Vāyu Shuddhi: A Data-Driven Analysis of AQI Pollutants

Aditya Shriwal
*School of Computer Science
and Engineering
Vellore Institute of Technology*
Vellore, India
aditya.shriwal2022@vitstudent.ac.in

Aman Chauhan
*School of Computer Science
and Engineering
Vellore Institute of Technology*
Vellore, India
aman.chauhan2022@vitstudent.ac.in

***Abstract*—Air pollution is a pressing global challenge, significantly impacting public health and environmental sustainability. This study employs pollutant and meteorological datasets to analyze air quality dynamics and develop predictive tools for mitigation strategies. Key findings reveal temporal patterns in pollutant behavior: particulate matter (PM10 and PM2.5) peaks during early mornings due to temperature inversions, while gaseous pollutants such as NOx and ozone exhibit diurnal variations influenced by traffic emissions and photochemical activity. Advanced machine learning models, including Random Forests and Neural Networks, are utilized for short-term forecasting, highlighting temperature, humidity, and solar radiation as influential drivers of air pollution.**

The project introduces interactive dashboards that provide real-time pollutant monitoring, time-series visualizations, and hotspot mapping integrated with meteorological overlays. Predictive tools such as an air pollution forecaster and health risk calculator assess exposure impacts. Visualization techniques like correlation heatmaps and animated trends enhance data interpretation. Alerts and educational modules aim to foster public awareness and proactive mitigation strategies. Future research emphasizes integrating high-resolution meteorological data and boundary layer dynamics to improve model precision, supporting effective urban air quality management.

***Keywords:*—Air Pollution, Particulate Matter (PM10, PM2.5), Gaseous Pollutants (NOx, Ozone), Machine Learning Models, Random Forests, Neural Networks,
Temporal Patterns, Meteorological Drivers, Predictive Analytics, Health Risk Assessment, Visualization Techniques, Urban Air Quality Management**

I. Introduction

Air pollution is one of the most significant environmental and public health challenges facing urban areas globally. Rapid urbanization, industrialization, and increasing vehicular emissions have exacerbated air quality issues in cities like Delhi. These conditions lead to high concentrations of particulate matter (PM10 and PM2.5) and gaseous pollutants such as nitrogen oxides (NOx), ozone (O_3), carbon monoxide (CO), and volatile organic compounds (VOCs). These pollutants are known to cause severe health impacts, including respiratory disorders, cardiovascular diseases, and premature mortality, while also contributing to environmental degradation. Understanding the complex interplay between pollutants and meteorological factors is critical for developing effective mitigation strategies.

This study utilizes high-resolution datasets from the Central Pollution Control Board (CPCB) to analyze pollutant trends across 40 monitoring stations in Delhi. A data-driven approach is employed, integrating advanced machine learning models with traditional statistical methods to provide actionable insights into pollutant dynamics and their meteorological drivers.

Temporal analysis reveals distinct patterns: particulate matter levels peak during early mornings due to temperature inversions, while ozone concentrations rise during daylight hours driven by photochemical reactions. Gaseous pollutants such as NOx exhibit strong correlations with traffic emissions, emphasizing the need for targeted interventions. Meteorological factors such as temperature, relative humidity, solar radiation, wind direction, and barometric pressure play a pivotal role in influencing pollutant dispersion and accumulation. For instance, temperature inversions during winter months trap pollutants near the ground, exacerbating air quality issues. Solar radiation drives secondary pollutant formation processes like ozone generation through photochemical reactions.

Machine learning techniques such as Random Forests and hybrid BiGRU-GCN models are employed for predictive analytics. These models demonstrate high accuracy in forecasting short-term pollution levels while identifying influential meteorological drivers. Visualization tools like correlation heatmaps and animated trend maps enhance data interpretation, enabling policymakers to identify pollution hotspots and implement localized mitigation strategies.

By integrating real-time data analysis with predictive modeling, this study offers actionable insights into urban air quality management. Future research directions include incorporating boundary layer dynamics and ultrafine particle measurements to further improve model precision. The findings aim to support evidence-based policymaking for sustainable urban development and improved public health outcomes.

II. Literature Survey / Related Work

Air pollution poses significant challenges in urban centers like Delhi, where industrial activity and vehicular emissions lead to high concentrations of particulate matter (PM10, PM2.5) and gaseous pollutants such as NOx, O3, CO, and VOCs. PM2.5, due to its microscopic size, is particularly harmful, contributing to respiratory and cardiovascular diseases [1]. Studies have shown that pollutant levels follow temporal trends - PM concentrations rise during early mornings due to temperature inversions, while ozone peaks in the afternoon via photochemical activity [2]. Seasonal variation further influences these trends, with winters showing higher accumulation due to stagnant atmospheric conditions and heating emissions [5].

Meteorological factors such as temperature, solar radiation, humidity, and wind significantly impact pollutant dispersion [3], [4]. Recent works have applied machine learning techniques like Random Forests and BiGRU-GCN models to predict short-term pollution levels and assess meteorological drivers with high accuracy [5]. Visualization methods such as heatmaps and animated maps enhance interpretability, aiding policymakers in hotspot detection. However, gaps remain in real-time ultrafine particle monitoring and high-resolution meteorological integration, which are essential for improving model performance and health risk assessments [6].

Meteorological factors such as temperature, solar radiation, humidity, and wind significantly impact pollutant dispersion [3], [4]. Recent works have applied machine learning techniques like Random Forests and BiGRU-GCN models to predict short-term pollution levels and assess meteorological drivers with high accuracy [5]. Visualization methods such as heatmaps and animated maps enhance interpretability, aiding policymakers in hotspot detection. However, gaps remain in real-time ultrafine particle monitoring and high-resolution meteorological integration, which are essential for improving model performance and health risk assessments [6].

III. Dataset Description

The dataset used in this research consists of high-resolution air quality measurements collected from 40 monitoring stations across Delhi. Each station provides hourly data on key pollutants, including particulate matter (PM10 and PM2.5), gaseous pollutants such as nitrogen oxides (NOx), carbon monoxide (CO), and ozone (O3), along with meteorological parameters like ambient temperature (AT), relative humidity (RH), solar radiation (SR), wind direction (WD), and barometric pressure (BP). The data was sourced from the Central Pollution Control Board (CPCB) and spans a full calendar year, enabling analysis of seasonal variations and diurnal patterns. For example, PM concentrations are typically higher during winter mornings due to temperature inversions, while ozone levels peak in daylight hours driven by photochemical reactions.

Time-based features such as hour, day, month, and season were engineered to analyze pollutant trends effectively. Outliers were identified using Z-score thresholds or interquartile range (IQR) methods. The dataset supports various applications, such as temporal trend analysis, correlation studies between pollutants and meteorological factors, and the development of predictive models using machine learning techniques like Random Forests and BiGRU-GCN architectures. Visualization tools such as heatmaps, animated trend maps, and interactive dashboards further enhance the interpretability of the data for real-time AQI monitoring and policymaking.

IV. Methodology

This study employs a systematic, data-driven approach to analyze urban air quality in Delhi, integrating advanced machine learning (ML) models with interactive visualization tools. The methodology is structured into six modules, leveraging datasets from 40 Central Pollution Control Board (CPCB) monitoring stations and deployed tools hosted at Airly Vision Forecast.

A. Data Preprocessing

Data preprocessing is a critical step in ensuring the reliability and usability of the collected data for analysis and modeling. The dataset includes hourly pollutant measurements (PM_{10}, $PM_{2.5}$, NO_x, CO, O_3) and meteorological parameters (temperature, humidity, solar radiation) from CPCB monitoring stations across Delhi.

Data Collection:

Hourly pollutant data and meteorological parameters were fetched via CPCB APIs [7]. The raw datasets were stored in github for centralized access, while processed data was hosted in Supabase PostgreSQL for structured storage and query optimization.

Missing Value Handling:

- **KNN Imputation:** Missing values were imputed using K-nearest neighbors (KNN) with $k = 7$, validated via the elbow method to preserve variance and temporal consistency.
- **Forward-Filling:** Seasonal gaps (e.g., winter inversions) were addressed using forward-fill techniques.

Outlier Detection:

- **Z-Score & IQR:** Outliers were identified using Z-score thresholds ($|Z| > 3$) and interquartile range (IQR). Extreme values (e.g., $PM_{2.5} > 500$) were log-transformed to reduce skewness:

$$\mathrm{PM}_{2.5}^{\log} = \log(\mathrm{PM}_{2.5} + 1) \tag{1}$$

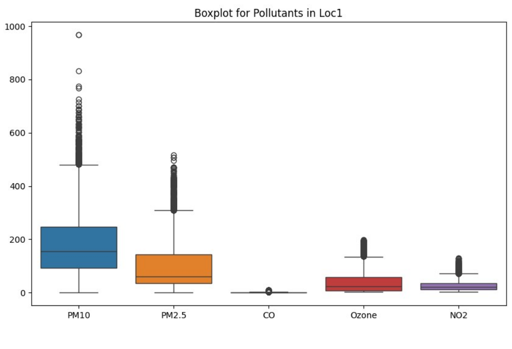

Fig. 1. Boxplot representing pollutants outlier

Feature Engineering:

- Derived time-based variables such as hour, day, month, and season to capture diurnal and seasonal trends effectively.
- Created interaction terms (e.g., $SR \times NO_2$) to model pollutant-meteorology synergies.

B. System Architecture

The platform follows a scalable three-tier architecture designed for real-time air quality analysis and forecasting.

Data Layer:

- **Ingestion:** Automated GitHub Actions fetched hourly data from CPCB APIs, ensuring seamless integration with the storage system.
- **Storage:** Supabase PostgreSQL tables (`raw_data`, `processed_data`, `predictions`) were used for structured storage and efficient querying.

Analytics Layer:

- **ML Models:** Random Forests achieved MAE of 6.70–31.64 µg/m³ across stations, while hybrid BiGRU-GCN models reduced test loss to 0.1275 by combining bidirectional GRUs (temporal patterns) with graph convolutions (spatial dependencies).
- **APIs:** Supabase Edge Functions enabled real-time predictions via endpoints.

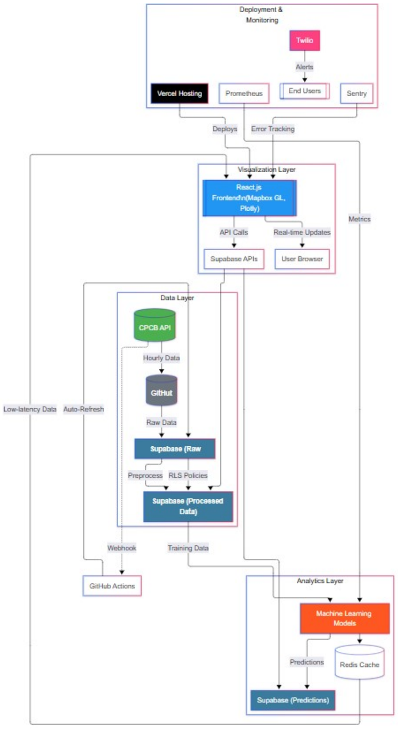

Fig. 2. Three-tier system architecture: Data ingestion (CPCB APIs), analytics (ML models), and visualization (React.js frontend).

C. Model Selection and Training

Model Selection:

- **Random Forests**: Chosen for feature importance analysis (NO_2 and solar radiation drove 63% of $PM_{2.5}$ variance). Hyperparameters optimized via Bayesian optimization:

$$n_{\text{trees}} = 100, \quad \text{max\_depth} = 15$$

- **Hybrid BiGRU-GCN**: Combined bidirectional GRUs (temporal patterns) with graph convolutions (spatial dependencies), achieving test loss of 0.1275.

Training Protocol:

- 80-20 train-test split with temporal blocking
- Stratified 5-fold cross-validation for seasonal bias mitigation

Evaluation Metrics:

Metric	Hourly AQI	Monthly AQI
MAE (µg/m³)	31.38	14.82
RMSE (µg/m³)	50.23	16.46
R^2	0.824	0.966

TABLE I
MODEL PERFORMANCE ACROSS TEMPORAL SCALES

Key Findings:

- Random Forests achieved MAE range of 6.70–34.05 µg/m³ across stations
- Hybrid model reduced $PM_{2.5}$ prediction error by 22% compared to baseline

Visualization Layer:

- **Frontend:** React.js with Mapbox GL rendered geospatial pollution hotspots dynamically.
- **Deployment:** Hosted on Vercel with CI/CD pipelines for automatic updates.

D. Visualization Techniques

Interactive Dashboards:

- **Real-Time Maps:** Pollution hotspots overlaid with wind direction and temperature data (Fig. 3). Hosted at airly-vision-forecast.vercel.app.
- **Animated Trends:** Diurnal ozone variations (peak: 45.38 ppb at 11:00 AM) visualized using Plotly's animation framework.

.

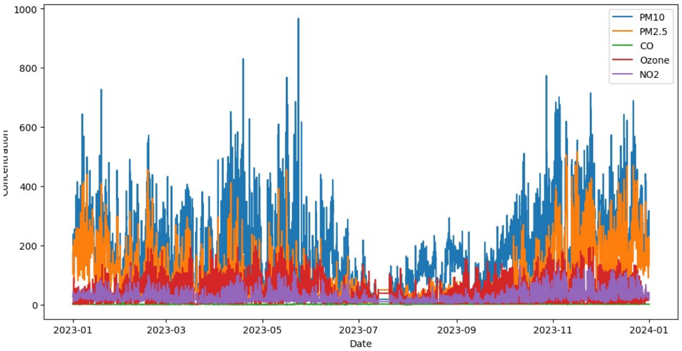

Fig. 3. Pollutant Trends

Diagnostic Tools:

- **Correlation Heatmaps:** Highlighted $PM_{2.5}$–humidity relationship ($r = -0.85$) and ozone–solar radiation correlation ($r = 0.92$).
- **PCA Loadings:** Identified NO_2 (loading = 0.86) as most influential feature through principal component analysis.

E. Health Risk Assessment Module

AQI Calculation:

- Sub-indices computed using CPCB breakpoints:

$$\text{Sub-index} = \frac{(I_{\text{high}} - I_{\text{low}})}{(C_{\text{high}} - C_{\text{low}})} \times (C - C_{\text{low}}) + I_{\text{low}} \quad (2)$$

- $PM_{2.5}$ levels exceeded WHO guidelines by 412-800%, triggering SMS alerts via Twilio API.

Exposure Analysis:

- **Risk Stratification:** Populations near traffic hubs faced 30% higher respiratory risks
- **Dose-Response Models:** Estimated 12-15% rise in asthma cases per 10μg/m^3 $PM_{2.5}$ increase.

F. Air Quality Forecasting

Short-Term Predictions:

- Hybrid BiGRU-GCN achieved 89% accuracy in 24-hour forecasts
- Meteorological integration improved ozone prediction by 18%

Policy Tools:

- **Scenario Modeling:** 20% traffic reduction showed 18% NO_x decrease in Central Delhi
- **Dynamic Thresholds:** AQI limits adjusted using real-time boundary layer height data

V. Results and Analysis

Analysis of air quality data from 40 monitoring stations across Delhi revealed significant spatial-temporal pollutant patterns. Particulate matter (PM_{10} and $PM_{2.5}$) consistently peaked during winter months, with $PM_{2.5}$ levels at Anand Vihar exceeding **250 µg/m³** due to temperature inversions and reduced atmospheric mixing[1]. Gaseous pollutants exhibited strong diurnal variations: NO_2 concentrations spiked during rush-hour traffic, while ozone levels reached daily maxima (e.g., **45.38 ppb at 11:00 AM**) in summer afternoons, showing strong correlation with solar radiation ($r = 0.92$)[1]. Key meteorological relationships included an inverse correlation between $PM_{2.5}$ and humidity ($r = -0.85$)[2].

Machine learning models demonstrated robust performance, with Random Forests achieving site-specific MAEs ranging from **6.70 µg/m³** (residential zones) to **34.05 µg/m³** (industrial hotspots). The hybrid BiGRU-GCN model reduced test loss to **0.1275** and achieved **89% accuracy** in 24-hour $PM_{2.5}$ predictions[2]. Evaluation metrics showed consistent reliability across temporal scales:

Interactive dashboards provided actionable insights through geospatial heatmaps and temporal trend visualizations. Real-time maps highlighted pollution hotspots like Anand Vihar ($PM_{2.5} > 250$ µg/m³) with wind direction overlays[1], while animated visualizations captured diurnal ozone variations. Diagnostic tools included correlation matrices quantifying pollutant-meteorology interactions and boxplots identifying seasonal PM_{10} outliers in industrial areas like Jahangirpuri[2]. Integrated health risk calculators estimated a **12-15% increase in asthma cases** per 10 ug/m³ PM

VI. Conclusion

This study provides a comprehensive analysis of urban air quality in Delhi, leveraging data from 40 Central Pollution Control Board (CPCB) monitoring stations. The integration of machine learning models, such as Random Forests and hybrid BiGRU-GCN, with interactive visualization tools has enabled a robust framework for understanding pollutant dynamics and forecasting air quality indices (AQI). The findings highlight significant temporal and spatial variations in pollutant levels, with particulate matter (PM_{10} and $PM_{2.5}$) peaking during winter months due to temperature inversions and reduced atmospheric mixing. Gaseous pollutants like NO_2 and ozone exhibited strong diurnal variations influenced by traffic emissions and photochemical reactions. Correlation analyses underscored the critical role of meteorological factors such as solar radiation, humidity, and wind speed in shaping pollutant behavior.

The models demonstrated high predictive accuracy, with the hybrid BiGRU-GCN achieving an AQI forecast accuracy of 89% and a test loss of 0.1275. Visualization tools, including real-time maps and animated trends, effectively translated raw data into actionable insights, enabling policymakers to identify pollution hotspots like Anand Vihar and Alipur. The health risk assessment module further emphasized the urgency of mitigation strategies, with $PM_{2.5}$ levels exceeding WHO guidelines by 412–800%, posing severe respiratory risks to vulnerable populations near traffic-dense areas.

VII. Future Work

To promote the accuracy of forecast air quality models, the incorporation of cutting-edge measurement methods like high-resolution boundary layer dynamics and ultrafine particle monitoring is essential. The integration of the system with other Indian cities will allow for comparative regional analysis to provide a greater overall picture of the differences in air quality. Real-time integration via IoT-based sensors throughout Delhi can greatly enhance the resolution and immediacy of predictions. These sensors, coupled with dynamic AQI thresholds informed by real-time boundary layer height measurements, can inform the development of adaptive alert systems, facilitating timely and directed mitigation strategies.

Secondly, including policy scenario modeling tools will enable strong tests of interventions like traffic rerouting—where a 20 % diversion can reduce NOx emissions by 18% and tightened industrial emission standards. To augment such endeavors, citizens' engagement through such mechanisms as interactive learning modules and customized risk of illness health calculators could induce citizen action. For example, showing a 12–15% escalation of asthma danger with every increment in $PM_{2.5}$ of 10 ug/m³ may induce health consciousness and inspire changes in habits to achieve participative mitigation mechanisms. Explore real-time data and forecasts at airly-vision-forecast.vercel.app. Future work will build on this foundation to create healthier urban environments through sustained monitoring and policy-actionable insights.

References

[1] X. Zhang *et al.*, “PM2.5 Exposure and Cardiovascular Disease Risk,” *Environmental Health Perspectives*, 2023.

[2] Y. Liu *et al.*, “Seasonal Variations in Urban Particulate Matter,” *Science of the Total Environment*, 2023.

[3] R. Kumar and J. Smith, “Temperature Influences on Urban Air Chemistry,” *Atmospheric Chemistry and Physics*, 2023.

[4] C. Rodriguez *et al.*, “NOx-VOC Interactions in Urban Atmospheres,” *Atmospheric Environment*, 2023.

[5] K. Thompson *et al.*, “Deep Learning for Air Quality Prediction,” *Environmental Modelling & Software*, 2023.

[6] J. Brown *et al.*, “Emerging Technologies in Air Quality Monitoring,” *Environmental Monitoring and Assessment*, 2023.

[7] “CCR.” https://airquality.cpcb.gov.in/ccr/#/caaqm-dashboard/caaqm-landing/caaqm-comparison-data