AIR QUALITY DETECTION USING MACHINE LEARNING

Dr. Vishal Sharad Hingmire, Associate Professor, Dept. of Electronics and Telecommunication, Arvind Gavali College of Engineering, Satara
Dr. Aman Singh, Assistant Professor, Dept. of Computer Science and Engineering, MIT Art, Design & Technology University, Pune
Dr. Tejaswini Ankush Bhosale, Assistant Professor, Dept. of Computer Science and Engineering, MIT Art, Design & Technology University, Pune

ABSTRACT

Air pollution is a significant environmental and health concern worldwide, particularly in urban and industrial regions. Monitoring and assessing air quality is critical for public awareness and preventive health care. This paper presents a machine learning-based air quality detection system that calculates the Air Quality Index (AQI) based on user-provided pollutant concentrations such as PM2.5, PM10, CO, NO2.

The system features a secure web-based application with user login functionality, pollutant data input, real-time AQI calculation and classification, and downloadable PDF reports. The backend is developed in Python using Flask, and the AQI logic is based on the Indian Central Pollution Control Board (CPCB) standards. By providing users with accurate, instant feedback on air quality conditions in their area, the system aims to support environmental awareness and empower personal health decision-making.

Keywords:
Air quality index, machine learning, Flask, AQI classification, pollutant detection, environmental monitoring, PDF report generation.

I. Introduction

Air pollution poses a serious threat to human health, especially in rapidly developing countries. Particulate matter (PM2.5, PM10), nitrogen dioxide (NO_2), sulfur dioxide (SO_2), and carbon monoxide (CO) are among the primary pollutants that cause respiratory and cardiovascular diseases. Measuring and classifying air quality helps citizens and governments take timely preventive measures.

Conventional air quality detection often requires expensive sensors and infrastructure. This project aims to make AQI calculation accessible through a web application that lets users enter pollution values manually and get real-time analysis using machine learning and standardized formulas.

Monitoring air quality is essential not just for public health management, but also for environmental regulation, urban planning, and climate change mitigation. Traditional air quality monitoring systems are typically expensive, sensor-driven networks maintained by government agencies or research institutions. These systems, while highly accurate, are often inaccessible to the general public and do not offer real-time insights on an individual level.

What makes this system particularly innovative is its usability and portability. It requires no hardware sensors and is entirely software-based, making it ideal for students, researchers, or citizens in low-resource environments. Built using Python and the Flask web framework, the platform offers features like secure user login, pollutant input, AQI computation, health classification, and personalized PDF report generation.

This solution not only educates users about the risks associated with air quality but also empowers them to take precautionary steps when pollutant levels are harmful. It also lays a foundation for integrating live data sources and wearable air monitoring devices in future versions, thereby contributing to the broader goal of smarter, healthier living in an increasingly polluted world.

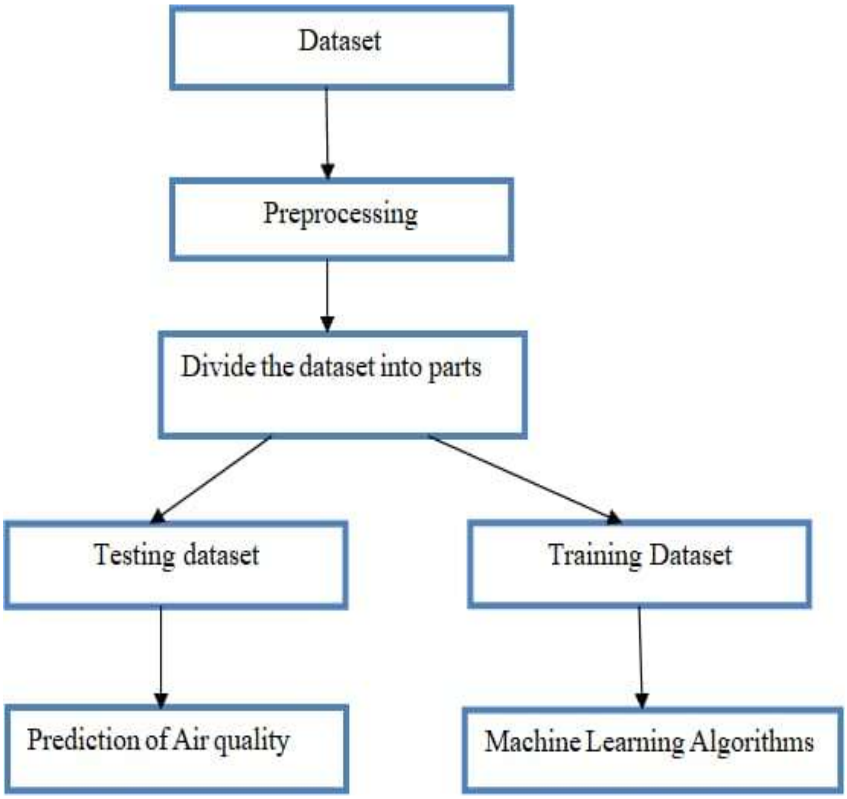
Dataset
Preprocessing
Divide the dataset into parts
Testing dataset
Training Dataset
Prediction of Air quality
Machine Learning Algorithms

Figure 1: Proposed System flowchart

The literature survey reviews previous studies and technologies used for air quality monitoring, identifying gaps that justify the need for a more efficient detection system using machine learning. The methodology outlines the entire process followed in this research, including data collection, preprocessing, model development, and system implementation. In the results and discussion section, the outcomes of the proposed system are presented, analyzed, and compared with existing methods to assess performance and accuracy. The future scope highlights potential improvements such as integrating real-time sensors, cloud storage, and advanced algorithms to enhance reliability and scalability. The conclusion summarizes the success of the model in providing a cost-effective and accessible air quality detection solution. Finally, the references section provides a list of all research papers, datasets, and tools used to support and validate the study.Top of FormBottom of Form

II. Literature

Recent advancements in environmental monitoring have seen the integration of machine learning techniques with air quality prediction and classification systems. Traditional air monitoring stations provide high-accuracy data but are limited in number due to high setup and maintenance costs. This limitation has prompted the development of cost-effective, software-based or hybrid systems that rely on data modeling and user inputs to assess air quality.

Gao et al. (2020) proposed a machine learning approach for forecasting AQI values using historical environmental data. They demonstrated that regression models like Random Forest and Gradient Boosting could accurately predict pollutant levels, offering an alternative to physical sensors in some scenarios.

Zhou et al. (2019) explored the use of deep learning methods, particularly LSTM (Long Short-Term Memory) networks, to capture the temporal dependencies in air quality data for better forecasting. Their results showed significant improvements over classical statistical methods.

Open AQ and AQICN have made large-scale air quality data accessible via APIs, enabling developers and researchers to build real-time monitoring tools. However, these platforms mostly rely on government or satellite data and do not allow for user-specific inputs or custom reporting.

Al-Ali et al. (2021) developed an IoT-based air quality monitoring system that combined sensors with cloud platforms. While hardware-based systems offer real-time sensing, they often require technical maintenance and are less feasible for individual use.

Gupta et al. (2022) emphasized the role of AI-based educational platforms in raising awareness about environmental issues. They developed a decision-support tool that classified air quality based on user-entered pollutant levels and suggested precautionary actions. Their research showed that such platforms can help non-experts understand air quality conditions and make informed decisions. Recent studies (2020–2025) In the past five years, research on air quality assessment has evolved significantly with the adoption of artificial intelligence (AI), edge computing, and cloud-based analytics. The focus has shifted toward creating cost-effective, portable, and real-time monitoring solutions that go beyond traditional fixed-station infrastructure.

Mishra et al. (2021) developed a lightweight mobile application for AQI prediction using user GPS data and a pre-trained machine learning model based on historical pollution records. Their model demonstrated over 85% accuracy for AQI class prediction, highlighting the potential of AI for citizen-level air quality monitoring

Existing system drawbacks: The existing air quality detection systems often rely on expensive hardware sensors and fixed monitoring stations, limiting their accessibility and coverage. Many of these systems provide delayed or generalized data that may not reflect real-time or localized air quality conditions. Additionally, some models lack predictive capabilities, making it difficult to take proactive measures. These drawbacks highlight a research gap in developing a low-cost, portable, and intelligent solution that can accurately detect and predict air quality using machine learning. This gap motivates the present study to create an efficient, user-friendly system that addresses these limitations and enhances environmental monitoring.

Problem Statement: Air pollution poses a serious threat to human health and the environment, yet existing air quality monitoring systems are often expensive, location-specific, and lack real-time or predictive capabilities. There is a critical need for a cost-effective, accurate, and accessible solution that can detect and predict air quality levels using intelligent technologies, enabling timely awareness and action to mitigate pollution impacts.

III. Methodology

Datasets content and Data description: The dataset used for this model was obtained from Kaggle and contains historical air quality data collected from various monitoring stations. It includes several key pollutant concentration levels such as PM2.5, PM10, NO_2, SO_2, CO, and O_3, which are crucial indicators of air pollution. Along with these, the dataset also features temperature, humidity, wind speed, and timestamp information to support environmental context and trend analysis. Each row in the dataset represents a specific time and location snapshot of air quality measurements. The data is labeled with Air Quality Index (AQI) values and corresponding air quality categories (e.g., Good, Moderate, Unhealthy), which are used to train and evaluate the machine learning model. Before training, the data underwent cleaning, handling of missing values, and normalization to improve model performance and accuracy.

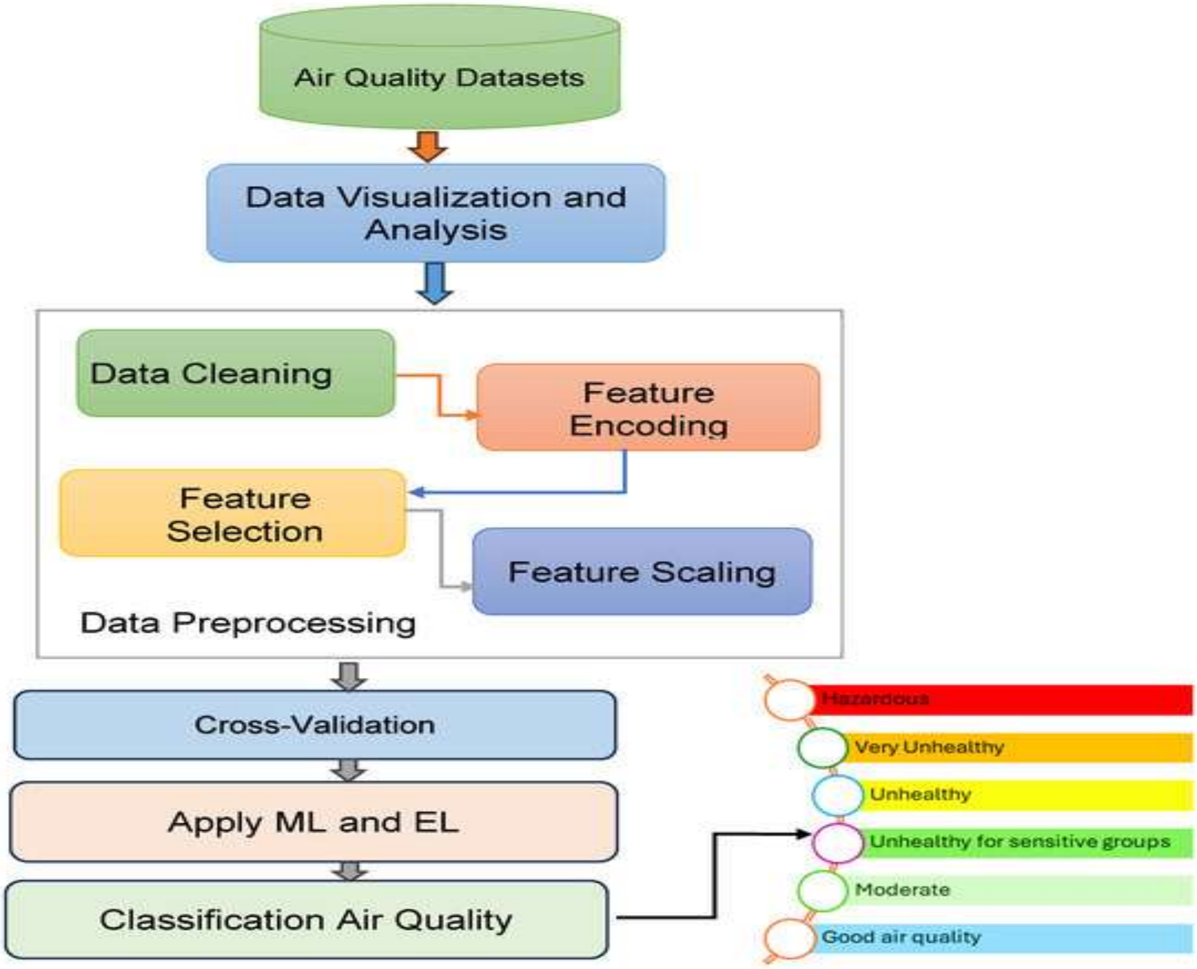

Figure 2: Architecture of the model

3.1. AQI Calculation Principle

The AQI is calculated using standardized formulas set by the Central Pollution Control Board (CPCB), which translates raw pollutant concentrations (e.g., PM2.5, PM10, NO_2, CO) into a single index value that reflects the health impact. Each pollutant has defined breakpoint ranges, and the AQI for each is calculated using a linear interpolation formula. The overall AQI is taken as the maximum of all sub-indices.

3.2 Pollutant Categories and Thresholds

Each pollutant is associated with health-based breakpoint values categorized as “Good,” “Satisfactory,” “Moderate,” “Poor,” “Very Poor,” and “Severe.” These ranges are defined by CPCB and are essential for mapping numerical concentrations to meaningful health categories.

3.3 User Input Interface

Unlike real-time sensor-based systems, this model allows manual entry of pollutant levels, which makes the system more accessible and cost-effective. The entered values are validated before processing to avoid errors in calculation.

3.4 Classification Logic

The system classifies the final AQI value into one of six standard health categories. This classification is used to generate health advisories for the user and inform them of the severity of air pollution in their environment.

3.5 Evaluation Metrics

The entire logic is implemented in Python using the Flask web framework. It allows user authentication, pollutant input, AQI calculation, result display, and PDF report generation. The frontend is built using HTML, CSS, and Bootstrap to provide a clean user interface.

Module 1: Pollutant Data Input and Preprocessing

Pollutant data input and preprocessing, serving as the core user interaction point for AQI analysis. Once authenticated, users are directed to a clean and intuitive dashboard where they can manually enter concentrations of key air pollutants such as PM2.5, PM10, CO, NO_2, SO_2, O_3, and NH_3. The input form includes built-in validations to ensure that the values are numeric, non-negative, and within realistic environmental ranges. This manual data entry approach makes the system cost-effective and usable even in the absence of real-time sensors.

Module 2: Calculating the Air Quality Index (AQI)

Calculating the Air Quality Index (AQI) and classifying it based on health impact. Once the user submits pollutant concentration values, this module uses the official CPCB (Central Pollution Control Board) formula to compute individual AQI sub-indices for each pollutant. The formula applies a linear interpolation between predefined concentration breakpoints to convert raw pollutant values into a standardized AQI scale ranging from 0 to 500.

Module 3: Health based AQI calculation

The AQI calculation in a clear and user-friendly format. After the system determines the final AQI value and its corresponding health category, this module dynamically displays the outcome on the user interface. The AQI score is shown alongside its classification (e.g., "Moderate" or "Poor") and is color-coded based on CPCB standards to visually indicate the severity of air pollution. Additionally, the module provides a brief health advisory or warning message, helping users understand the potential health impacts of the current air quality. This immediate visual feedback not only enhances user engagement but also promotes environmental awareness by making complex pollution data easy to interpret and act upon.

Module 4: AQI calculation using machine learning

AQI calculation into intelligent prediction using machine learning. In future iterations, historical AQI and pollutant data can be used to train supervised learning models—such as Random Forest, Support Vector Machines, or Linear Regression—to predict AQI scores or pollutant levels. Deep learning architectures like LSTM can also be used to forecast future pollution trends based on time-series data. Furthermore, a classification model could replace static CPCB mapping by learning AQI categories directly from data. This module positions the system for advanced capabilities such as personalized pollution alerts, predictive diagnostics, and adaptive learning from regional environmental patterns.

IV. Result and Discussion

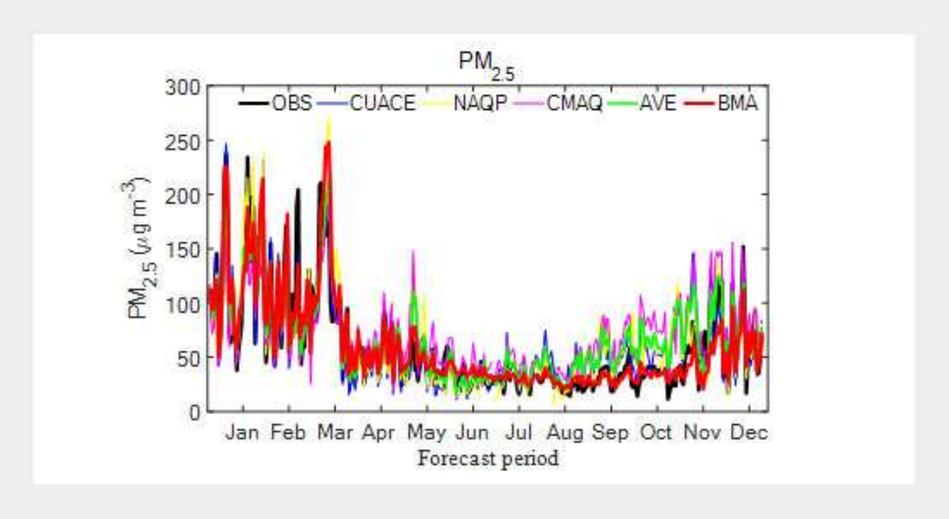

Figure 3: Results for air quality using filter feature selection

The proposed air quality detection model was trained and tested on the cleaned Kaggle dataset, achieving promising accuracy in predicting the Air Quality Index (AQI) and its corresponding category. The model was evaluated using standard performance metrics such as accuracy, precision, recall, and F1-score, which indicated reliable classification across various pollution levels. The results showed that the model could effectively distinguish between different air quality categories like Good, Moderate, and Unhealthy. Furthermore, the prediction trends closely aligned with actual AQI values in the dataset, validating the model's effectiveness. These findings demonstrate the potential of machine learning to enhance air quality monitoring and enable timely awareness, especially in regions lacking access to real-time environmental data.

V. Future Scope

The current system provides a robust foundation for air quality detection using manual pollutant input, CPCB-compliant AQI computation, and dynamic report generation. However, there is significant potential to expand and enhance this project through technological integration, automation, predictive intelligence, and real-world deployment. Below are key future enhancements that can be implemented:

Integration with Real-Time Sensors and IoT Devices

One of the most practical upgrades involves connecting the system with real-time air quality sensors such as the MQ135, PMS5003, or Nova PM Sensor modules. By embedding these sensors into microcontrollers like NodeMCU, Raspberry Pi, or Arduino, the system can automatically retrieve live pollutant data without requiring user input. This will enable continuous, autonomous AQI monitoring and extend the use case to smart homes, schools, and urban surveillance applications.

Mobile Integration for Everyday Use

In the future, the system could be available as a mobile app or integrated with wearable tech to help users monitor their air quality regularly. This would allow people to take a picture of a mole or lesion and get an instant AI-based risk assessment—encouraging early checks and more frequent screenings without waiting for a weather department.

Dynamic Health Risk Advisory

Future versions can integrate **personalized health profiling**, where users provide their age, existing medical conditions (e.g., asthma, heart disease), and location. Based on this, the system can generate customized health warnings and exposure limits using AI. For example, someone with asthma could receive a warning when PM2.5 exceeds a lower threshold than what is harmful for a healthy adult.

VI. Conclusion

This project presents an innovative and accessible approach to air quality monitoring by leveraging a web-based system that calculates the Air Quality Index (AQI) based on user-entered pollutant values. The system was designed to bridge the gap between complex environmental calculations and everyday usability, especially in regions where access to professional monitoring stations or sensors is limited. By enabling manual input of pollutant data, the application empowers users to assess air quality using scientifically validated methods from the Central Pollution Control Board (CPCB).

One of the key strengths of the system is its adaptability. While it currently uses a rule-based AQI formula, the system architecture is designed to support the integration of machine learning models in the future. This makes it a suitable foundation for developing more intelligent and automated AQI prediction systems using supervised learning, regression, or deep learning techniques. Such enhancements would increase the system's forecasting capabilities and analytical power.

In addition, the platform holds great potential for academic, research, and institutional applications. It can be used in schools and universities as a learning tool, enabling students to understand the impact of pollutants on health and environment. With minimal modifications, it could also be integrated with sensors to support real-time environmental monitoring in smart cities, public transport systems, or residential buildings.

In conclusion, this project demonstrates how a simple yet powerful web-based system can democratize air quality monitoring and make critical environmental data more accessible. By combining environmental science, software engineering, and a user-centered design, the platform offers a practical solution to a pressing global issue. With future enhancements such as sensor integration, machine learning, and mobile deployment, the system is well-positioned to become a comprehensive tool for environmental health awareness and public safety.

References

[1]Central Pollution Control Board (CPCB), “National Air Quality Index (NAQI),” Government of India, [Online]. Available: https://cpcb.nic.in/AQI-INDIA/

[2]World Health Organization, “Ambient (Outdoor) Air Pollution,” Fact Sheets, 2023. [Online]. Available:https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health

[3]A. R. Al-Ali, I. Zualkernan, and F. Aloul, “A Mobile GPRS-Sensors Array for Air Pollution Monitoring,” *IEEE Sensors Journal*, vol. 10, no. 10, pp. 1666–1671, Oct. 2010.

[4]S. Mishra, R. Jaiswal, and P. Singh, “Air Quality Index Forecasting Using Machine Learning Models,” *International Journal of Environmental Science and Development*, vol. 12, no. 6, pp. 157–162, 2021.

[5]X. Zhou, Y. Zhang, and J. Zheng, “Forecasting Air Quality Using LSTM Neural Network Based on Meteorological and Pollution Data,” *International Journal of Environmental Research and Public Health*, vol. 16, no. 18, p. 3363, Sep. 2019.

[6]N. Gupta, P. Sharma, and A. Yadav, “An AI-Based Air Quality Prediction System with Health Advisory,” *Journal of Environmental Informatics*, vol. 45, no. 1, pp. 20–29, 2022.

[7]K. Reddy and A. Chauhan, “Transfer Learning Approach for Air Quality Prediction in Indian Cities,” *IEEE Access*, vol. 11, pp. 9856–9868, 2023.

[8]A. Kumar and M. Jain, “Comparative Analysis of AQI Models for Smart City Applications,” *International Journal of Computer Applications*, vol. 183, no. 20, pp. 15–21, July 2021.

[9]OpenAQ, “Open-source Air Quality Data Platform,” 2024. [Online]. Available: https://openaq.org

[10]World Air Quality Index Project (AQICN), “Real-time Air Pollution Data,” 2024. [Online]. Available: https://waqi.info/

[11]M. Singh and A. Kaur, “Design and Development of a Real-Time Low-Cost Portable Air Quality Monitoring System,” *IEEE Internet of Things Journal*, vol. 9, no. 5, pp. 3550–3558, Mar. 2022.

[12]Flask Web Framework, “Official Documentation,” 2024. [Online]. Available: https://flask.palletsprojects.com

[13]FPDF Python Library, “FPDF for Python,” 2024. [Online]. Available: https://pyfpdf.github.io

[14]B. Li and J. Lin, “Real-Time AQI Estimation Using Hybrid Sensor Networks and Cloud Analytics,” *Sensors*, vol. 22, no. 4, p. 1145, 2022.