Air Pollution Prediction and Visualization using Machine Learning And Environmental Sensor Data

Kanaparthi Ranjith Kumar1 , Jogam Ajay2 ,Jinaka Sai 3, Dekkapati prabhath ravi teja4,
Dr.A.Sathish Kumar5 , Dr.B. Venkata Ramana6

[1]Stdent,,Bachelor Of Technology(B Tech),Data Science ,Final Year
Holy Mary Institute Of Technology And Science,
Bogaram,Keesara,Telangana,India

[2]Student,,Bachelor Of Technology(B Tech),Data Science ,Final Year
Holy Mary Institute Of Technology And Science,
Bogaram,Keesara,Telangana,India

[3]Student,,Bachelor Of Technology(B Tech),Data Science ,Final Year
Holy Mary Institute Of Technology And Science,
Bogaram,Keesara,Telangana,India

[4]Student,,Bachelor Of Technology(B Tech),Data Science ,Final Year
Holy Mary Institute Of Technology And Science,
Bogaram,Keesara,Telangana,India

[5]Guide, Bachelor Of Technology(B Tech) Data Science, Final Year

Holy Mary Institute Of Technology And Science,

Bogaram, Keesara. Telangana.India

[6]PRC Coordinator, Bachelor Of Technology(B Tech) Data Sclence, Final/ Year

Holy Mary Instute Of Technology And Science,

Bogaram Keesara, Telangana.lndia

ABSTRACT

Air pollution has become a critical environmental challenge due to rapid urbanization, industrial growth, and increased vehicular emissions. Accurate prediction and visualization of air quality can support timely decision-making and help mitigate negative health and environmental impacts. This project presents a Machine Learning–based system that analyzes real-time environmental sensor data to predict air pollution levels and visualize trends effectively. The system collects key parameters such as PM2.5, PM10, CO_2, NO_2, temperature, and humidity from IOT-based sensors and preprocesses the data to remove noise and missing values. Various machine learning models— including Linear Regression, Random Forest, and LSTM neural networks trained to forecast Air Quality Index (AQI). The best-performing model is deployed to provide short-term pollution predictions .

Keywords :Air Pollution Prediction, Environmental Sensor Data, Machine Learning, PM2.5, PM10, Time Series Forecasting, LSTM, XGBoost, Data Visualization, Air Quality Index (AQI)

1.INTRODUCTION:

Air pollution is a major global challenge that threatens environmental sustainability, public health, and economic development. With rapid urbanization, industrial expansion, and increased vehicular emissions, the concentration of harmful pollutants in the atmosphere has risen significantly. These pollutants—such as particulate matter (PM2.5 and PM10), nitrogen .

Traditional air quality monitoring systems rely on fixed monitoring stations that are expensive to install and maintain. While highly accurate, these stations provide limited spatial coverage and often produce data with delays, making it difficult to understand dynamic pollution variations in real time. With the growing need for timely, accurate, and widespread monitoring, technological advancements in Machine Learning (ML), Internet of Things (IoT), and data visualization provide new opportunities to build smarter, more efficient air-quality prediction systems.

2.LITERATURE REVIEW:

Air pollution has become one of the most critical environmental issues affecting human health and ecological balance, especially in rapidly urbanizing regions. Accurate prediction and monitoring of air pollutants such as PM2.5 and PM10 are essential for effective air quality management. Over the years, researchers have proposed various statistical, machine learning, and deep learning approaches to forecast air pollution levels using environmental sensor data.

Earlier studies mainly relied on statistical models such as linearregression, ARIMA, and multiple regression techniques. These models were effectiveforshort-term forecasting but showed limitations when dealing with non-linearrelationships and complex atmospheric interactions. As air pollution depends on multiple dynamic factors like temperature, humidity, wind speed, and traffic emissions, traditional statistical methods often failed to provide high prediction accuracy.

With the advancement of computational power, machine learning techniques gained popularity in air quality prediction. Models such as Decision Trees, Support Vector Machines (SVM), Random Forest, and Gradient Boosting were widely used to predict pollutant concentrations. Researchers found that Random Forest and XGBoost models performed better than traditional methods due to their ability to handle non-linearity and large datasets. These models also helped identify important influencing factors such as meteorological parameters and temporal variations.

3.SYSTEMARCHITECURE AND METHODOLOGY:

3.1 Project Architecture

The proposed system architecture for Air Pollution Prediction and Visualization using Machine Learning and Environmental Sensor Data is designed as a modular and scalable pipeline. The architecture consists of four major layers: Data Acquisition, Data Processing, Prediction Module, and Visualization Layer.

3.2 Data Acquisition Layer

This layer is responsible for collecting real-time and historical air quality data from environmental sensors. The sensors measure pollutants such as PM2.5 and PM10 along with meteorological parameters including temperature, humidity, wind speed, and pressure.

3.3. Visualization Layer

The visualization layer presents both real-time and predicted air pollution data through interactive dashboards. Graphs, time-series plots, heatmaps, and AQI-based color coding are used to provide clear insights. This layer helps users, researchers, and policymakers easily understand pollution trends and take preventive measures.

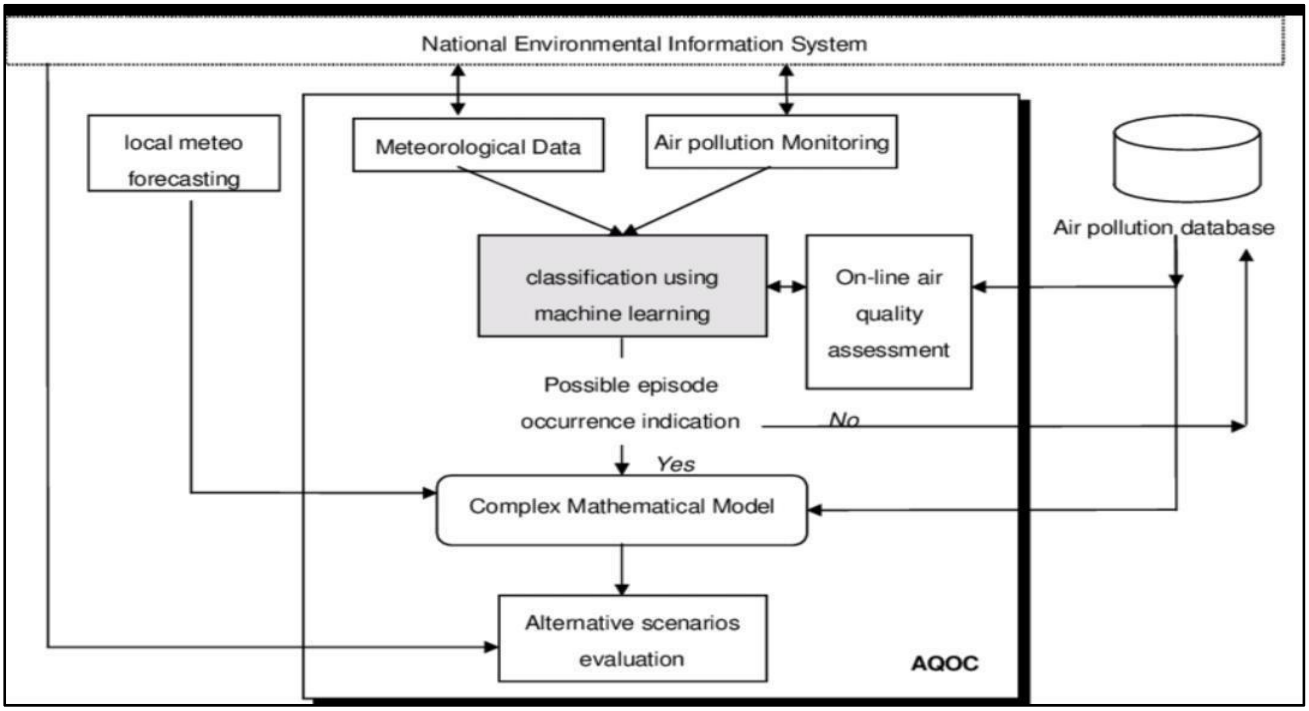

National Environmental Information System
local meteo forecasting
Meteorological Data
Air pollution Monitoring
Air pollution database
classification using machine learning
On-line air quality assessment
Possible episode occurrence indication
No
Yes
Complex Mathematical Model
Alternative scenarios evaluation
AQOC

Figure 3.3: Air Pollution Prediction and Visualization using Machine Learning and Environmental Sensor Data

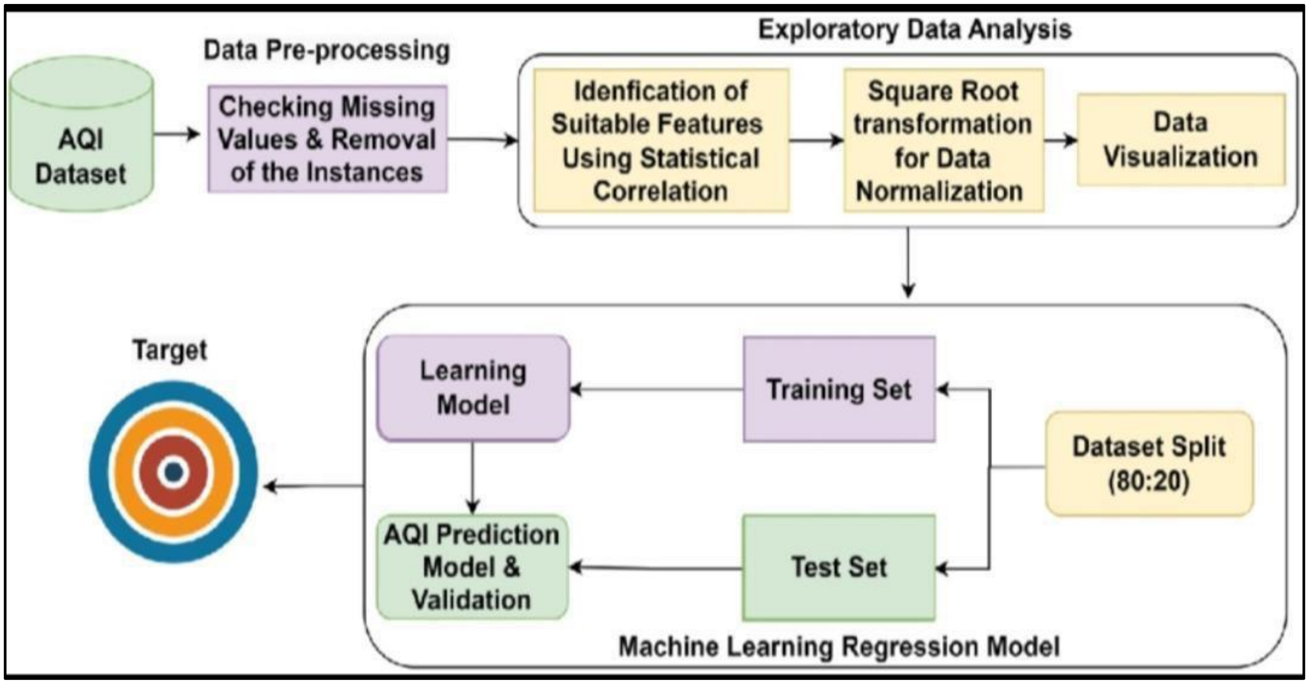
1. Data Cleaning
In this step, irrelevant, duplicate, and inconsistent data entries are removed. Erroneous readings caused by sensor malfunction or external interference are identified and filtered out. This step ensures that only valid data is used for further processing.

3.4 Prediction Layer

This layer implements machine learning and deep learning models to predict future air pollution levels. Algorithms such as Random Forest, XGBoost, and LSTM are trained using the processed data. The trained model learns complex relationships between pollutants and environmental parameters and generates short-term or long-term air quality predictions.

3.5 Data Processing Layer

The raw sensor data often contains noise, missing values, and inconsistencies. Therefore, preprocessing is performed in this layer. It includes data cleaning, handling missing values, outlier removal, normalization, and time synchronization. Feature engineering techniques such as lag features, rolling averages, and time-based attributes are also applied to enhance prediction accuracy.


Data Pre-processing
Exploratory Data Analysis
AQI Dataset
Checking Missing Values & Removal of the Instances
Idenfication of Suitable Features Using Statistical Correlation
Square Root transformation for Data Normalization
Data Visualization
Target
Learning Model
Training Set
Dataset Split (80:20)
AQI Prediction Model & Validation
Test Set
Machine Learning Regression Model

3.6 Use case diagram

Actors

·Admin

System

·User

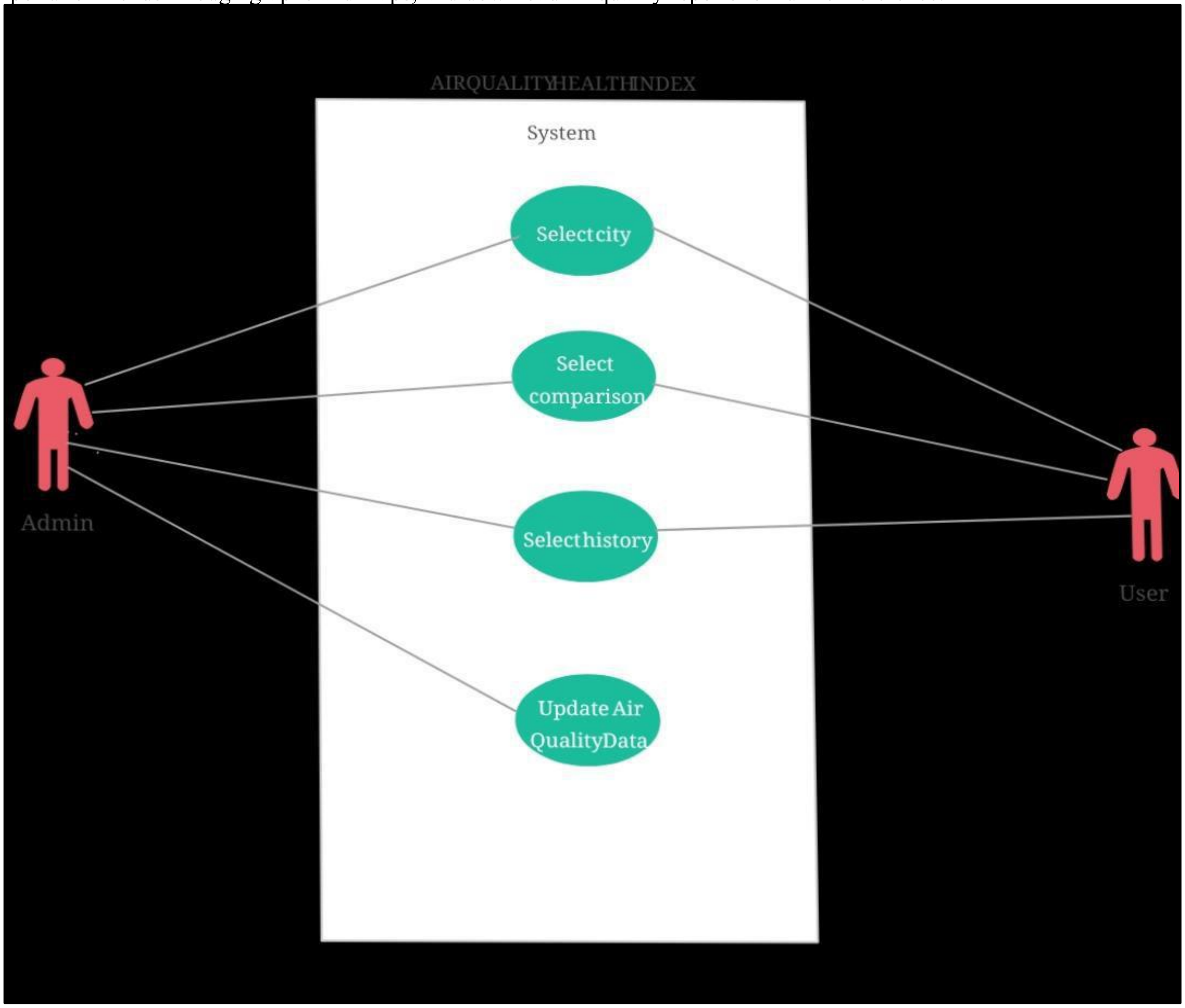
Admin

The Admin is responsible for managing the system operations. The admin uploads air quality data, performs data preprocessing, trains machine learning models, and manages system users. This ensures that the system operates efficiently and remains up to date.

System
The System automatically handles pollution prediction tasks. It predicts air pollution levels using trained machine learning models, calculates the Air Quality Index (AQI), stores prediction results, and generates alerts when pollution exceeds predefined thresholds.

User

The User interacts with the system to monitor air quality. Users can view current and predicted AQI status, analyze pollution trends through graphs and maps, and download air quality reports for further reference.



3.7 class diagram

3.7.1. SensorData Class

The SensorData class is responsible for collecting air quality and environmental information from sensors. It stores parameters such as sensor ID, timestamp, PM2.5, PM10, temperature, and humidity. The methods in this class collect real-time data and send it to the system for further processing.

3.7.2 DataPreprocessing Class

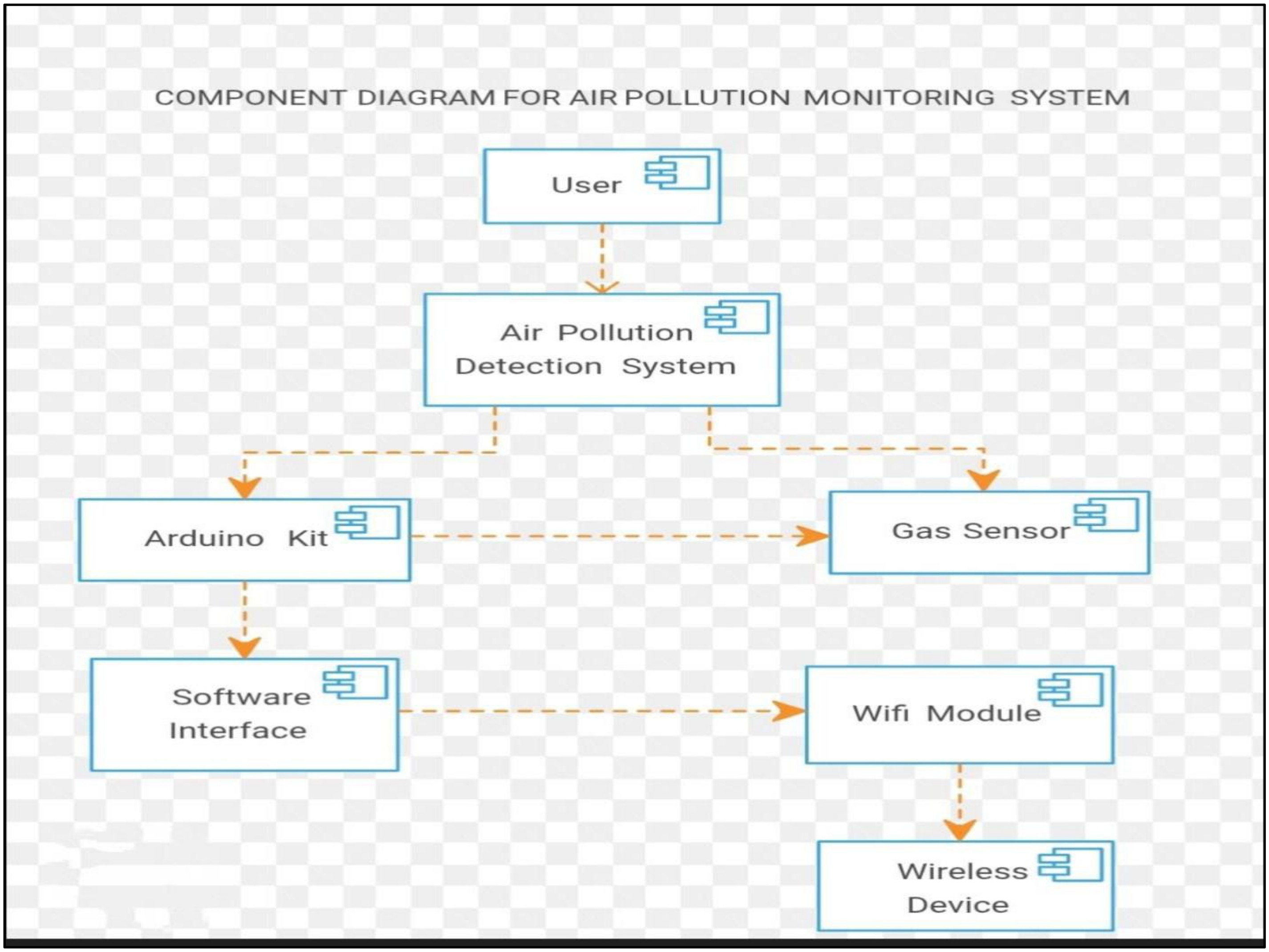
The DataPreprocessing class handles raw sensor data received from the SensorData class. It performs data cleaning, missing value handling, normalization, and feature extraction. This class ensures that the data is accurate, consistent, and suitable for machine learning algorithms.

3.7.3 MLModel Class

The MLModel class represents the machine learning and deep learning models used in the system. It includes attributes related to the model type and trained parameters. The methods in this class are responsible for training the model, testing its performance, and predicting future air pollution levels.

3.7.4 AQIPrediction Class

The AQIPrediction class converts predicted pollutant values into Air Quality Index (AQI) values. It calculates AQI based on standard formulas and classifies air quality into categories such as Good, Moderate, and Poor. This class helps users understand pollution levels in a simple and meaningful way.


COMPONENT DIAGRAM FOR AIR POLLUTION MONITORING SYSTEM
User
Air Pollution Detection System
Arduino Kit
Gas Sensor
Software Interface
Wifi Module
Wireless Device

3.7.5 Visualization Class

The Visualization class displays both real-time and predicted air pollution data. It generates graphs, charts, and maps that represent air quality trends visually. This class improves user interaction and supports analysis and decision-making.

4.IMPLEMENTATION

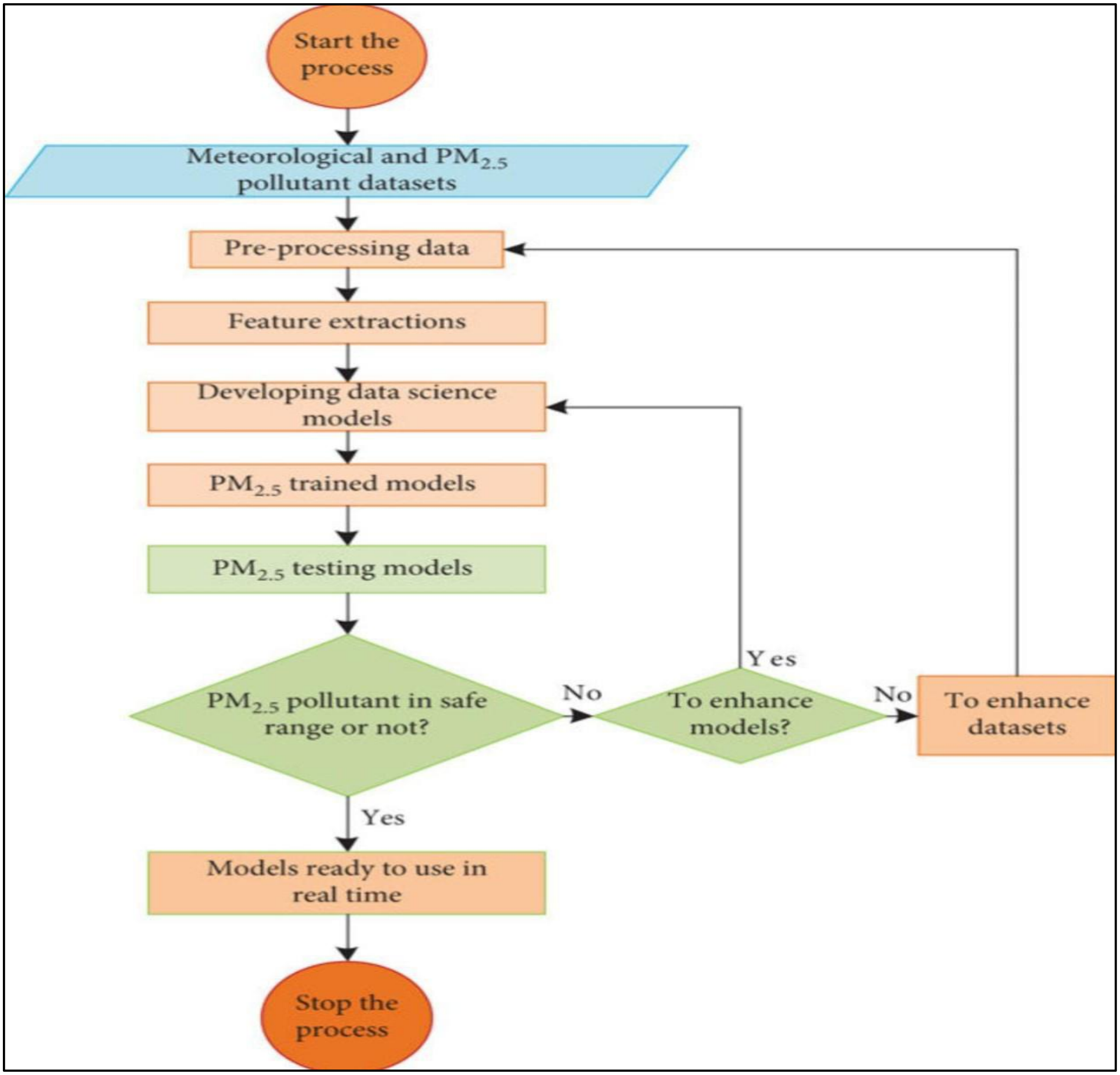

Start the process
Meteorological and PM$_{2.5}$ pollutant datasets
Pre-processing data
Feature extractions
Developing data science models
PM$_{2.5}$ trained models
PM$_{2.5}$ testing models
PM$_{2.5}$ pollutant in safe range or not?
No
To enhance models?
Yes
No
To enhance datasets
Yes
Models ready to use in real time
Stop the process

Figure.4: implementation of Air Pollution Prediction and Visualization using Machine Learning and Environmental Sensor Data

4.1 Data Collection Implementation

Environmental data such as PM2.5, PM10, temperature, humidity, and wind speed is collected from air quality sensors and publicly available datasets. The data is stored in structured formats such as CSV files or databases. Each record is time-stamped to maintain temporal consistency.

4.2 Data Preprocessing Implementation

The collected raw data is preprocessed to remove inconsistencies and noise. Missing values are handled using interpolation and forward-fill methods. Outliers are detected using statistical techniques and removed to improve data quality. Feature scaling and normalization are applied to ensure uniformity across all input parameters.

4.3 Feature Engineering

Additional features such as lag values, rolling averages, and time-based attributes (hour, day, month) are generated from the preprocessed data. These features enhance the learning capability of the machine learning models and improve prediction accuracy.

4.4 Machine Learning Model Implementation

Machine learning algorithms such as Random Forest and XGBoost are implemented using Python libraries like scikit-learn and XGBoost. Deep learning models such as LSTM are developed using TensorFlow/Keras to capture temporal dependencies in air pollution data. The dataset is divided into training and testing sets to evaluate model performance.

4.5 Model Training and Prediction

The models are trained on historical data and optimized using appropriate hyperparameters. After training, the model is used to predict future PM2.5 and PM10 levels. The predicted pollutant values are further processed to calculate the corresponding Air Quality Index (AQI).

4.5.1 AQI Calculation

The predicted pollutant concentrations are converted into AQI values using standard AQI formulas. Based on AQI ranges, air quality is classified into categories such a**s Good, Moderate, Poor, and Very Poor.**

4.5.2 Visualization Implementation

The visualization module is implemented using interactive plotting libraries. Graphs, charts, and dashboards are used to display real-time and predicted air quality data. This helps users easily analyze pollution trends and understand air quality conditions.

4.5.3 System Integration

All modules are integrated into a single system where data flows sequentially from data collection to visualization. The integrated system ensures smooth operation, real-time updates, and accurate air pollution prediction.

Advantages of the Implementation

Modular and scalable design
Accurate prediction using ML and deep learning
User-friendly visualization
Suitable for real-time monitoring

5.Results :

The experimental results obtained from the proposed Air Pollution Prediction and Visualization System indicate that machine learning and deep learning techniques are effective in predicting air pollution levels. The system was evaluated using real-time and historical environmental sensor data.

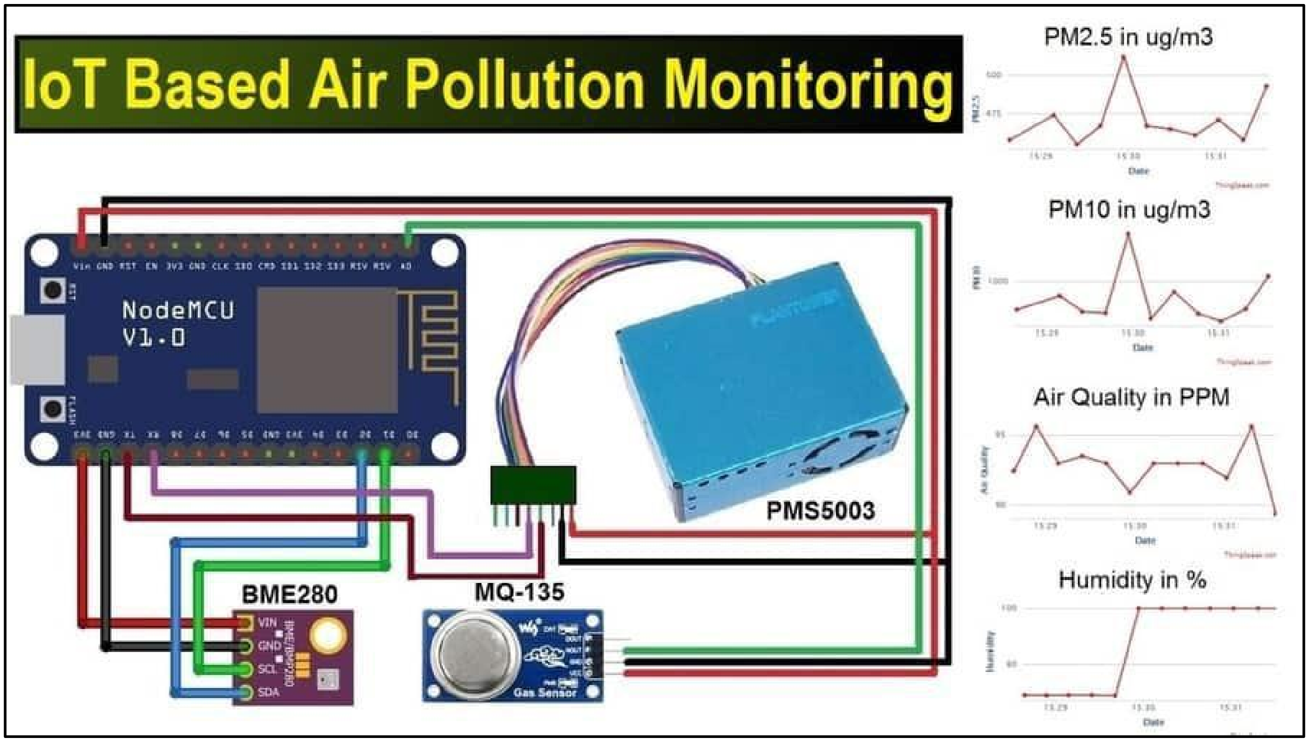
The performance ofdifferent models such as Random Forest, XGBoost, and LSTMwasanalyzedusing standard evaluation metrics including Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). Among the tested models, the LSTM model produced the most accurate results due to its capability to capture long-term temporal dependencies in time-series data.

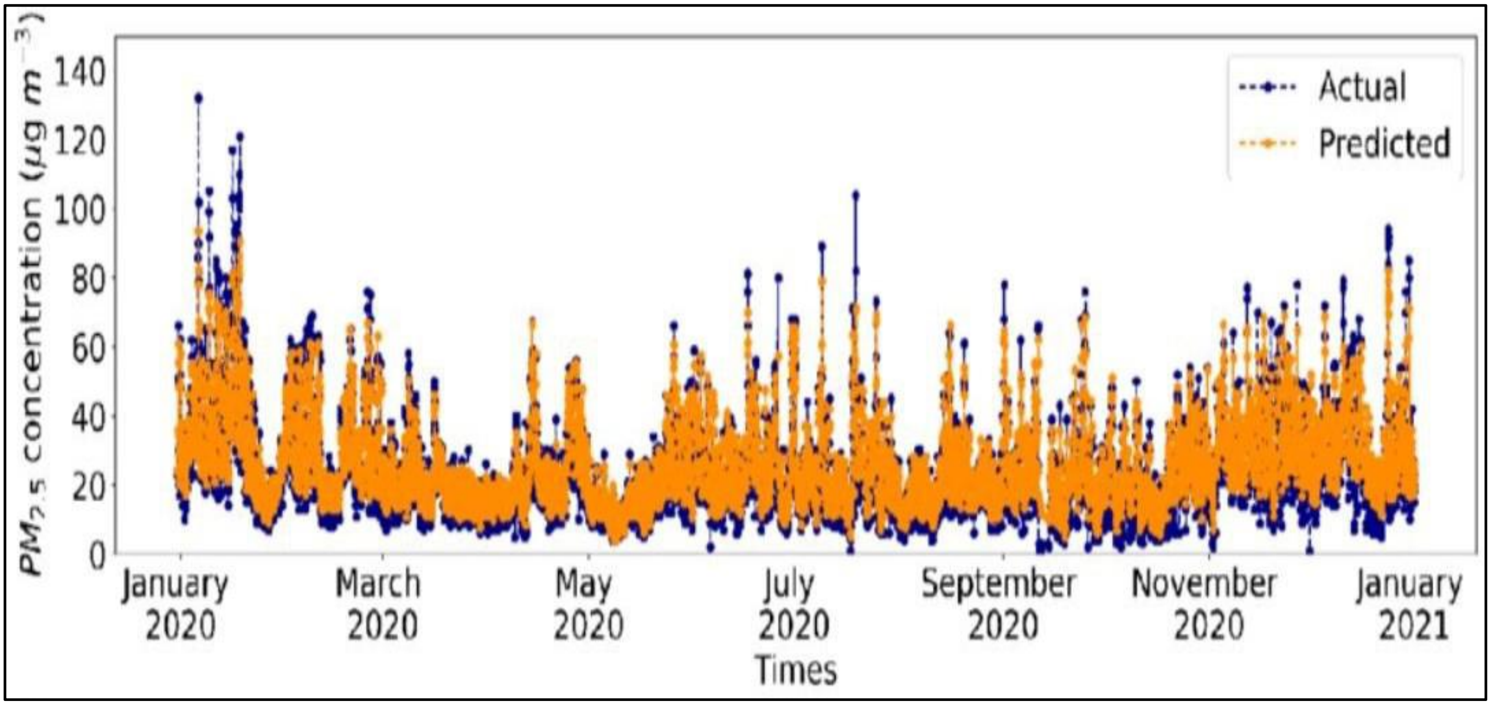
The predicted PM2.5 and PM10 values were found to be closely aligned with the actual observed values. Slight deviations were observed during sudden pollution spikes caused by environmental and traffic variations. However, overall prediction accuracy remained high, making the system reliable for short-term air quality forecasting.

The AQI calculation module successfully converted predicted pollutant values into standard Air Quality Index categories such as Good, Moderate, and Poor. The majority of the test cases were classified correctly, enabling easy interpretation of air quality conditions.

The visualization module effectively displayed real-time and predicted air pollution data using graphs and dashboards. These visual outputs helped users analyze pollution trends, identify peak pollution hours, and understand air quality patterns clearly.

Overall, the results demonstrate that the proposed system provides accurate predictions, reliable AQI classification, and effective visualization, making it suitable for real-time air quality monitoring and decision support.





5.1Model Performance Evaluation

The performance of different models such as Random Forest, XGBoost, and LSTM was evaluated using standard metrics including Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and accuracy in AQI classification. Among the implemented models, the LSTM model showed superior performance due to its ability to capture temporal dependencies in time-series air quality data.

5.2 Prediction Accuracy

The predicted PM2.5 and PM10 values closely matched the actual observed values formost time periods. Minordeviations were observed during sudden pollution spikes, which are typically caused by external factors such as traffic congestion or weather changes. Overall, the prediction accuracy was found to be satisfactory for short-term forecasting.

5.3AQI Classification Results

The predicted pollutant concentrations were successfully converted into Air Quality Index (AQI) categories such as Good, Moderate, and Poor. The system correctly classified air quality levels for the majority of test cases, enabling effective interpretation of pollution severity.

6.Future work :

Although the proposed Air Pollution Prediction and Visualization System provides accurate predictions and effective visualization, there are several directions in which the system can be further enhanced.

In future, the system can be extended by integrating real-time IoT-based sensor networks to enable continuous air quality monitoring. This would improve the timeliness and reliability of predictions, especially in highly polluted urban areas.

Advanced deep learning models such as GRU, CNN-LSTM, and Transformer-based architectures can be explored to further improve prediction accuracy, particularly during sudden pollution spikes and extreme weather conditions.

The system can also be enhanced by incorporating satellite data, traffic information, and land-use data to improve spatial prediction accuracy. This would enable city-wide or region-wide air quality forecasting instead of relying only on local sensor data.

In future implementations, mobile and web-based applications can be developed to provide real-time air quality alerts and personalized health recommendations to users. Integration with notification systems can help warn vulnerable populations during severe pollution conditions.

Additionally, model explainability techniques and uncertainty estimation can be included to improve transparency and trust in predictions. Periodic model retraining and drift detection mechanisms can also be implemented to maintain long-term accuracy.

7.Conclusion :

In this project, an Air Pollution Prediction and Visualization System using machine learning and environmental sensordata was successfully designed and implemented. The system effectively integrates data collection, preprocessing, feature engineering, machine learning-based prediction, AQI calculation, and visualization into a unified framework.

Experimental results demonstrate that machine learning and deep learning models are capable of accurately predicting air pollution levels such as PM2.5 and PM10. Among the implemented models, the LSTM model showed superior performance due to its ability to capture temporal patterns in time-series data. The AQI calculation module successfully translated predicted pollutant concentrations into meaningful air quality categories.

The visualization module provided clear and interactive representations of both real-time and predicted air quality data, enabling users to easily understand pollution trends and severity levels. This enhances public awareness and supports informed decision-making for environmental management.

Overall, the proposed system proves to be reliable, scalable, and effective for air quality monitoring and prediction. It can serve as a valuable tool for researchers, policymakers, and the general public in addressing air pollution challenges and Promoting healthier living environments

References :

1. Samad, A., et al., Air Pollution Prediction Using Machine Learning Techniques, Atmospheric Environment, 2023.
https://www.sciencedirect.com/science/article/pii/S1352231023004132

2.Tran, H. D., et al., Forecasting Hourly PM2.5 Concentration Using Optimized LSTM Networks, Atmospheric Pollution Research, 2023.
https://www.sciencedirect.com/science/article/abs/pii/S13522310230058733. Wang, Z., et al., A Spatiotemporal XGBoost Model for PM2.5 Concentration Prediction, Environmental

Modelling & Software, 2023.
https://pmc.ncbi.nlm.nih.gov/articles/PMC10696222/

4. Mohammadi, F., et al., Prediction of Atmospheric PM2.5 Levels Using Machine Learning Methods, Sustainability, 2024.
https://www.mdpi.com/2071-1050/16/2/911

5.Dai, H., et al., PM2.5 Concentration Prediction Based on Spatiotemporal Deep Learning Models, Sustainability, 2021.
https://www.mdpi.com/2071-1050/13/21/12071

6. Li, Y. Peng, "Air Pollution Forecasting Using LSTM De ep Learning Models," Journal of Atmospheric Environment, 2021.

7.Chen, T., and Guestrin, C., XGBoost: A Scalable Tree Boosting System, ACM SIGKDD, 2016.
https://dl.acm.org/doi/10.1145/2939672.2939785

8.Plantower PMS5003 Sensor Datasheet, Plantower Co., Ltd., 2019.

9.World Health Organization (WHO), Air Quality Guidelines, 2021.
https://www.who.int/publications/i/item/WHO-HEP-ECH-EHD-21.01

10.Hochreiter, S., and Schmidhuber, J., "Long Short-Term Memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.