Machine Learning-Based Real-Time Air Pollution Monitoring and Forecasting Using IoT and Environmental Sensor Data: A Survey

Sarita Bharat Kaphare
Dept. of Artificial Intelligence & Data Science (M.E),
Jaihind College of Engineering
Savitribai Phule Pune University
Kuran, Maharashtra, India.
sarita.kaphare07@gmail.com

Prof. Sachin Bhosale
Dept. of Computer Engineering
Jaihind College of Engineering
Savitribai Phule Pune University
Kuran, Maharashtra, India.
ssachinbhosale@gmail.com

Dr. Anand Khatri
Dept. of Computer Engineering
Jaihind College of Engineering
Savitribai Phule Pune University
Kuran, Maharashtra, India.
khatrianand@gmail.com

Dr. Shubhangi Gunjal
Dept. of Computer Engineering
Jaihind College of Engineering
Savitribai Phule Pune University
Kuran, Maharashtra, India.
shubhangi.gunjal83@gmail.com

Abstract:

Air pollution remains a serious environmental and public health problem, driven mainly by vehicle emissions, industrial activity and fast urban growth. Most monitoring setups report only the pollution level at the present moment and say little about what the air will be like later. This paper surveys recent work on predicting air quality with machine learning and deep learning, and then describes a proposed system that joins an Arduino-based sensing unit with historical air pollution data. The hardware unit uses particulate matter, environmental and gas sensors to record $PM_{2.5}$, PM10, temperature, humidity and selected gases, while a Java web-based application presents the readings and the forecasts to the user. Historical data from Kaggle supports model training and testing. After cleaning, missing-value handling, normalization and feature selection, Random Forest, Decision Tree, Support Vector Machine and Artificial Neural Network models are to be compared for predicting the Air Quality Index (AQI) and pollutant concentrations. The system is intended to offer real-time monitoring, forecasting and early warnings of poor air quality.

Keywords: Air pollution prediction, Air Quality Index (AQI), Arduino, environmental monitoring, Internet of Things (IoT), machine learning, predictive analytics, regression.

1. Introduction

Poor air quality is linked to breathing disorders, climate change and wider damage to the environment. Pollutants such as $PM_{2.5}$, NO_2, SO_2, CO and O_3 affect both people and ecosystems. Present monitoring arrangements generally depend on sensors and laboratory analysis, and they mainly report the pollution level as it is now.

Machine learning offers a way to move from reporting to forecasting. By learning from environmental records, including temperature, humidity, wind speed and pollutant levels, models can estimate future pollution and the AQI. Such estimates help organizations plan ahead for air quality and the health problems that follow from it.

Low-cost IoT sensors make it practical to gather live readings, and machine learning makes it possible to learn from past records. Combining the two is the central idea of the system considered in this paper. An Arduino microcontroller connected to PM, environmental and gas sensors supplies real-time readings, historical datasets from Kaggle supply training material, and a Java web-based application gives the user access to the results.

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 states the problem and objectives. Section 4 describes the proposed system, and Section 5 concludes the paper.

2. Literature Survey

Rosa-Bilbao et al. [1] built an indoor air quality management system for smart educational settings. A wireless sensor network picks up temperature, humidity, CO_2, TVOC and $PM_{2.5}$, and a low-code platform ties these devices to complex event processing, machine learning and deep learning modules, a blockchain and a NoSQL database. For prediction, an LSTM forecasts temperature, CO_2 and $PM_{2.5}$, whereas a Decision Tree is used to study how temperature, humidity, CO_2 and $PM_{2.5}$ are related to each other. The system was installed in university lecture rooms and laboratories, where it handled live monitoring, prediction and detection of complex events, and kept a secure record of important events on the blockchain.

Zhang et al. [2] wrote a broad survey of deep learning methods for air quality prediction. They went through how models are designed, what data goes into them and how performance is judged across many earlier studies, and then described new trends and problems that are still open. In their view, models meant for environmental use need to become easier to interpret, more dependable and better suited to real-time use.

Dong et al. [3] put Empirical Mode Decomposition, transformer models and BiLSTM networks together to predict air quality over the short term. The combination picked up the nonlinear patterns and time dependence in pollution data, and their experiments showed clearly better accuracy than any single model on its own. Because it gives quick forecasts, it suits cities where conditions change fast.

Houdou et al. [4] reviewed interpretable machine learning for forecasting air pollution. For them, a model that can be explained matters, particularly when the results have to be shared with the public or used in policy making. They compared a range of ML and DL methods and pointed out the ones that manage to be accurate and understandable at the same time. They concluded that explainable models can help guide environmental decisions and raise public awareness.

Kumar and Pande [5] studied pollution prediction for a number of large Indian cities. Past pollution records were combined with weather data to sharpen the forecasts. The models predicted pollution trends reliably, which makes it possible to take preventive steps in time, and the authors saw good scope for such methods in city management and policy making.

Zhu et al. [6] built a deep learning framework aimed at air and water pollution together in crowded urban areas. Drawing on several environmental datasets, it could locate pollution hotspots and predict how contamination would develop. It gave better accuracy than many traditional analytical models and, according to the authors, helps with sustainable city planning and better use of resources.

Maltare and Vahora [7] forecast the AQI of Ahmedabad using several machine learning algorithms. They tried both regression and classification approaches to see which worked well for local AQI prediction, and some techniques proved more accurate for regional assessment. Municipal authorities that plan to manage air quality ahead of time can take practical guidance from this work.

Gupta et al. [8] compared machine learning models for AQI prediction, testing random forests, support vector machines and neural networks. Ensemble and hybrid methods came out ahead consistently. The authors also noted that the model architecture should be chosen to match the pollution behaviour of the region, so that predictions stay reliable.

Liang et al. [9] proposed Airformer, a transformer-based model for forecasting air quality across China. It makes use of the spatial and temporal links between environmental factors and gave highly accurate predictions. On large-scale data in particular, it beat conventional ML and DL approaches, so it can support broad environmental monitoring and national policy.

Bekkar et al. [10] presented a deep learning method for predicting air pollution in smart cities, working with data from real-time IoT sensors. Their model forecast pollution levels precisely and was meant to help urban planners and health officials in their decisions. In the comparison they carried out, deep learning did better than conventional machine learning, especially on complex real-world data.

3. Problem Statement and Objectives

Air pollution keeps increasing with urbanization, industrial activity and vehicle emissions. Existing monitoring systems mostly give current air quality information and have limited ability to forecast future levels. The work considered here is to build an Arduino-based real-time monitoring and forecasting system that uses PM, environmental and gas sensors, combines its readings with historical Kaggle data, and applies machine learning to predict AQI and pollutant concentrations with early warnings for poor air quality.

Objectives:

- To build a machine learning framework that forecasts the AQI from historical air pollution and weather data, together with an Arduino-based real-time monitoring system using PM, environmental and gas sensors.
- To collect real-time parameters such as $PM_{2.5}$, PM10, temperature, humidity, and gas concentrations.
- To integrate real-time sensor data and historical air pollution datasets from Kaggle.
- To preprocess, analyze, and select relevant features from the collected data.
- To develop and compare Machine Learning models such as Random Forest, Decision Tree, SVM, and ANN for AQI/pollutant prediction.
- To provide real-time forecasting and early warnings for poor air quality.

4. Methodology

The proposed system has a hardware part and a Java web-based part. The hardware collects live readings, the data-processing part prepares them and applies the machine learning models, and the web-based application presents the monitoring results and forecasts.

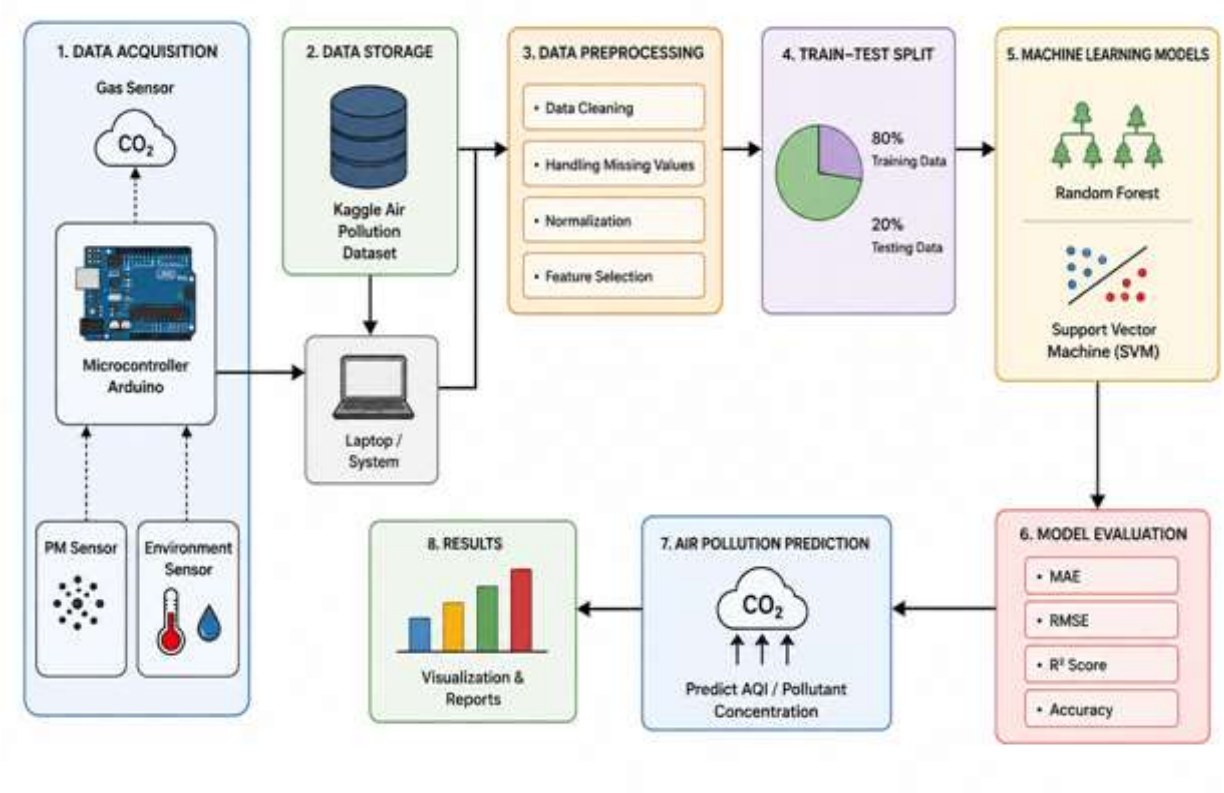


Fig: - System Architecture

4.1 Hardware Unit

An Arduino microcontroller reads values periodically from the connected PM sensors, environmental sensors and gas sensors, and produces time-stamped observations. Table 1 lists a typical record.

Table 1. Typical Data Record

Parameter	Source
$PM_{2.5}$	PM sensor
PM10	PM sensor
CO / selected gases	Gas sensor
Temperature	Environmental sensor
Humidity	Environmental sensor
Timestamp	Arduino / system
AQI	Calculated / predicted

4.2 Dataset

Two sources are used. The first is the real-time data gathered by the Arduino-based system. The second is the Kaggle air pollution dataset, which holds several pollutant and weather parameters such as $PM_{2.5}$, temperature, humidity and CO, and is used for training and testing the models.

4.3 Data Preprocessing

Preprocessing has four steps in order: data cleaning to remove noise, handling of missing values, normalization, and feature selection to identify the important attributes.

4.4 Train-Test Split

The dataset is divided into a training set containing 80% of the data and a testing set containing the remaining 20%.

4.5 Machine Learning Models

Random Forest, Decision Tree, Support Vector Machine and Artificial Neural Network models are compared. The prediction target is the AQI or the concentration of individual pollutants.

4.6 Model Evaluation

Performance is measured with Mean Absolute Error (MAE), Root Mean Square Error (RMSE), the R2 score and accuracy.

4.7 Real-Time Monitoring

The Arduino continuously acquires sensor readings and transmits them to the data-processing system. Through the Java web-based application, the system provides real-time PM readings, environmental readings and gas sensor readings, along with visualization of both historical and real-time data. Results are shown as graphs and charts, and final reports and insights are drawn from them.

The work is expected to yield the following:

- A functional Arduino-based real-time air pollution monitoring prototype.
- Collected real-time environmental and air pollution data.
- A comparison of Random Forest, Decision Tree, SVM and ANN models.
- Identification of a suitable machine learning model based on prediction performance.
- A real-time visualization and monitoring mechanism.
- A scalable framework suitable for smart-city and environmental monitoring applications.
- A demonstration of IoT, environmental sensors and machine learning working together for air pollution forecasting.

5. Conclusion

The surveyed studies show that machine learning and deep learning can forecast air quality and the AQI with useful accuracy, and that IoT sensors can supply the live data such models need. They also show that ensemble and hybrid methods often perform well, and that interpretability and real-time usability remain important. Building on these findings, this paper described an Arduino-based monitoring unit combined with historical data, a comparison of Random Forest, Decision Tree, SVM and ANN models, and a Java web-based application for real-time display and early warnings. The system aims to move air quality monitoring from reporting present conditions toward forecasting future ones.

References

[1] J. Rosa-Bilbao, F. S. Butt, D. Merkl, M. F. Wagner, J. Schäfer, and J. Boubeta-Puig, “IoT-based indoor air quality management system for intelligent education environments,” IEEE, 2025.

[2] Z. Zhang et al., “A systematic survey of air quality prediction based on deep learning,” Alexandria Eng. J., vol. 93, 2024.

[3] J. Dong, Y. Zhang, and J. Hu, “Short-term air quality prediction based on EMD-transformer-BiLSTM,” Sci. Rep., vol. 14, no. 1, 2024.

[4] A. Houdou et al., “Interpretable machine learning approaches for forecasting and predicting air pollution: A systematic review,” Aerosol Air Qual. Res., vol. 24, no. 1, 2024.

[5] K. Kumar and B. P. Pande, “Air pollution prediction with machine learning: A case study of Indian cities,” Int. J. Environ. Sci. Technol., vol. 20, no. 5, 2023.

[6] L. Zhu et al., “Deep learning method for minimizing water pollution and air pollution in urban environment,” Urban Climate, vol. 49, 2023.

[7] N. N. Maltare and S. Vahora, “Air quality index prediction using machine learning for Ahmedabad city,” Digit. Chem. Eng., vol. 7, 2023.

[8] N. S. Gupta et al., “Prediction of air quality index using machine learning techniques: A comparative analysis,” J. Environ. Public Health, vol. 2023, no. 1, 2023.

[9] Y. Liang et al., “Airformer: Predicting nationwide air quality in China with transformers,” in Proc. AAAI Conf. Artif. Intell., vol. 37, no. 12, 2023.

[10] A. Bekkar et al., “Air-pollution prediction in smart city, deep learning approach,” J. Big Data, vol. 8, no. 1, 2021.