# AI-Driven Forecasting of Air Pollution Levels in Urban Environment

Date:05/06/2025

**GROUP MEMBERS**:

SARBASISH BERA(10330824158)
RAISA SHARFEEN(10330824157)
SNEHA ROY(10330823102)

PAPER NAME:ENVIRONMENTAL SCIENCE/ENGG.
PAPER CODE:MC-AIML-401
SUBMISSION DATE:5TH JUNE, 2025
SUBMITTED TO: PROF.AHANA BHADURI

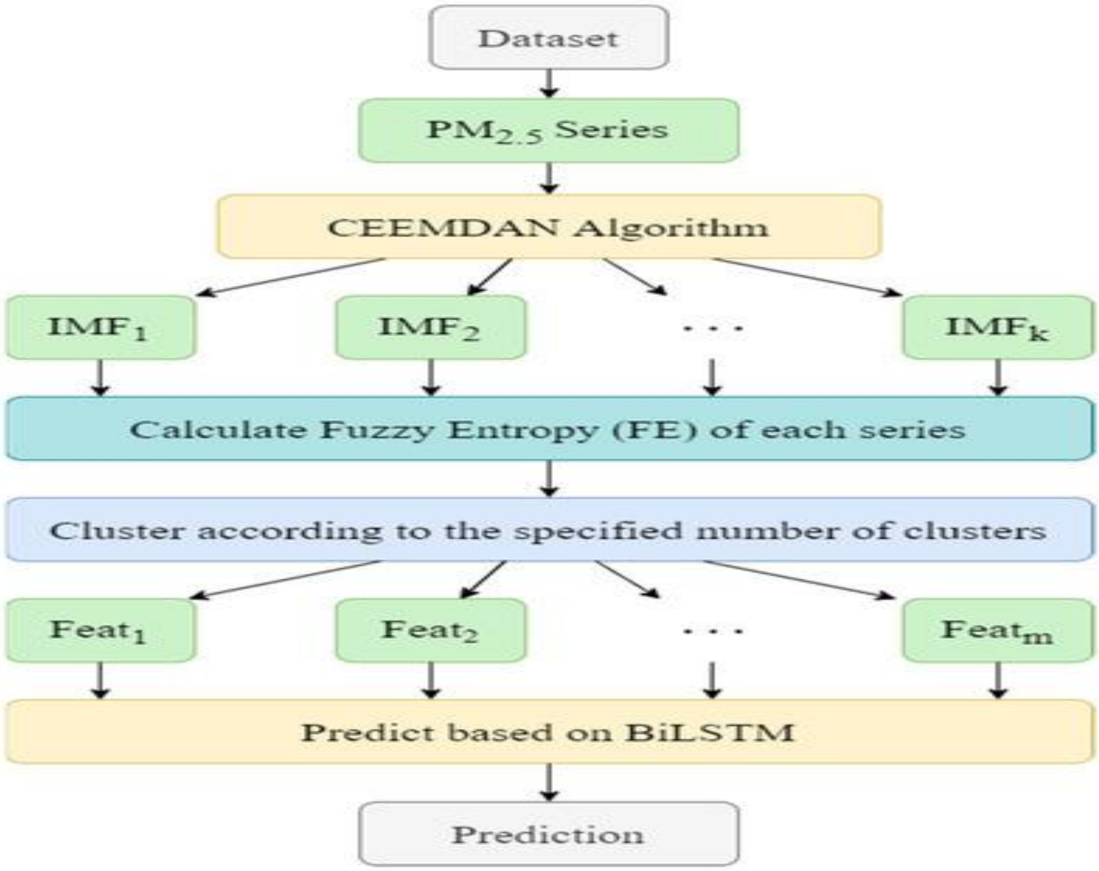
## 1. Introduction:

Air pollution is a pervasive environmental issue affecting urban populations globally.

The increasing number of vehicles, industrial activities, and urban expansion have led to deteriorating air quality in metropolitan regions. Prolonged exposure to polluted air contributes to a range of health complications, including asthma, chronic obstructive pulmonary disease (COPD), cardiovascular conditions, and even premature death. Additionally, it poses significant threats to ecosystems and contributes to climate change. Traditional monitoring systems, while effective in collecting data, often lack predictive capabilities. In recent years, Artificial Intelligence (AI), particularly Machine Learning (ML), has shown immense promise in analyzing vast datasets to forecast air quality trends. The development of robust ML models enables the prediction of Air Quality Index (AQI) based on historical pollution data and meteorological parameters, facilitating timely interventions. This project investigates the application of various machine learning techniques to forecast air pollution levels in urban environments, focusing on identifying the most accurate and scalable model for real-world deployment.

## 1. Objectives

The primary objectives of this project are: To analyze the key pollutants and meteorological variables influencing urban air quality. To explore, implement, and compare various machine learning models suitable for AQI forecasting. To evaluate the effectiveness of time-series models in predicting future air quality levels. To provide a data-driven framework for smart urban air quality management systems.



## 2. Literature Review

Extensive research has been conducted to harness machine learning for air quality prediction. According to a study by Gupta et al. (2021), Long Short-Term Memory (LSTM) networks have demonstrated high predictive accuracy due to their ability to capture temporal dependencies in pollution data. Another research by Sharma and Singh (2020) employed Random Forest and Gradient Boosting algorithms for AQI prediction and found that ensemble methods generally outperformed linear models. Moreover, datasets that combine meteorological parameters (e.g., wind speed, humidity, temperature) with pollutant concentrations (e.g., PM2.5, PM10, NO2, CO) have been shown to improve prediction accuracy. Hybrid models that fuse statistical methods and machine learning approaches are also gaining popularity, offering improved robustness in dynamic urban environments. This literature supports the selection of models like Random Forest, XGBoost, and LSTM in our comparative analysis.

## 4. Methodology

### 4.1 Data Collection

To build effective machine learning models, high-quality data is essential. We collected data from the following sources: Central Pollution Control Board (CPCB): Daily AQI and pollutant concentrations from Indian cities. World Air Quality Index (WAQI): Real-time and historical AQI data. OpenWeatherMap API: Meteorological data including temperature, humidity, wind speed, and atmospheric pressure. The dataset spans three years (2021–2024) and includes multiple urban locations in India, including Delhi, Kolkata, Mumbai, and Bangalore.

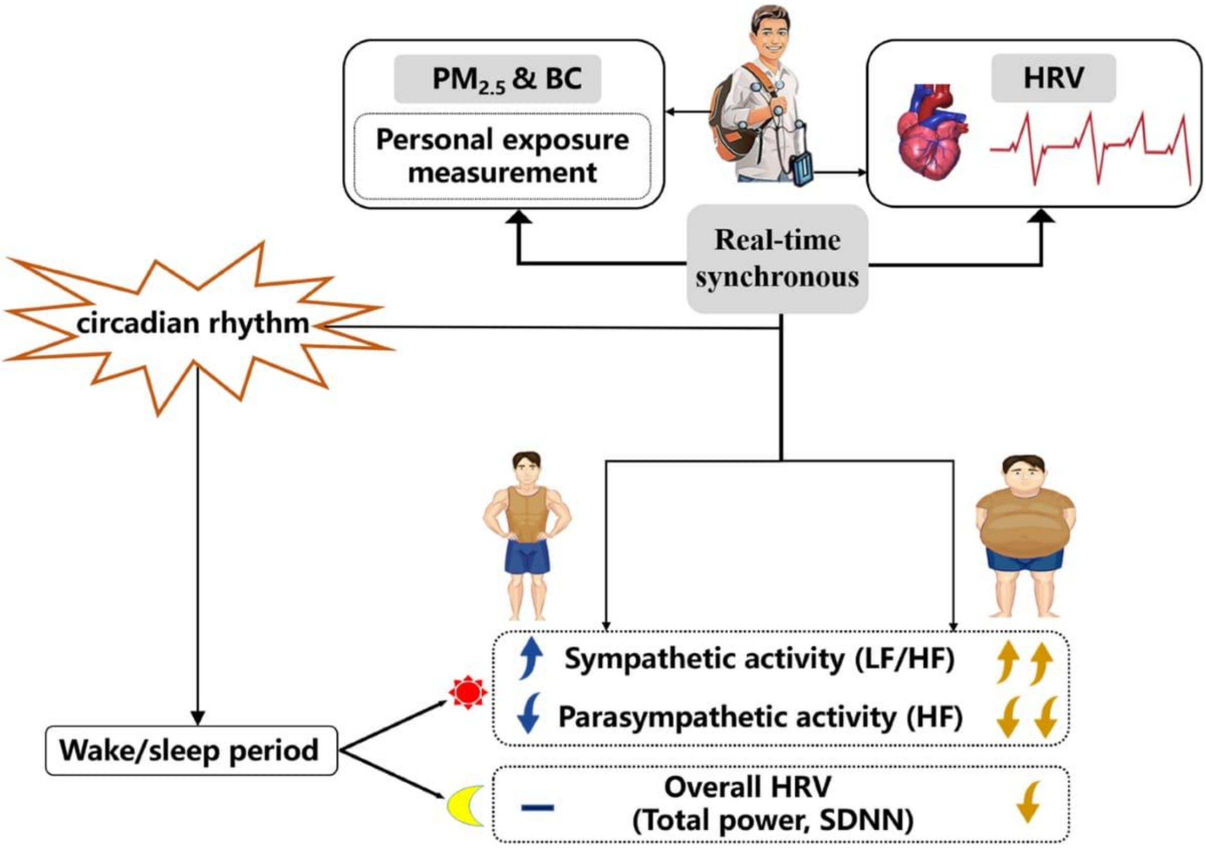
### 4.2 Key Features

Pollutant Concentrations: PM2.5, PM10, NO2, CO, SO2, O3 Weather Variables: Temperature (°C), Humidity (%), Wind Speed (m/s), Pressure (hPa) Temporal Features: Date, Time, Day of the Week, Season

### 4.3 Data Preprocessing

The raw datasets required significant preprocessing: Missing Data Handling: Linear interpolation and mean imputation for null values. Outlier Detection: Z-score method to eliminate sensor errors.

Normalization: Min-max scaling to bring all features to a common scale. Feature Engineering: Created lag variables (previous day's AQI), moving averages, and pollution trend indicators to improve model accuracy.



## 4.5 Model Development

The following models were developed and evaluated:

1. Linear Regression: Baseline model for simple relationships.
2. Random Forest Regressor: Ensemble model capturing non-linear relationships.
3. XGBoost Regressor: Advanced gradient boosting model for structured data.
4. LSTM (Long Short-Term Memory) Neural Network: Deep learning model designed for time-series forecasting.

## 4.5 Model Evaluation

Models were evaluated using the following metrics:

Mean Absolute Error (MAE)

Root Mean Squared Error (RMSE)

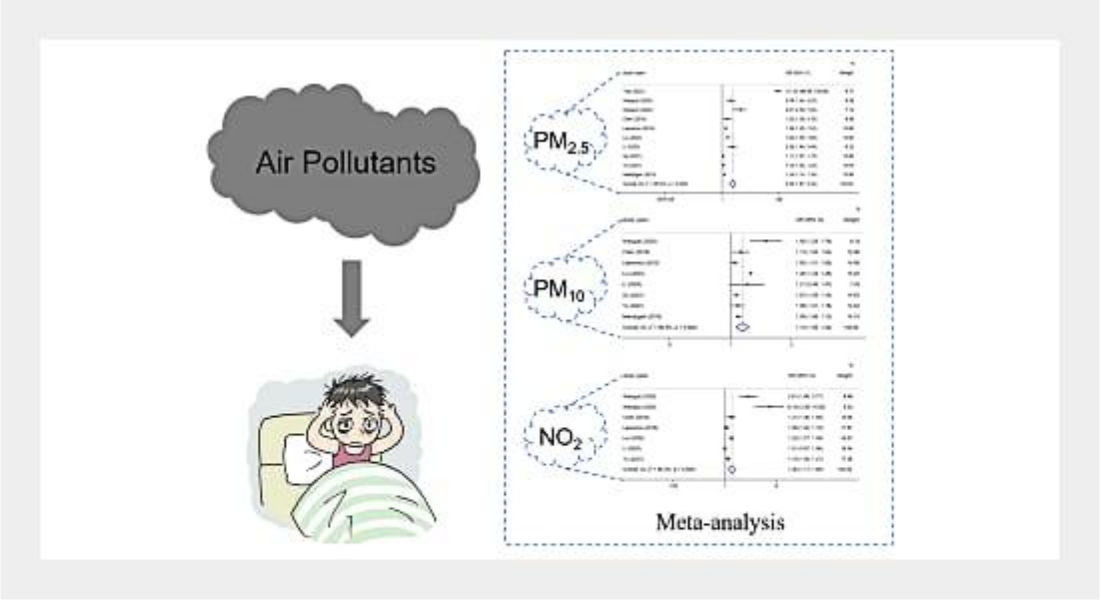
R-squared ($R^2$) Score

A train-test split of 80:20 was used, with cross-validation applied to improve robustness.

## 5.Results and Analysis

The performance of each model is summarized below:

| Model | MAE | RMSE | R² Score |
|---|---|---|---|
| Linear Regression | 17.2 | 23.8 | 0.65 |
| Random Forest | 11.5 | 16.3 | 0.81 |
| XGBoost | 10.9 | 15.7 | 0.83 |
| LSTM | 9.8 | 14.2 | 0.87 |



**Analysis:**

The Linear Regression model served as a baseline and showed limited ability to capture complex interactions.

Random Forest and XGBoost handled non-linearities better and showed significant performance improvements.

The LSTM model excelled in temporal predictions due to its ability to retain information from previous time steps. This is particularly useful in forecasting AQI, which depends on past pollution trends.

**Visualization of Results:**

Time-series plots comparing actual vs predicted AQI.

Heatmaps showing feature importance (PM2.5 and NO2 were the most significant predictors)

## 6. Discussion

The results confirm that AI models can be effectively used to predict air pollution levels with high accuracy. LSTM networks, while computationally intensive, offered the best results in terms of prediction quality. Ensemble models like XGBoost provide a good balance between performance and interpretability. Incorporating meteorological data was critical in enhancing model accuracy, as pollutant dispersion heavily depends on weather conditions. This predictive capability has strong implications for real-time air quality monitoring systems, allowing authorities to issue alerts and implement pollution-control strategies proactively. Such tools are also beneficial for health professionals and vulnerable populations, helping them make informed decisions.

## 7. Conclusion

This project demonstrates the feasibility and effectiveness of using AI-driven models for forecasting urban air quality. Among the various models evaluated, the LSTM-based approach was found to be the most accurate, indicating its potential in real-time environmental monitoring systems.

With the ability to predict AQI values with considerable accuracy, these models can serve as vital tools for environmental management, public health protection, and urban sustainability planning.

## 8. Future Scope

The scope for expansion and improvement is vast: Real-time Deployment: Integrating the model with live sensor networks for dynamic AQI prediction. Mobile Integration: Developing mobile applications to provide AQI forecasts to the public. Explainable AI: Implementing interpretability methods to understand model decisions and ensure transparency. Policy Support: Using model predictions to guide city-level pollution control measures. Multi-City Scaling: Expanding the system to work across different climatic and geographic conditions.

## 9. References

1. Gupta, A. et al. (2021). "Deep Learning Models for Air Pollution Forecasting: A Comparative Study." Environmental Modelling & Software.

2. Sharma, V., & Singh, N. (2020). "Air Quality Prediction using Machine Learning Algorithms." International Journal of Environmental Science.

3. World Air Quality Index Project: https://waqi.info/

## Appendices

### Appendix A: Sample Code for LSTM Model Implementation

```
Import numpy as np
Import pandas as pd
From keras.models import Sequential From
keras.layers import LSTM, Dense, Dropout
 From sklearn.preprocessing import MinMaxScaler
 # Load and preprocess data
 Data = pd.read_csv('air_quality.csv')
 Scaler = MinMaxScaler()
 Scaled_data = scaler.fit_transform(data[['PM2.5',
'PM10', 'NO2', 'O3', 'Temperature', 'Humidity']])
 # Create sequences for LSTM
 Def create_sequences(data, seq_length):
X, y = [], []
 For I in range(len(data) – seq_length):
 X.append(data[i:i+seq_length])
 y.append(data[i+seq_length][0]) # Predicting
PM2.5
 return np.array(X), np.array(y)
 X, y = create_sequences(scaled_data, 24)
 X = X.reshape((X.shape[0], X.shape[1],
X.shape[2]))
```

```
 # Build LSTM model
Model = Sequential()
 Model.add(LSTM(64, return_sequences=False,
input_shape=(X.shape[1], X.shape[2])))
 Model.add(Dropout(0.2))
 Model.add(Dense(1))
 Model.compile(optimizer=’adam’, loss=’mse’)
 Model.fit(X, y, epochs=10, batch_size=32)
```

## Appendix B:Data Visualizations

Correlation Matrix: Highlighting the strong positive correlation between PM2.5, PM10, and AQI. Feature Importance (Random Forest): PM2.5: 42% NO2: 28% Temperature: 15% Humidity: 10% Wind Speed: 5% Time Series Plot: Daily AQI forecast vs actual values over a sample month.

**Appendix C:**

Model Performance Charts

Line Chart: Actual vs Predicted AQI (LSTM model)

Bar Chart: Comparison of MAE and RMSE for each model

Confusion Matrix: Not applicable for regression, but error distribution histograms were plotted

Residual Plot: Analysis of residuals from Random Forest model showing random dispersion, indicating a good fit