Journals Pub

International Journal of
Distributed Computing and Technology

ISSN: 2455-7307
Volume 11, Issue 1, 2025
DOI (Journal): 10.37628/IJDCT

https://journalspub.com/journal/ijdct/

Research IJDCT

Comparative Analysis of Air Quality Indices Across Five Eastern States of India Using ML Models and Pollutant Based Insights

Vansh Mathur[1], Shalavya Agrawal[2,*], Naina Agrawal[3], Udit Pandey[4], Parth Kedawat[5]

Abstract

This paper presents a comparative study of Air Quality Index (AQI) levels in five eastern Indian states–Arunachal Pradesh, Assam, Meghalaya, Nagaland, and Tripura. Using historical data, the study explores the concentration of key pollutants, including PM2.5, PM10, NO_2, SO_2, CO, and Ozone, highlighting regional variations in air quality. Assam, with its higher levels of urbanization and industrial activity, shows elevated AQI levels, especially for pollutants like PM10 and NO_2. In contrast, Arunachal Pradesh and Nagaland, with their rural and forested landscapes, exhibit lower AQI levels, indicating better air quality. The study utilizes AdaBoost (Adaptive Boosting) and XGBoost (Extreme Gradient Boosting) models to predict future AQI trends, providing valuable insights for policymakers to anticipate air quality changes. Monthly and yearly comparisons of pollutant levels reveal significant temporal variations, with some states experiencing seasonal spikes in pollutants, particularly in winter. Through detailed graphs and AdaBoost/XGBoost-based predictions, the study emphasizes the impact of urbanization, industrialization, and geographical factors on air quality, offering crucial data for environmental policy and public health initiatives.

Keywords: Air Quality Index (AQI), AdaBoost, CO, eastern states of India, graphical abstract, machine learning, NO_2, Ozone, PM2.5, PM10, pollution, SO_2, XGBoost

INTRODUCTION

Air pollution has emerged as a critical global issue, affecting millions of people across different regions and posing severe risks to both environmental sustainability and public health [1]. The rapid pace of urbanization and industrialization, particularly in developing countries like India, has exacerbated this issue, making it one of the leading contributors to premature deaths and diseases worldwide [2, 3]. According to the World Health Organization (WHO), air pollution is responsible for approximately 7 million deaths annually [4]. In many Indian cities, particularly in the northern plains, residents are frequently exposed to hazardous air quality, which can trigger respiratory problems, cardiovascular diseases, and even reduce life expectancy [5]. Therefore, understanding air pollution levels and their long-term effects has become a priority for governments, environmental agencies, and researchers [4]. This understanding is crucial for devising effective policies aimed at mitigating the adverse effects of pollution on public health and the environment.

The Air Quality Index (AQI) serves as a standardized metric to simplify the assessment of air quality by converting pollutant concentrations into

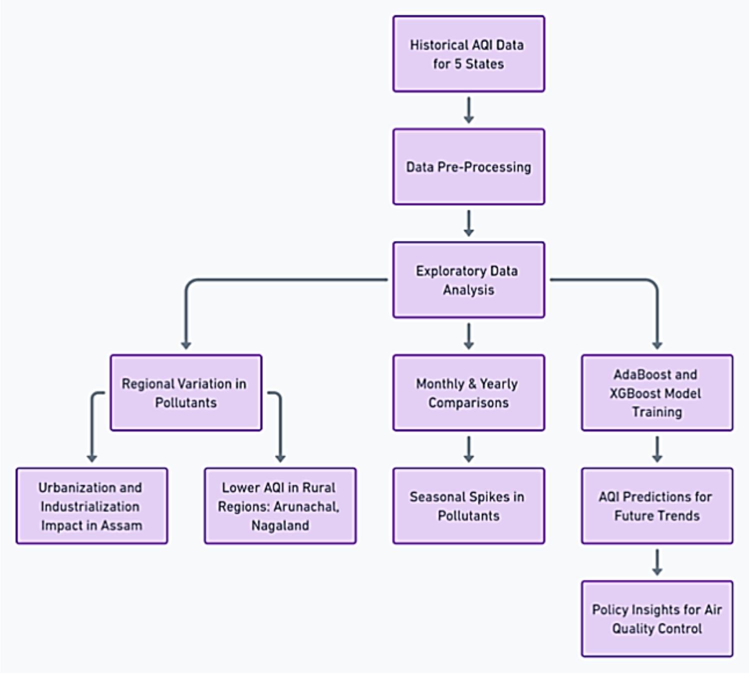
***Author for Correspondence**
Shalavya Agrawal
E-mail: shalavyaagrawal@gmail.com

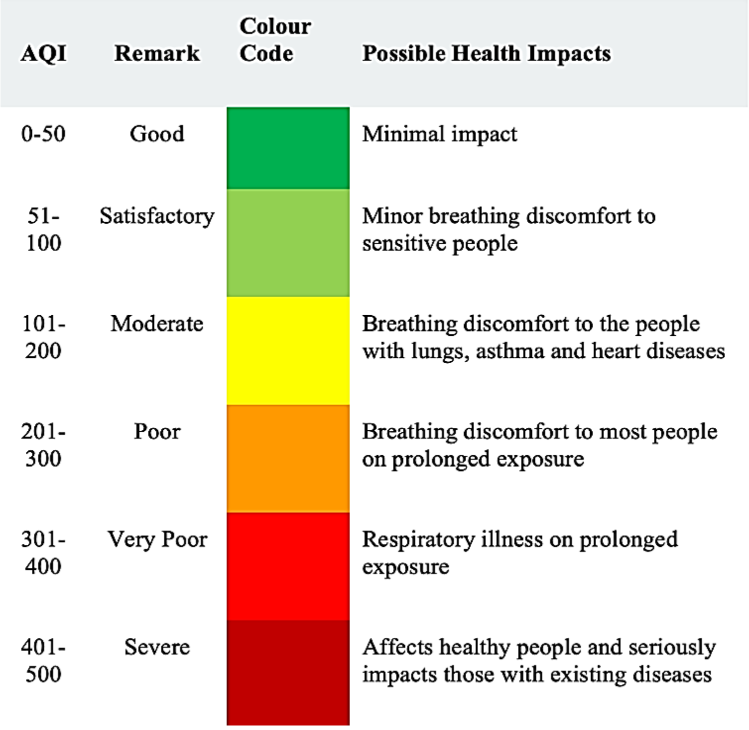
Student, School of Computer Science and Engineering, Vellore Institute of Technology, Vellore, Tamil Nadu, India

Received Date: August 29, 2024
Accepted Date: December 06, 2024
Published Date: December 17, 2024

Citation: Vansh Mathur, Shalavya Agrawal, Naina Agrawal, Udit Pandey, Parth Kedawat. Comparative Analysis of Air Quality Indices Across Five Eastern States of India Using ML Models and Pollutant Based Insights. International Journal of Distributed Computing and Technology. 2025; 11(1): 31–42p.

a single, easily interpretable number. This metric encompasses several key pollutants, such as particulate matter (PM2.5 and PM10), nitrogen dioxide (NO_2), sulfur dioxide (SO_2), carbon monoxide (CO), and ozone (O_3). These pollutants, each presenting unique risks to human health, are closely monitored to ensure public safety. For example, PM2.5 and PM10 are known to penetrate deep into the respiratory system, exacerbating conditions like asthma and other lung diseases. Meanwhile, NO_2 and SO_2 contribute to respiratory irritation and cardiovascular stress, especially in vulnerable populations. The AQI system categorizes air quality from “Good” to “Hazardous,” thereby providing essential information to the public and policymakers to help mitigate risks associated with high pollution levels (Figure 1) [6].


Historical AQI Data for 5 States
Data Pre-Processing
Exploratory Data Analysis
Regional Variation in Pollutants
Monthly & Yearly Comparisons
AdaBoost and XGBoost Model Training
Urbanization and Industrialization Impact in Assam
Lower AQI in Rural Regions: Arunachal, Nagaland
Seasonal Spikes in Pollutants
AQI Predictions for Future Trends
Policy Insights for Air Quality Control



AQI	Remark	Colour Code	Possible Health Impacts
0-50	Good	🟩	Minimal impact
51-100	Satisfactory	🟩	Minor breathing discomfort to sensitive people
101-200	Moderate	🟨	Breathing discomfort to the people with lungs, asthma and heart diseases
201-300	Poor	🟧	Breathing discomfort to most people on prolonged exposure
301-400	Very Poor	🟥	Respiratory illness on prolonged exposure
401-500	Severe	🟥	Affects healthy people and seriously impacts those with existing diseases

Figure 1. AQI Metric.

India, as one of the worlds' fastest-growing economies, faces significant challenges related to air pollution, especially in its urban and industrial areas. The rapid urbanization and escalating energy demands have led to increased emissions from transportation, industries, and residential sources. Cities like Delhi, Kolkata, and Mumbai frequently report some of the highest pollution levels globally, with AQI values regularly breaching safe limits. However, the air quality varies greatly across India's diverse regions due to differences in geography, industrial activity, and population density. The northern regions, especially those in the Indo-Gangetic Plain, suffer from severe pollution due to vehicle emissions, industrial activities, and seasonal agricultural practices like stubble burning. In contrast, northeastern states, which are less industrialized, are believed to have relatively better air quality. However, detailed studies on these regions remain limited [6, 7]. This study aims to fill this gap by analyzing and comparing AQI levels across five eastern Indian states: Arunachal Pradesh, Assam, Meghalaya, Nagaland, and Tripura [8].

STUDY AREA

The study focuses on AQI across five northeastern states of India – Arunachal Pradesh, Assam, Meghalaya, Nagaland, and Tripura – each characterized by distinct geographical and environmental features. Arunachal Pradesh, with its expansive forest cover and minimal industrial presence, is generally considered to have cleaner air. However, the state's increasing developmental projects, including infrastructure expansion, could pose future air quality challenges. Assam, being more urbanized and industrialized, has significant sources of pollution, such as vehicular emissions and industrial activities in cities like Guwahati. This leads to higher concentrations of pollutants like PM10 and NO_2 [9]. Meghalaya, known for its hilly terrain, faces localized air quality issues, especially due to coal mining activities, which contribute to elevated pollutant levels in certain areas [10].

Nagaland and Tripura, primarily rural states with dense forest cover, generally experience lower pollution levels. The absence of heavy industry in these regions contributes to better air quality; however, seasonal agricultural practices, such as crop burning, and local vehicular emissions can affect the AQI temporarily [11]. Understanding the AQI trends in these varied environments is crucial for identifying region-specific pollution sources and implementing effective air quality management strategies.

DATA METHODOLOGY

Data Extraction

Data on pollutant concentrations, including PM2.5, PM10, NO_2, SO_2, CO, and O_3, were extracted from the Central Pollution Control Board (CPCB), India's main body for air quality monitoring. The CPCB operates a vast network of monitoring stations that record pollutant levels across various regions [5]. The data for this study was gathered from stations situated within or near the target states, focusing on daily readings of pollutants over a defined period [11]. This comprehensive dataset provides insights into both short-term variations and long-term trends in air quality, making it valuable for predictive modeling. The data collection process aimed to ensure the inclusion of diverse climatic conditions and geographic factors to provide a holistic view of air quality in these states.

Data Pre-Processing

Data preprocessing is a crucial step to ensure the quality and consistency of datasets used in analysis and modeling. The air quality data initially contained missing entries and outliers that could distort model accuracy [12]. To address this, missing values were imputed using statistical techniques like mean substitution or interpolation, while instances where imputation was not feasible were excluded. Outliers, identified through Z-score analysis and visualized using boxplots, were managed to maintain the integrity of the data [13]. These steps were vital in preparing a clean dataset suitable for training predictive models.

Normalization of numerical features, such as pollutant concentrations, was performed to ensure that variables with different scales do not disproportionately influence the model's learning process [8]. This step was especially important for pollutants like PM2.5, which can have a wide range of values compared to gases like CO or NO_2. Additionally, categorical variables were encoded to ensure compatibility with machine learning models, resulting in a refined dataset ready for analysis and model training [14].

Feature engineering played a critical role in enhancing the dataset for improved model performance. This involved creating new features, such as daily pollutant averages, weekly trends, and pollutant ratios, to better capture temporal variations and pollutant dynamics [15]. Moreover, interactions between pollutants and meteorological parameters, such as the impact of temperature on PM2.5 dispersion or the effect of humidity on NO_2 concentration, were encoded as additional features. This approach aimed to provide the models with a richer understanding of underlying patterns in air quality changes. The dataset also included lagged variables to account for delayed effects of meteorological conditions on pollutant levels, which is particularly useful in time-series predictions [3]. By implementing these steps, the study ensured that the models could better identify complex relationships in the data, ultimately leading to more accurate AQI forecasts.

Computing AQI

To compute the AQI, the study followed CPCB guidelines, considering pollutants like PM2.5, PM10, NO_2, SO_2, CO, and O_3 [3]. The AQI was calculated using the highest concentration values among any three pollutants, providing a comprehensive assessment of air quality for each location. The inclusion of meteorological parameters like Temperature, Relative Humidity (RH), Wind Speed (WS), Wind Direction (WD), Solar Radiation (SR), Air Pressure (BP), Ambient Temperature (AT), and Rainfall (RF) allowed for a nuanced understanding of how weather conditions impact air quality [16, 17]. These factors play a significant role in the dispersion or concentration of pollutants, with phenomena like temperature inversions during winter months often leading to higher pollution levels [16].

Machine Learning Methods to Predict AQI

With the rise in availability of real-time air quality data, machine learning models have become a valuable tool for predicting AQI trends and understanding pollution dynamics. This study employs two robust models, AdaBoost and XGBoost, chosen for their ability to handle complex datasets with varying relationships between pollutants and environmental factors [17]. While traditional statistical models offer insights, machine learning models like these can better capture the non-linear interactions between pollutants and meteorological influences [18].

Adaboost Model

AdaBoost, short for Adaptive Boosting, is an ensemble learning technique that creates a strong predictive model by iteratively improving the accuracy of weak learners – typically simple decision trees [13]. By focusing on the misclassified instances from previous iterations, AdaBoost adjusts its predictions, thereby enhancing overall accuracy. This model is particularly effective in cases where the dataset contains noise or irregular patterns, making it a suitable choice for air quality prediction.

The final prediction of the AdaBoost model is a weighted sum of the predictions of weak classifiers ht(x). The formula is:

$$H(x) = \operatorname{sign}\left(\sum_{t=1}^{T} \alpha_t \cdot h_t(x)\right) \tag{1}$$

where:

H(x): Final prediction (either –1 or 1 for binary classification).

T: Total number of iterations (weak classifiers).

α_t: Weight of the t-th weak classifier, calculated as:

$$\alpha_t = \frac{1}{2} \ln\left(\frac{1-\epsilon_t}{\epsilon_t}\right) \tag{2}$$

$h_t(x)$: The t-th weak classifier's prediction.
ϵt: Error of the t-th weak classifier.

XGBoost Model

XGBoost, an advanced gradient boosting framework, is known for its efficiency and ability to handle large datasets with complex interactions. By using decision trees and gradient descent, XGBoost optimizes predictions through minimizing residual errors and incorporating regularization techniques to prevent overfitting [8]. Its ability to handle missing values and incorporate parallel processing makes XGBoost a powerful model for AQI forecasting.

The model's robustness enables it to deliver high accuracy across different states with varying pollution sources and meteorological conditions (Figure 2).

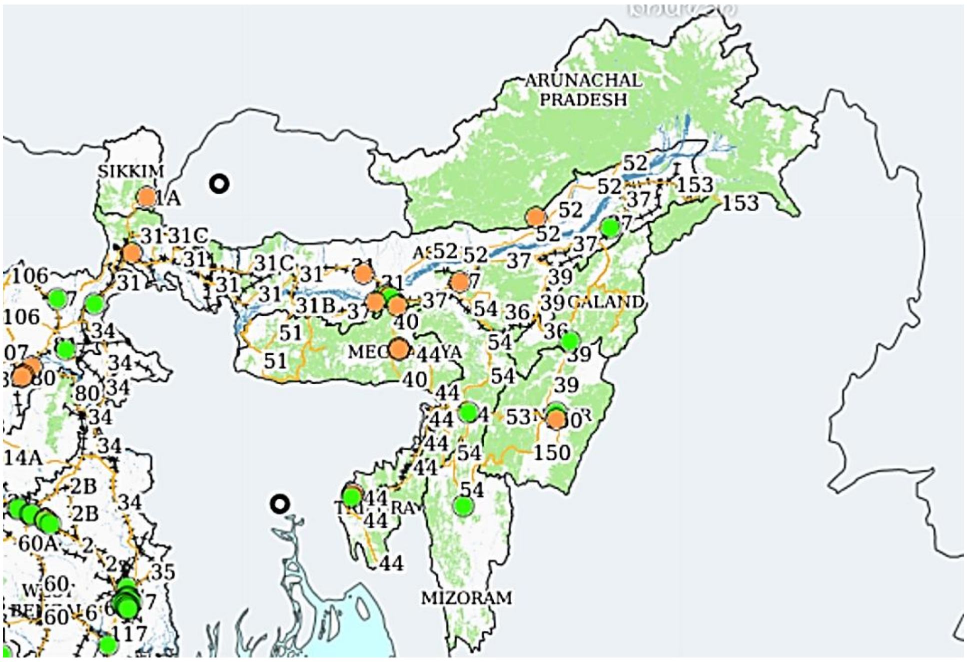


Figure 2. Location map of Northeastern States of India.

By training AdaBoost and XGBoost on historical AQI data, the study forecasts future air quality trends, offering insights into potential periods of heightened pollution. These forecasts can guide policymakers in issuing warnings or implementing control measures during high-risk periods, thereby supporting public health and environmental management [19].

The XGBoost model makes predictions by adding the outputs of all the decision trees (boosted trees). The formula for the final prediction is:

$$\hat{y} = \sum_{k=1}^{T} f_k(x) \tag{3}$$

where:
$\hat{Y}$: Final predicted value for input x.
T: Total number of trees.
$f_k(x)$: The prediction of the k-th tree, optimized using gradient boosting.

The objective function minimized during training is:

$$\mathcal{L}(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{T} \Omega(f_k) \tag{4}$$

where:

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \sum_{j} w_j^2 \tag{5}$$

Exploratory Data Analysis

After EDA provides an initial understanding of the air quality data across the five states by exploring trends, distributions, and relationships. Graphical methods like histograms and boxplots revealed variations in pollutant levels across different states and time periods. For instance, Assam showed significantly higher levels of PM10 and NO_2, indicative of its urban and industrial activities, whereas Arunachal Pradesh and Nagaland displayed lower concentrations, reflecting their rural settings [11]. Correlation analysis between pollutants and meteorological factors, visualized through heatmaps, highlighted how temperature, humidity, and wind conditions affect pollutant dispersion. For example, wind speed was found to have a negative correlation with PM2.5 levels, suggesting that stronger winds help disperse particulate matter [9, 10]. Additionally, seasonal trends were examined, with winter months showing higher pollutant levels due to temperature inversions that trap pollutants close to the surface [3]. These insights helped in selecting the most relevant features for training the AdaBoost and XGBoost models, ensuring that the models effectively capture the nuances of air quality variations in these regions [13, 17].

Building on the EDA findings, the analysis also identified distinct patterns in the pollutant distribution that vary not only by geography but also by seasonal changes. For example, Assam’s urban centres like Guwahati are heavily influenced by vehicular emissions and industrial activities, which contribute to sustained high levels of NO_2 and PM10 throughout the year [11]. In contrast, states like Arunachal Pradesh and Nagaland, with their dense forest cover and lower population density, experience more stable and lower pollutant levels, except during seasonal agricultural practices such as crop burning [9]. The study further revealed that in Meghalaya, mining activities contributed to localized spikes in pollutants like PM10, emphasizing the impact of specific economic activities on air quality [10]. These patterns underscore the need for tailored air quality management strategies for each state, considering both the local economic drivers of pollution and the seasonal atmospheric conditions that can exacerbate air quality issues [14]. By incorporating these insights into the feature selection process for AdaBoost and XGBoost models, the study ensured that the models are well-equipped to predict AQI trends with a high degree of accuracy, ultimately aiding policymakers in designing more effective intervention measures [8, 17].

RESULTS

AQI Comparisons Across the States

The comparative analysis of AQI levels in the five eastern states is shown in Table 1. The graphs represent the average AQI across these states based on historical data and predictions. The graphs for individual pollutants (Ozone, CO, SO_2, NO_2, PM10, and PM2.5) offer a detailed comparison of how each pollutant contributes to the overall air quality in these states.

Data Transformation

Table 1 illustrates the impact of data transformation on the skewness and kurtosis of various air quality attributes, including CO, NO_2, Ozone, PM10, PM2.5, and SO_2. Before transformation, most attributes exhibit positive skewness, indicating a longer right tail in their distribution, and high kurtosis values, which suggest the presence of outliers or heavy-tailed distributions.

Transformations are often applied to normalize such distributions, making the data more suitable for predictive modeling. However, in this case, the skewness and kurtosis values remain unchanged after transformation, indicating that the applied transformations did not significantly alter the distribution shapes of these variables. This suggests that the transformations were either ineffective or not aimed at normalizing these specific metrics but potentially focused on other aspects like scaling or range adjustments to improve model training stability.

Additionally, the unchanged skewness and kurtosis values after transformation might imply that the original data distributions were already stable enough for certain modeling techniques, reducing the necessity for aggressive normalization. In such cases, transformations like standard scaling or min-max scaling may have been prioritized to ensure consistency across features rather than altering their distribution shapes. This approach can be particularly useful when the goal is to maintain the inherent variability of the data while ensuring that all features contribute evenly during model training.

Table 1. Skewness and Kurtosis Values of Selected Features Before and After the Transformation.

Attributes	**Before Transformation**		**After Transformation**	
	Skewness	***Kurtosis***	***Skewness***	***Kurtosis***
CO	1.37	6.9	1.37	6.9
NO_2	1.96	9.34	1.96	9.34
Ozone	2.5	11.18	2.5	11.18
PM10	1.48	5.5	1.48	5.5
PM2.5	1.53	6.12	1.53	6.12
SO_2	1.5	6.85	1.5	6.85

Particulate Matters (PM2.5 and PM10)

The comparison of particulate matter levels (PM2.5 and PM10) across Arunachal Pradesh, Assam, Meghalaya, Nagaland, and Tripura reveals notable trends over both monthly and yearly scales. Tripura consistently exhibits the highest levels of both PM10 and PM2.5, especially during the months from June to December [8, 14], indicating significantly poorer air quality compared to the other states. This trend is mirrored in the yearly data, where Tripura shows a substantial increase in PM10 levels from 2022 to 2023, accompanied by a rise in PM2.5 concentrations, suggesting deteriorating air quality over time. In contrast, Arunachal Pradesh and Meghalaya show relatively lower and more stable levels of particulate matter throughout the months, with both states showing a decrease in PM10 levels year-over-year, signalling an improvement in air quality [3, 6]. Assam, though showing steady levels during the months, demonstrates a notable decline in both PM10 and PM2.5 over the year, indicating positive progress in reducing particulate pollution. Nagaland, while relatively stable, exhibits moderate levels of both PM10 and PM2.5, though it is slightly higher than the cleaner states like Arunachal Pradesh and Meghalaya. Overall, Tripura remains the region of highest concern due to its elevated and increasing particulate pollution levels, while Assam, Arunachal Pradesh, and Meghalaya show improvements or relatively stable, better air quality. The data suggests that while some regions are managing to control particulate matter effectively, Tripura faces a growing challenge, particularly in the latter half of the year pollution status (Figures 3–6).

Gaseous Pollutants (CO, SO_2, NO_2)

Based on the monthly and yearly visualizations for the gaseous pollutants – CO, SO_2, and NO_2 – across Arunachal Pradesh, Assam, Meghalaya, Nagaland, and Tripura, the comparison reveals interesting trends in air quality.

For CO (Carbon Monoxide), the levels in all states remain relatively low throughout the year. However, Tripura sees a notable spike in CO concentration during the winter months, especially between October and December, while Assam, Nagaland, and Arunachal Pradesh maintain relatively stable and lower levels. Year-over-year, Tripura continues to show slightly elevated CO levels in 2023, with Nagaland also showing a minor increase, while the other states remain steady [11–20].

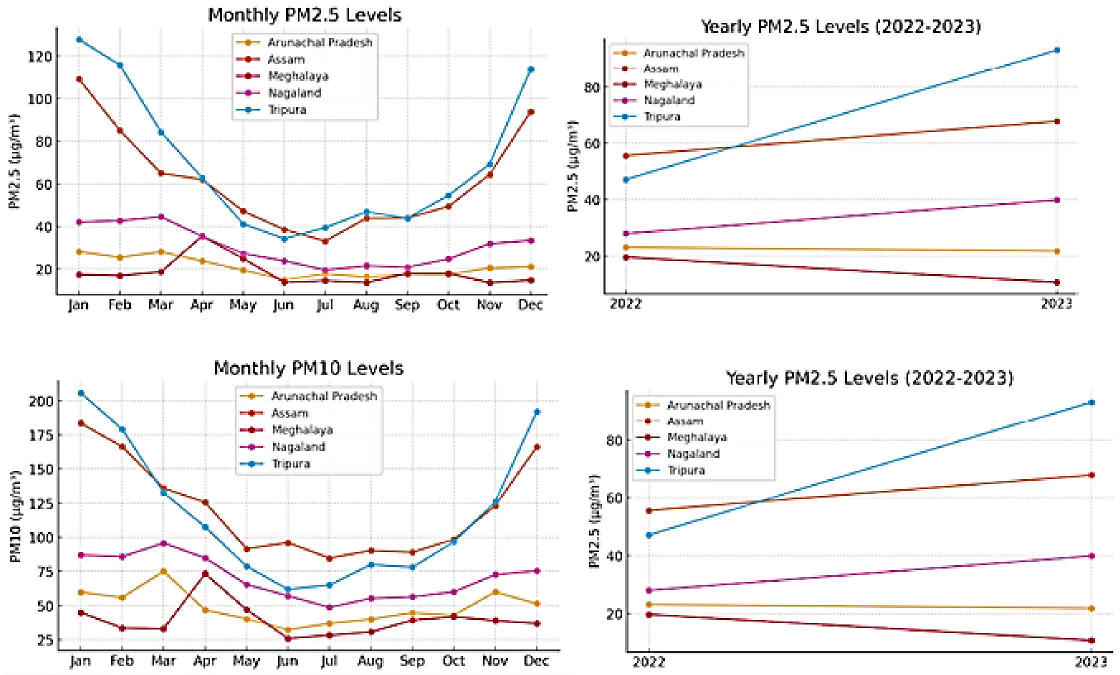


Figure 3. Annual and monthly variation of particulate matter (PM2.5 and PM2.10).

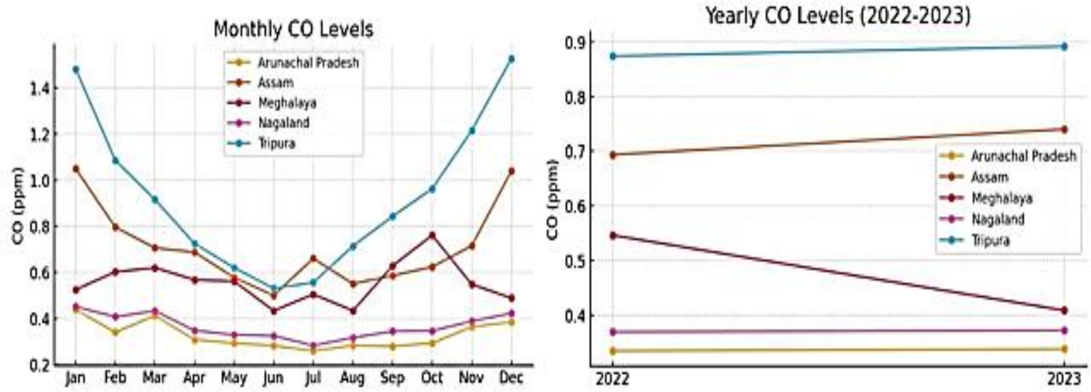


Figure 4. Annual and monthly variation of gaseous pollutant CO.

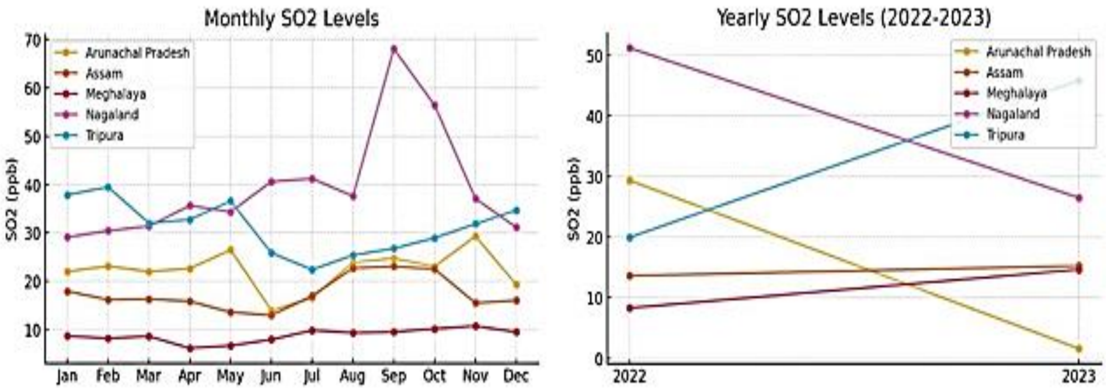


Figure 5. Annual and monthly variation of gaseous pollutant $SO_{2.}$

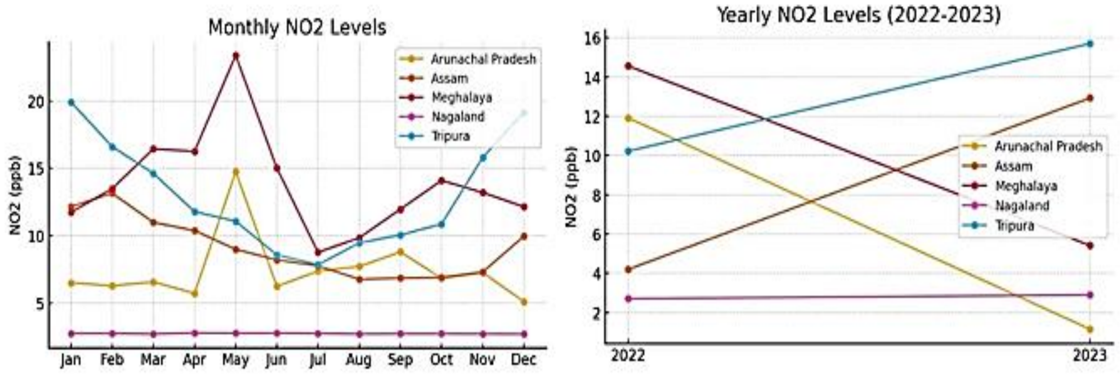


Figure 6. Annual and monthly variation of gaseous pollutant NO_2.

For SO_2 (Sulphur Dioxide), Nagaland exhibits the highest concentrations, particularly peaking around August and September. The other states maintain significantly lower SO_2 levels throughout the year, with little fluctuation. Assam, Arunachal Pradesh, and Meghalaya show relatively stable SO_2 levels both on a monthly and yearly basis. From 2022 to 2023, Tripura shows a consistent rise, while Nagaland, despite starting high, sees a slight decline in 2023. Assam and Arunachal Pradesh show stable SO_2 concentrations, reflecting little change in air quality regarding this pollutant.

As for NO_2 (Nitrogen Dioxide), Nagaland experiences significant monthly spikes, especially in April and September, suggesting seasonal variations or specific pollution events. Assam and Arunachal Pradesh display relatively stable, lower levels, with no significant monthly variations, while Tripura shows an increase during the end of the year. Yearly trends indicate that NO_2 levels in Nagaland and Tripura remain high, with Nagaland seeing an increase into 2023, while Arunachal Pradesh and Assam have maintained or slightly decreased their levels. Overall, Nagaland and Tripura stand out with elevated levels of all gaseous pollutants, indicating localized pollution events or industrial activities, while the other states, particularly Assam and Arunachal Pradesh, show better air quality with lower and stable concentrations of CO, SO_2, and NO_2.

Ozone

The comparison of Ozone (O_3) levels across Arunachal Pradesh, Assam, Meghalaya, Nagaland, and Tripura shows significant variations both on a monthly and yearly basis. Nagaland consistently has the highest ozone levels, with prominent peaks in April and November, likely due to seasonal factors or localized pollution. In contrast, Assam and Arunachal Pradesh display lower, more stable ozone levels throughout the year, while Meghalaya maintains consistently low levels, indicating better air quality.

Tripura shows moderate levels but experiences a notable increase in ozone concentration during the colder months. On a yearly scale, Nagaland continues to report the highest ozone levels, with Tripura showing an upward trend into 2023. Arunachal Pradesh and Assam have seen improvements with a decrease in ozone levels, and Meghalaya remains stable with the lowest levels overall. Thus, while Nagaland and Tripura face rising ozone pollution, Assam and Arunachal Pradesh are experiencing improvements, with Meghalaya maintaining its low pollution status (Figure 7).

Machine Learning Models to Predict AQI

The prediction of Air Quality Index (AQI) using machine learning methods has become increasingly effective with the growing availability of environmental data. In this study, two prominent machine learning models, XGBoost and AdaBoost, were utilized to forecast AQI across several northeastern Indian states, including Arunachal Pradesh, Assam, Meghalaya, Nagaland, and Tripura. The performance of these models was evaluated using key metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Square Error (RMSE), and the R^2 score, to determine their accuracy in predicting future air quality levels.

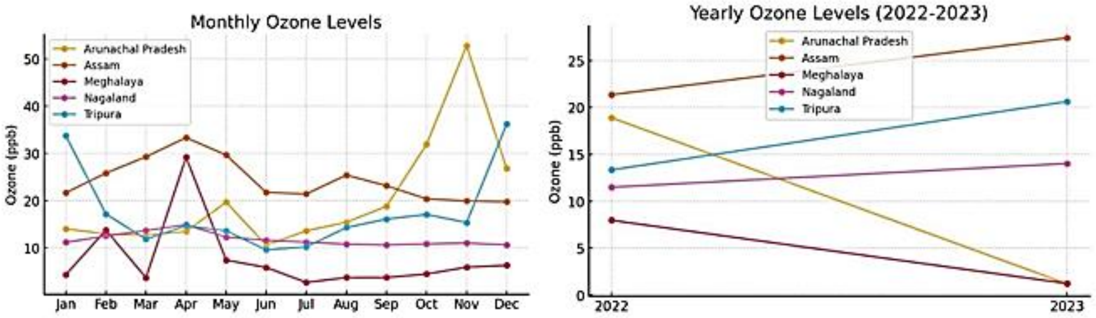


Figure 7. Annual and monthly variation of ozone.

XGBoost, known for its gradient boosting framework, performed exceptionally well across most states, particularly in Nagaland, where it achieved an R² score of 0.985, demonstrating its ability to capture complex patterns in the data. AdaBoost, a boosting algorithm that combines weak learners to improve model performance, also provided strong results, with its highest performance observed in Tripura, where it achieved an R² score of 0.990.

By training these models on historical AQI data, both XGBoost and AdaBoost were able to identify significant temporal patterns and predict future AQI trends with high accuracy. This allows policymakers and environmental agencies to anticipate air quality fluctuations and implement timely interventions to mitigate pollution, thus supporting better environmental and public health strategies.

IMPLICATIONS AND PERSPECTIVES

The implications of this study's comparative analysis between XGBoost and AdaBoost highlight significant insights (Table 2) for both environmental policy and public health. In this analysis, XGBoost consistently outperformed AdaBoost across most states, particularly in Nagaland and Assam, where it demonstrated superior predictive accuracy with R² scores of 0.985 and 0.965, respectively. XGBoost's strength lies in its ability to handle larger datasets with complex relationships, such as multiple pollutants and meteorological factors, making it highly effective for AQI prediction. Conversely, AdaBoost delivered competitive results in certain cases, notably in Tripura, where it achieved a slightly higher R² score of 0.990. However, on average, XGBoost proved to be the more robust model overall [12–13].

Table 2. Machine learning Models with their performance factors in prediction of AQI.

State	Model	MAE	MSE	RMSE	R² Score
ARUNACHAL PRADESH	XGBoost	3.51	226.02	15.03	0.896
	AdaBoost	4.37	232.33	15.24	0.893
ASSAM	XGBoost	5.24	475.58	21.81	0.965
	AdaBoost	8.76	246.38	15.70	0.982
MEGHALAYA	XGBoost	5.00	86.73	9.31	0.583
	AdaBoost	5.42	65.01	8.06	0.687
NAGALAND	XGBoost	1.90	19.23	4.39	0.985
	AdaBoost	3.56	24.14	4.91	0.981
TRIPURA	XGBoost	5.77	255.84	15.99	0.979
	AdaBoost	4.78	129.52	11.38	0.990

Note: MAE – Mean Absolute error; MSE- Mean squared error: RMSE- Root Mean square Error R^2 - Correlation Coefficient.

The predictive power of these models allows policymakers to anticipate AQI trends and implement timely interventions, such as issuing public health advisories or adjusting industrial operations during periods of poor air quality. The ability of both models to capture regional variations in air quality, particularly in urbanized regions like Assam or rural areas like Arunachal Pradesh, provides tailored insights for targeted policy actions. The comparative analysis emphasizes the importance of selecting the most suitable machine learning model for the task, with XGBoost standing out as the top performer. These insights empower decision-makers to address air quality challenges more effectively, contributing to improved public health and more sustainable urban planning strategies (Figure 8).

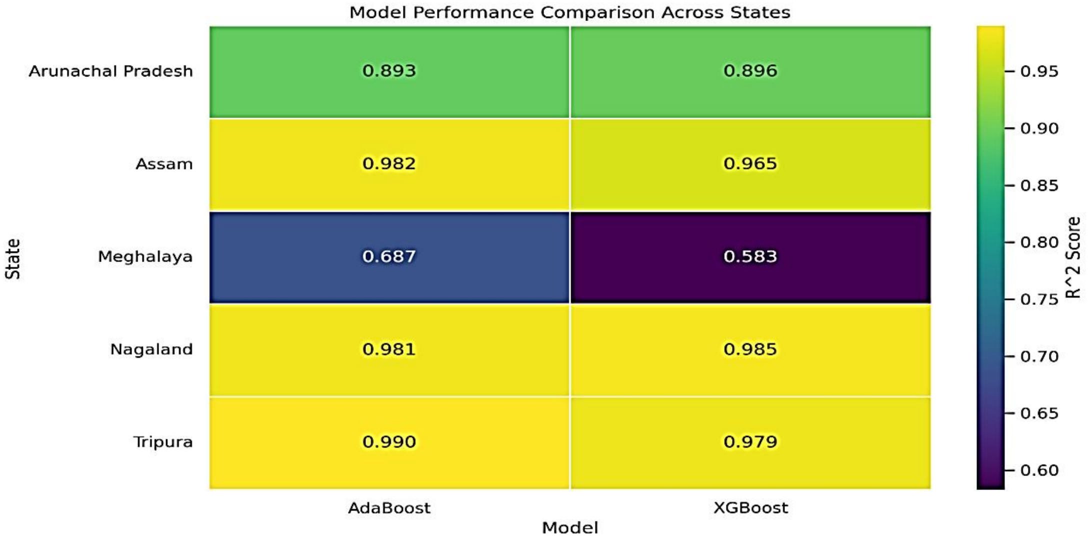


Figure 8. R^2 Score Model Heat Map.

CONCLUSIONS

Based on the yearly analysis and the performance of XGBoost and AdaBoost models across various states, it is evident that both models offer substantial accuracy in forecasting AQI levels over time. The yearly AQI trends, especially between 2017 and 2022, revealed rising AQI levels due to urbanization and industrialization in states like Assam and Tripura, while rural and less industrialized states like Arunachal Pradesh and Nagaland exhibited better air quality. The impact of external factors, such as the 2020 nationwide lockdown due to the COVID-19 pandemic, was also evident, with AQI levels dipping temporarily during that period before resuming their upward trajectory.

XGBoost consistently outperformed AdaBoost in most regions, particularly in Nagaland, where it achieved an R^2 score of 0.985, and Assam, with an R^2 score of 0.965, demonstrating its ability to handle complex datasets over multiple years. However, AdaBoost provided competitive results, particularly in Tripura, where it slightly outperformed XGBoost with an R^2 score of 0.990.

In conclusion, the combination of yearly and model-based analysis illustrates the growing need for robust machine learning frameworks like XGBoost and AdaBoost to predict AQI trends accurately. These insights allow policymakers to anticipate pollution spikes and take preventive measures. To further improve these models, validation across a broader range of regions and years is essential, ensuring their scalability and adaptability to various air quality conditions.

REFERENCES

1. Alrashed S. Key performance indicators for Smart Campus and Microgrid. Sustain Cities Soc. 2020 Sep 1;60:102264.
2. Iskandaryan D, Ramos F, Trilles S. Air quality prediction in smart cities using machine learning technologies based on sensor data: A review. Appl Sci. 2020 Apr 1;10(7):2401.
3. Zhang Y, Xue W, Long R, Yang H, Wei W. Acetochlor affects zebrafish ovarian development by producing estrogen effects and inducing oxidative stress. Environ Sci Pollut Res. 2020 Aug;27:27688–96.

4. Chowdhury S, Dey S. Cause-specific premature death from ambient PM2. 5 exposure in India: Estimate adjusted for baseline mortality. Environ Int. 2016 May1;91:283–90.
5. Guttikunda SK, Jawahar P. Atmospheric emissions and pollution from the coal-fired thermal power plants in India. Atmos Environ. 2014 Aug 1;92:449–60.
6. Mahato S, Pal S, Ghosh KG. Effect of lockdown amid COVID-19 pandemic on air quality of the megacity Delhi, India. Sci Total Environ. 2020 Aug 15;730:139086.
7. Ramsey NR, Klein PM, Moore III B. The impact of meteorological parameters on urban air quality. Atmos Environ. 2014 Apr 1;86:58–67.
8. Dawar I, Singal M, Singh V, Lamba S, Jain S. Air Quality Prediction Using Machine Learning Models: A Predictive Study in the Himalayan City of Rishikesh. SN Comput Sci. 2024 Dec;5(8):1–6.
9. Sothea K, Oanh NT. Characterization of emissions from diesel backup generators in Cambodia. Atmos Pollut Res. 2019 Mar 1;10(2):345–54.
10. Baklanov A, Zhang Y. Advances in air quality modeling and forecasting. Global Transitions. 2020 Jan 1;2:261–70.
11. Hardini M, Sunarjo RA, Asfi M, Chakim MH, Sanjaya YP. Predicting air quality index using ensemble machine learning. ADI J Recent Innov. 2023Aug 22;5(1Sp):78–86.
12. Ren L, Song C, Wu W, Guo M, Zhou X. Reservoir effects on the variations of the water temperature in the upper Yellow River, China, using principal component analysis. J Environ Manage. 2020 May 15;262:110339.
13. Wu X, Kumar V, Ross Quinlan J, Ghosh J, Yang Q, Motoda H, et al. Top 10 algorithms in data mining. Knowl Inf Syst. 2008Jan;14:1–37.
14. Zheng Y, Liu F, Hsieh HP. U-air: When urban air quality inference meets big data. In Proceedings of the 19th ACM SIGKDD Int Conf Knowl Discov Data Min. 2013 Aug 11:1436–1444.
15. Li C, Li Y, Bao Y. Research on air quality prediction based on machine learning. In 2021 2nd Int Conf Intell Comput Hum-Comput Interact. IEEE. 2021 Nov 17:77–81.
16. Breiman L. Random forests. Machine Learning. 2001;45:5–32.
17. Chen T, Guestrin C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm Sigkdd Int Conf Knowl Discov Data Min. 2016 Aug 13:785–794.
18. Decroos T, Van Haaren J, Davis J. Automatic discovery of tactics in spatio-temporal soccer match data. In Proceedings of the 24th acm Sigkdd Int Conf Knowl Discov & Data Min. 2018 Jul 19:223–232.
19. Freund Y, Schapire RE. A decision-theoretic generalization of on-line learning and an application to boosting. J Comput Syst Sci. 1997Aug1;55(1):119–39.
20. Solomos S, Bougiatioti A, Soupiona O, Papayannis A, Mylonaki M, Papanikolaou C, et al. Effects of regional and local atmospheric dynamics on the aerosol and CCN load over Athens. Atmos Environ. 2019 Jan 15;197:53–65.