sharp decline in ASI and a high MSI.

Table 6: Seasonal Variation of Air Quality Indicators produced by the proposed methodology

Season	Period	ASI Trend	MSI Trend	PM2.5 Trend	Dominant Factors
Spring	March-April	Moderate	Low to Moderate	Fluctuating	Residual winter pollution, dust transport
Pre-Monsoon	May-June	High	Moderate to High	Moderate, with surges	Dust storms, high temperatures, industrial emissions
Monsoon	July-September	Sharp Decline	High	Annual Low	Rainfall, aerosol wet deposition
Post-Monsoon	October-November	Rising	Decreasing	Substantial Increase	Agricultural residue burning, stagnant air

Journal of Information Systems Engineering and Management
2025, 10(54s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Autumn	Late Nov-Early Dec	Elevated	Low	Frequently Hazardous	Transportation, industry, biomass burning
Winter	December-February	Peak	Persistently Low	Highest, Worst Quality	Industrial activity, vehicular emissions

This research examined the seasonality of PM2.5 concentrations in the Gurugram region and saw distinct trends influenced by the weather (Figure 12). Seasonal fluctuations were found to be highly correlated with ASI and MSI. Precisely, we saw that increased MSI during the winter was associated with increased PM2.5 concentrations, reflecting on the synergistic effects of emission sources as well as static air conditions.[25][26] This research have seen that the graph indicates very clearly a surge in PM2.5 concentrations in weeks 40-50 in Gurugram. The season is in accordance with the stubble burning season, and PM2.5 concentration is observed to be the highest at this time of the year.

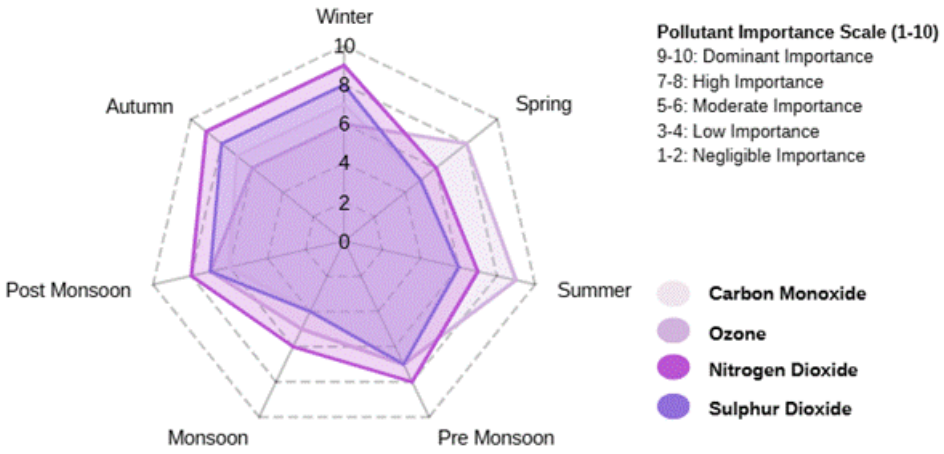

Figure 12: Seasonal Importance of Top 4 Pollutants Contributing to PM2.5 by the proposed methodology

4.4 Stubble Burning Period Analysis

The presence of flat maxima emissions signal towards episodic, intense emission pulses after agricultural crops have been harvested, thus revealing the seasonality of anthropogenic emissions resulting from the burning of crops. [27][28] These emissions are considered to be a major threat to the quality of air and in most cases, exceed the acceptable limits therefore, requiring immediate action to be taken. Superimposed on these, the annual difference of maxima height could also be due to the shifts in farming activities, changes in weather, or the success of the implementation of pollution control policies (Figure 13).

Journal of Information Systems Engineering and Management
2025, 10(54s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

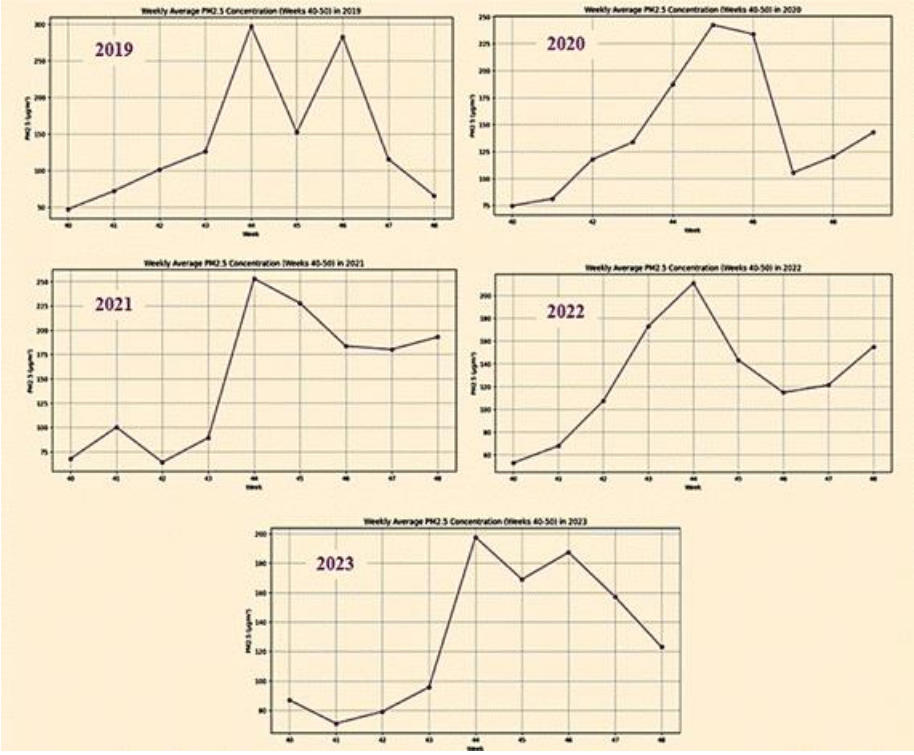

Figure 13: Time Series Analysis of PM2.5 During Stubble Burning Period (Week 40-50) generated by the proposed methodology

4.5 Comparing Temporal Resolution and Spatial Granularity in Pollutant Data

This research performed a performance comparison between the Sentinel-5P and CPCB datasets for CO, O3, NO2, and SO2 concentrations. This research observed that while CPCB data yielded higher R-squared values, Sentinel data exhibited reduced absolute errors in MSE, MAE, and RMSE (Table 7). This research views these results as indicating CPCB data contains greater explanatory capacity for localized variation in pollution and hence is a better choice for trend analysis and short-term observation of pollution. Nonetheless, we do recognize Sentinel-5P has greater spatial detail for the trend analysis within a region. Additionally, ground stations measured not only PM2.5 contaminants but a more varied set as well, thereby augmenting satellite-based assessments of air quality.

Pollutants	Metric	Mean Squared Error	Mean Absolute Error	R-squared Error	Root Mean squared Error
CO	**Ground Data**	2.526	0.481	0.979	1.495
	Sentinel Data	0.040	0.005	0.846	0.061
O3	**Ground Data**	249.534	1.999	0.839	13.103
	Sentinel Data	0.044	0.005	0.856	0.063

Journal of Information Systems Engineering and Management
2025, 10(54s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

NO2	**Ground Data**	628.780	2.215	0.886	17.038
	Sentinel Data	0.052	0.005	0.871	0.069
SO2	**Ground Data**	505.667	2.726	0.881	17.614
	Sentinel Data	0.050	0.006	0.837	0.067

Table 7: Evaluation Metrics produced by the proposed methodology for Sentinel and CPCB Data

This research concludes that the complementary capabilities of Sentinel's high-resolution geographical data and CPCB's high-frequency temporal data highlight the necessity of a hybrid modeling approach for air pollution assessment.[29][30]

4.6 Evaluating the Proposed Predictive Model's Relative Effectiveness by Comparing it with Current Research

This research evaluates our proposed predictive model by comparing its performance with established methodologies, highlighting its relative advantages and the value it adds to existing approaches, as presented in Table 8.

S No.	**Author and year**	**Dataset**	**ML/DL Algorithms**	**Results (Performance Parameters)**
1	Saba Ameer et al (2019)	5 cities of China	Gradient boosting, Decision tree regression	MAE=29.3%
2	Vo Thi Tam Minh et al (2021)	HCM City	Linear regression ,Neural network regression	R-squared = 0.67,0.66
3	Abdullah Kaviani Rad et al (2022)	3 Cities of Iran	XG Boost	R-squared=0.36
4	Doreswamy et al (2020)	Taiwan Air Quality Monitoring data sets	LASSO, RIDGE	R-squared=0.5652,0.1104
5	Tran Trung Tin et al (2021)	Tan son nhat station-HCM City	Decision forest regressor , XG Boost	R-squared = 0.67,0.63
6	A. Masih et al (2019)	Makkah city	Support Vector Machine ,	R-squared=0.67

Journal of Information Systems Engineering and Management
2025, 10(54s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

7	Harish Kumar K S et al (2020)	Taiwan	KNN Regressor	R-squared=0.6755
8	Zamani joharestani et al (2019)	Tehran city	DL-AODs	R-squared=0.63
9	Li et al (2020)	Beijing	Feed forward neural neural network	R-squared=0.74
8	**RegBoost (**Proposed Methodology)	**CPCB , NASA , Sentinel 5P**	**RR-XGBOOST**	**R-squared=0.847**

Table 8: Overview of ML and DL algorithms and their performance in various studies for air quality prediction

This research addressed these key questions (refer, section 1, Introduction), tackling the major issues: (1) the high MSI during the monsoon season caused by humidity, low wind, and inversions degrades air quality; (2) how efficient remote sensing will be utilized in monitoring pollutants will be judged based on comparisons of the top four pollutants determined from combined data and remote sensing data; (3) higher ASI in the monsoon season and lower pollutant dispersal, with CPI registering more pollution in the monsoon, post-monsoon, and autumn; and (4) the new hybrid imputation and modeling methods came in handy when enhancing PM2.5 forecast and analysis.

In summary, this research endeavored the importance of combining various data sources and employing cutting-edge analytical methods to better comprehend the air pollution dynamics. Our results reflect the intricate relationships between Gurugram's seasonal patterns, pollutant levels, and meteorological parameters, providing essential information for the design of effective air pollution abatement measures. This research found that Gurugram's seasonal patterns, pollutant levels, and weather conditions are all deeply connected. This knowledge is crucial for creating strong air pollution reduction plans. Looking ahead, this research will use the latest analysis methods to make our predictions more accurate and to better understand how pollutants spread.

CONCLUSION

This research endeavor was initiated with the objective of dissecting the intricate relationship between meteorological factors and air pollution dynamics within the complex urban environment of Gurugram. New hybrid imputation and prediction methods significantly improved PM2.5 data accuracy. The identification of key pollutant species and the study of derived indices revealed the complex link between PM2.5 concentration and meteorological factors. Investigations of seasonal variations in PM2.5 also demonstrated the substantial impact of weather, especially the compounding effects of stable atmospheric conditions. A detailed comparison of CPCB and Sentinel-5P data showed clear benefits for each modality. Sentinel-5P performed better in absolute error readings, indicating that it was able to detect more regional patterns. CPCB data, however, showed a stronger ability to account for differences in local pollution. New hybrid imputation and prediction methods significantly improved PM2.5 data accuracy. The identification of key pollutant species and the study of derived indices revealed the complex link between PM2.5 concentration and meteorological factors. The significant influence of weather, particularly the compounding effects of steady atmospheric conditions, was also shown by investigations of seasonal fluctuations in PM2.5. Sentinel-5P and CPCB data were thoroughly compared, revealing distinct advantages for each modality. Sentinel-5P demonstrated its ability to catch more regional patterns by doing better in absolute error measurements' data, however, showed a stronger ability to account for differences in local pollution. In addition, deep learning models like TimesNet, GNN, and GNN+XGBoost failed to perform well as a result of the small dataset, verifying the requirement of large data sizes for them to achieve the best performance. However, the methodology presented herein performed better in terms of processing the data available, resulting in higher predictive accuracy. The model will be parameterized with ground data and Sentinel-5P retrievals in subsequent research to add more to this general

assessment. More sophisticated methods will be utilized in order to boost the accuracy of the forecasts and present a broader picture of the movement and spreading of the contaminants. This, ultimately, will lead to improved methods for regulating the quality of the air.

REFERENCES

[1] Yu, This research nhua, et al. "Deep ensemble machine learning framework for the estimation of PM 2.5 concentrations." Environmental health perspectives 130.3 (2022): 037004.
[2] Chen, Gongbo, et al. "A machine learning method to estimate PM2. 5 concentrations across China with remote sensing, meteorological and land use information." Science of the Total Environment 636 (2018): 52-60.
[3] Zamani Joharestani, Mehdi, et al. "PM2. 5 prediction based on random forest, XGBoost, and deep learning using multisource remote sensing data." Atmosphere 10.7 (2019): 373.
[4] Ma, Jun, et al. "Identification of the most influential areas for air pollution control using XGBoost and Grid Importance Rank." Journal of Cleaner Production 274 (2020): 122835.
[5] Wang, J., Du, W., Cao, W., Zhang, K., Wang, W., Liang, Y., & This research, Q. (2024). Deep learning for multivariate time series imputation: A survey. arXiv preprint arXiv:2402.04059.
[6] Yeo, Inchoon, et al. "Efficient PM2. 5 forecasting using geographical correlation based on integrated deep learning algorithms." Neural Computing and Applications 33.22 (2021): 15073-15089.
[7] Tonion, F., and F. Pirotti. "SENTINEL-5P NO2 data: Cross-validation and comparison with ground measurements." The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences 43 (2022): 749-756.
[8] Bekkar, Abdellatif, et al. "Air-pollution prediction in smart city, deep learning approach." Journal of Big Data 8 (2021): 1-21.
[9] X. Liu, C. Yan, S. Liu, and S. R. Hanna, "Machine learning-based PM2.5 concentration forecasting: A review," Atmospheric Environment, vol. 254, p. 118395, 2021.
[10] S. Ghimire, Y. J. Kim, and M. L. Shrestha, "Machine learning-based prediction of PM2.5 concentration in Kathmandu Valley, Nepal," Atmospheric Pollution Research, vol. 11, no. 11, pp. 1845-1855, 2020.
[11] X. Li, L. Peng, X. Yao, Z. Qian, and W. Zhang, "Deep learning for traffic as time series: A survey," IEEE Transactions on Knowledge and Data Engineering, vol. 30, no. 9, pp. 1605-1623, 2017.
[12] M. Stafoggia, J. Schwartz, C. Badaloni, T. Bellander, E. Alessandrini, M. Jerrett, et al., "Long-term exposure to traffic-related air pollution and heart failure in 11 European cohorts: a meta-analysis of 100 000 subjects," European Heart Journal, vol. 35, no. 42, pp. 3428-3439, 2014.
[13] C. Brokamp, M. B. Rao, K. C. Dannemiller, and R. Jandarov, "Exposure assessment for particulate matter air pollution using random forest," Environmental Health Perspectives, vol. 127, no. 5, p. 057006, 2019.
[14] Bali, Kunal, Sagnik Dey, and Dilip Ganguly. "Diurnal patterns in ambient PM2. 5 exposure over India using MERRA-2 reanalysis data." Atmospheric Environment 248 (2021): 118180.
[15] Wang, Yuan, et al. "Full-coverage spatiotemporal mapping of ambient PM2. 5 and PM10 over China from Sentinel-5P and assimilated datasets: Considering the precursors and chemical compositions." Science of The Total Environment 793 (2021): 148535.
[16] Nouri, Amir, Mehdi Ghanbarzadeh Lak, and Morteza Valizadeh. "Prediction of PM2. 5 concentrations using principal component analysis and artificial neural network techniques: a case study: Urmia, Iran." Environmental Engineering Science 38.2 (2021): 89-98.
[17] Zhao, Chao, and Zongwei Cai. "Mass spectrometry-based omics and imaging technique: a novel tool for molecular toxicology and health impacts." Reviews of Environmental Contamination and Toxicology 261.1 (2023): 10.
[18] Zhao, Qin, et al. "Spatiotemporal PM2. 5 forecasting via dynamic geographical Graph Neural Network." Environmental Modelling & Software (2025): 106351.
[19] Tsai, I-Chun, et al. "Climate change-induced impacts on PM2. 5 in Taiwan under 2 and 4° C global warming." Atmospheric Pollution Research 15.6 (2024): 102106.
[20] Zheng, Mei, et al. "Seasonal trends in PM2. 5 source contributions in Beijing, China." Atmospheric Environment 39.22 (2005): 3967-3976.

[21] Vignesh, P. Preetham, Jonathan H. Jiang, and Pangaluru Kishore. "Predicting PM2. 5 concentrations across the USA using machine learning." Earth and Space Science 10.10 (2023): e2023EA002911.
[22] Belachsen, Idit, and David M. Broday. "Imputation of Missing PM2. 5 Observations in a Network of Air Quality Monitoring Stations by a New k NN Method." Atmosphere 13.11 (2022): 1934.
[23] Zeng, Qiaolin, et al. "Full-coverage estimation of PM2. 5 in the Beijing-Tianjin-Hebei region by using a two-stage model." Atmospheric Environment 309 (2023): 119956.
[24] Mantilla, Johan Andrés Oblitas, and Edwin Jhonatan Escobedo Cárdenas. "Prediction of PM2. 5 and PM10 Concentrations Using XGBoost and LightGBM Algorithms: A Case Study in Lima, Peru." Interfases 020 (2024): 183-206.
[25] Alarcon, V. "Spatial study of the health risk index by inhalation of PM 2.5 during the dry season in the metropolitan area of the Toluca Valley."
[26] Das, Rupesh M. "Seasonal and diurnal variability of PM2. 5 concentration along with the role of wind patterns over different locations of Delhi during the year 2018 to 2022." Environmental Monitoring and Assessment 197.4 (2025): 1-29.
[27] Shu, Zhuozhi, et al. "Impact of deep basin terrain on PM2. 5 distribution and its seasonality over the Sichuan Basin, Southwest China." Environmental Pollution 300 (2022): 118944.
[28] Singh, Gourav Kumar, et al. "Assessment of particulate matters especially PM2. 5 and PM10 concentration during and before lockdown in the various metropolitan cities of India." Journal of Environmental Management & Tourism 13.3 (2022): 665-673.
[29] Athira, T., and V. Agilan. "Analysing Long-Term Trends in Monthly PM2. 5 Concentrations Over India Using a Satellite-Derived Dataset." Aerosol Science and Engineering (2024): 1-15.
Sahu, Shovan Kumar, and Sri Harsha Kota. "Significance of PM2. 5 air quality at the Indian capital." Aerosol and air quality research 17.2 (2017): 588-597.
[30] https://power.larc.nasa.gov/
[31] https://cpcb.nic.in/real-time-air-qulity-data/
[32] https://dataspace.copernicus.eu/explore-data/data-collections/sentinel-data/sentinel-5p