

## Automated Intelligent System for Air Quality Detection

**Nor Hazlyna binti Harun<sup>1,2</sup>, Mohamad Loqmanul Hakim bin Hisyam<sup>1</sup>,  
Muhammad Daniel Haikal bin Wazi<sup>1</sup>, Muhammad Hafizuddin bin Sha'ari<sup>1</sup>,  
Muhammad Amirul Amsyar bin Halid<sup>1</sup>, Farith Shazry bin Rosli<sup>1</sup>**

*1 Data Science Research Lab, School of Computing,  
University Utara Malaysia, 06010 Sintok, Kedah, Malaysia*

*2 Institute for Advanced and Smart Digital Opportunities,  
School of Computing, University Utara Malaysia, 06010 Sintok, Kedah, Malaysia*

\*Corresponding Author: loqmanh1410@gmail.com  
DOI: <https://doi.org/10.30880/emailt.2023.00.00.000>

### Article Info

Received: Day Month Year  
Accepted: Day Month Year  
Available online: Day Month Year

### Keywords

Random Forest, Machine Learning,  
Air Quality Detection, Intelligent  
System

### Abstract

Air pollution is the term used to describe any chemical, physical, or biological factor that tampers with the natural properties of the atmosphere to contaminate the indoor or outdoor environment[8]. Air pollution is in critical need of both intelligence and automation control in order to maintain a good air quality. Air quality prediction plays an important role in addressing environmental challenges. Traditional methods often rely on complex models and manual data analysis, leading to delays in obtaining accurate predictions. In this study, the research focused on addressing the pressing need for intelligent and automated control in combating air pollution and maintaining optimal air quality levels. The primary objective was to enhance air quality prediction through the utilization of artificial intelligence, specifically employing the Random Forest algorithm. The study involved the development of an intelligent system designed to predict air quality by analyzing various contributing factors. The Random Forest algorithm was chosen for its ability to handle complex datasets and provide robust predictions. The system included data on contaminants, acknowledging multiple elements that impact air quality. Using this method, an intelligent system capable of assessing a variety of parameters impacting air quality is created. This model demonstrated a high prediction accuracy of over 99.91%, showcasing its potential to improve air quality assessments. This research contributes to the advancement of efficient and timely air quality monitoring, providing valuable insights for environmental management and public well-being.

## 1. Introduction

Nowadays, the study of Artificial Intelligent (AI) applications in environmental domain is crucial for innovation that can increase productivity, reduce environmental risk and increase citizens safety. Over the last decade, air quality has been one of the critical global problems with major implications for public health, socio economics and even for ecosystems. Air pollution in smart cities in the world has been

drastically increasing and the increase in the concentration of particulate matter (PM<sub>2.5</sub>) in the air is a threat for the country and citizens health [6]. Many respiratory diseases were reported in most urban cities and premature deaths of children were reported, citing the reasons for consistent exposure to toxic air [1]. Toxic air and bad air quality are often adversely affected by surges in the emission of pollutants and harmful gasses from the industry into the atmosphere and worsening the air quality level, creating a demand for innovative approaches to address this problem. Hence early detection of the air quality level should be prioritized.

In most of the country nowadays, fixed monitoring stations and manual data collection have been the mainstays of air quality monitoring. These techniques have various drawbacks even though they have given insightful information about air quality levels. The first is that real-time monitoring and quick response approach are faced with new challenges due to the quickly changing landscape of pollution sources, which includes growing industrialization and increased vehicle traffic. Second, even though they are accurate, it only provides localized data, which may not fully capture the spatial variability of pollution in dynamic urban environments. Third, these methods might not be able to quickly adjust to shifting pollution patterns or new pollutants. Hence, due to world population expands and air quality becoming worse, an efficient and accurate system is needed to make the prediction of air quality better. Therefore, this system's existence can help the experts make an early detection for air pollution and this approach may directly increase the efficiency for detecting the air pollution especially in industrial areas.

This study aims to design and construct an AI-based system for air quality detection that can improve the prediction of the air quality especially in high population and industrialization areas. One of the most significant aspects of this air quality detection system is detecting the abnormal value of the air substances. Some problems like inaccurate values of ozone and carbon dioxide may seriously affect the classification stages, therefore it is crucial to implement measures to enhance the accuracy of data collection methods. Finally, this system can precisely classify air quality level and improve air pollution detection. This could lead to more precise, timely, and through insights into air quality issues and in the end, a healthier and more sustainable future for urban populations across the globe.

## 2. Related Work

A study has been proposed of **Smart City Air Quality Prediction using Machine Learning** where the research addresses the prediction of concentration of PM<sub>2.5</sub>, in the Air Pollutant Index (API) of Kuala Lumpur and Johor Bharu in order to predict the concentration of PM<sub>2.5</sub> for each smart city[6]. The research emphasizes 2 machine learnings which is Multi-Layer Perceptron Neural Networks (MLP) and Random Forest (RF). In the end of the study, the author justify that Random Forest has outperformed MLP as it provides a better interpretation of many decision trees and provides a better result than MLP required more data to train its neuron and predict it. Thus, the study resulted in Random Forest has 97% accuracy while MLP resulted in with 92% accuracy.

Recently, in India, there was also a research focusing on application of random forest to estimate the air quality[1]. All both 4 regression techniques(Machine learning algorithms) which includes Linear Regression (LR), Decision Tree (DT), Support Vector Regression (SVR) and Random Forest (RF) were implemented to evaluate the accuracy of Mean Absolute Error(MAE), Mean Squared Error(MSE) and Root Mean Squared Error(RMSE). The result ended up with Random Forest being the most consistent growth in all the stages, which be able maintain a high accuracy score of 80% because of its ensemble learning technique.

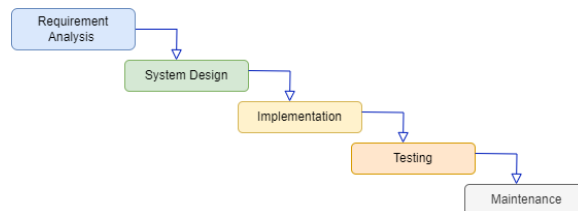
Another research also has been discussed in India to evaluate the best machine learning algorithm to predict the Air Quality Index (AQI) [4]. The research relies on these tools to perform evaluation for the ML algorithms which includes Accuracy, Precision, Recall, and F1 score. Amongst all the six Machine Learning algorithms involved (Logistic Regression, Naive Bayes, Random Forest, Support Vector Machine, K-Nearest Neighbors, and Decision Tree), Random Forest and Decision Tree classifiers ended up performed the best with both accuracy of 99% followed by Naive Bayes classifier being the worst of the six algorithms, with a classification accuracy of 84%. However, Random Forest is still better as it is proved to be more stability as compared to the Decision Tree and compared them to the actual class.

Prediction of Air Pollutants Using Supervised Machine Learning was another research done also in India where the work focused on components such as , PM10, PM2.5, SO2, CO, and NO2 to investigate the air quality index[5]. There were 6 machine learning programs mentioned by author during this paper which is logistic regression, decision tree, support vector machine, random forest tree, Nave Bayes theorem, and K-nearest neighbor. Amongst all 6 machine learning algorithms, decision tree is proven to be the most effective methodology as it reached 99.88% followed by Random Forest of 99.16%.

Based on the reviewed works with the same background of study, there are many machine learning algorithms implemented by the authors to reach the same goals, which is to attain the best machine learning algorithm to predict the most accurate and efficient Air Quality Index (AQI). All works mentioned resulted in Random Forest mostly being the best ML algorithm. Based on the findings, the project will utilize Random Forest to be used to predict air quality. Random Forest being the backbone of the project was the best decision as it is proven to provide the most accurate result of the research. Moreover, Random Forest has been fully utilized to its full potential in this project. Compared to other works mentioned, the result provided in the project has shown a higher accuracy which is about 99.91%. This proves that the project succeeded in achieving the best efficiency in predicting Air Quality Index (AQI) using the Random Forest algorithm.

### 3. Methodology

The study for this project was conducted based on the Waterfall Software Development Approach. This approach is chosen to ensure a clear understanding for all group members involved in this project. Figure 1 shows the five phases for this methodology that were conducted during the system development process. The initial step for the requirement analysis phase is to collect data. The second step for the system design phase includes data preprocessing. The third step involves training and testing data, as well as system interface development for the implementation phase. The complete system is then tested during the fourth step in the testing phase. Finally, the maintenance phase is the last step for this approach.



**Figure 1 : Waterfall Flowchart**

#### i) Data Collection

The Global Air Pollution dataset was obtained from the Kaggle website [3] to complete the first phase of this approach. A total of 23,464 data points were obtained from every country globally, consisting of 12 columns for attributes, including meta, feature, and target data. For this project, the main focus is on the feature data, which includes Carbon Monoxide, Ozone, Nitrogen Dioxide, and Particulate Matter in Air Quality Index values and categories. All these features are required to succeed in predicting the targeted data for implementing this system.

#### ii) Data Preprocessing

The dataset provided by the Kaggle website was cited from the trusted company called Elichens [2]. Thus, the dataset was clean, and no noisy data was found. Since, apart from the data being alphanumeric, all the alphabet data needed to be encoded into integers to enable it to run in the classifier models. The entire AQI category for all features data, such as "Good," "Satisfactory," "Moderately Polluted," "Poor," "Very Poor," and "Severe" was encoded into integers. The irrelevant data for attributes such as country and city were dropped in this phase to make the prediction accuracy for the system more precise. All the data were structured into numeric format because it will run in supervised learning classifiers, which have targeted data which is the Air Quality Index Category for this system.

### iii) Algorithm Machine Learning

The Random Forest Classifier machine learning algorithm was employed for accurate prediction of Air Quality Index (AQI) categories. Random Forest is a powerful and versatile supervised machine learning algorithm that involves growing and combining multiple decision trees to create a forest. The code imports the "RandomForestClassifier" from the sklearn.ensemble library to apply the algorithm for training and testing data, as well as predicting data for the targeted categories. The data for this experiment is split into a 70-30 ratio for training and testing, respectively. The classifiers were trained and tested using labeled data. Once the classifier has learned patterns and relationships within the data by training on 70% of the dataset, it is evaluated using the remaining 30% as the test set. The predictions made on the test set then compared with the actual categories in the test set to assess and evaluate the model's performance. The outcome of the model is evaluated using confusion matrix's classification report.

### iv) Design and Development

Following the phases of the Waterfall model, this section explains the design and development of the Air-Quality Prediction System (AQPS). The AQPS is developed as a standalone system, and software prototyping is employed as a common method for displaying software requirements. In the design phase, four input boxes were placed near the feature labels, allowing users to enter values for each feature. The system also provides a range for values under the feature labels to guide users when entering data. The development of the AQPS system was carried out using the Visual Studio Code software. Figure 2 displays the interface of the system.



Figure 2 : Air-Quality Prediction System Interface

## 4. Result and Evaluation

The confusion matrix, shown in Table 2, would serve as the foundation and evaluation for assessing the system's performance. There are thirty-six elements in this matrix. Orange and white colour indicate incorrect predictions, whereas blue indicates accurate predictions. Equation 1 formula was used to determine the suggested model's accuracy based on the confusion matrix.

$$\text{Accuracy(\%)} = \frac{\sum \text{blue}}{\sum \text{blue} + \sum \text{orange} + \sum \text{white}} \times 100$$

Equation 1. Accuracy formula

Six categories related to air quality were examined in total. The total instances, category, and accuracy of the instances tested are displayed in the table below. In screening the air quality, the percentage of the classification model is defined by the category and accuracy listed below.

Total instances	Category	Detection accuracy (%)
2981	Good	100.00
1130	Satisfactory	99.91
86	Moderately Polluted	100.00
2785	Poor	100.00
39	Very Poor	94.87
18	Severe	83.33

Table 1. Air Quality Testing Results

		Prediction					
		Good	Satisfactory	Moderately Polluted	Poor	Very Poor	Severe
Actual	Good	2981	0	0	0	0	0
	Satisfactory	0	1129	0	1	0	0
	Moderately Polluted	0	0	86	0	0	0
	Poor	0	0	0	2785	0	0
	Very Poor	0	0	0	0	37	2
	Severe	0	0	2	0	1	15

**Table 2. Confusion Matrix of Testing Result**

As can be seen from Table 1 above, three of the six categories have reached 100%, the ideal accuracy, as determined by Equation 1, and only one category has the lowest accuracy of 83.33%. to demonstrate that it could serve as a useful early screening tool, confusion matrix accuracy has been measured to assess the system's effectiveness. Accordingly, Table 2 above shows that 7033 of the 7039 instances categories had accurate detection. In the meantime, one satisfactory instance, two very poor instances, and three severe instances were mistakenly identified. Equation 1 states that 99.91% accuracy has been attained in the air quality category screening process. The accuracy findings have demonstrated how well this effort has worked to identify the various categories of air quality based on what has been recorded.

Figure 3 show the example of result when input of four pollutants AQI value which are carbon monoxide, ozone, nitrogen dioxide and particulate matter 2.5 were entered into the system. The input data were retrieved from the World's Air Pollution website of Hongkou Liangcheng, Shanghai city [7]. The result of the system match with the result on the website.

**Figure 3 : Air-Quality Prediction System Modul**

## 5. Conclusion

This paper presents the performance Random Forest algorithm in predicting the air quality based on four pollutants AQI value. In conclusion, the results achieved by the system to predict air quality are believed to help various parties in identifying the level of air pollution in their respective areas. This system can help classify lots of different things that affect air quality, and it is better than the usual methods that are used, which are fixed monitoring stations and manual data collection. This AI system that implements Random Forest algorithm will help authorities determine and classify the air quality level faster and more accurately. The system has been designed to be adaptable, allowing for enhancements through the incorporation of a more extensive and up-to-date dataset, including new types of pollutants. These improvements could make the system work better to predict the problems with air quality. The idea of bringing in artificial intelligence to keep an eye on air quality is a game-changer. It's not just about making air quality checks faster and more on time, but it's also about making sure people's growth is sustainable and making sure everyone stays healthy. As countries around the world balance between getting things done and taking care of the environment, the use of AI in air quality checks becomes an important part of building a better and healthier future for everyone.

## References

- [1] Application of Random Forests for Air quality estimation in India by adopting terrain features | IEEE Conference Publication | IEEE Xplore. (n.d.). Ieeexplore.ieee.org. Retrieved January 16, 2024, from <https://ieeexplore.ieee.org/document/9315252>
- [2] eLichens - Air Quality Solutions - smart gas sensors and stations. (2024, January 2). ELichens. Retrieved 16 January 2024, from <https://www.elichens.com/>
- [3] Muzdadid, H. A. (2022, November 8). Global Air Pollution Dataset. Kaggle. Retrieved 16 January 2024, from <https://www.kaggle.com/datasets/hasibalmuzdadid/global-air-pollution-dataset/data>
- [4] Pant, A., Sharma, S., & Pant, K. (2023). Evaluation of Machine Learning Algorithms for Air Quality Index (AQI) Prediction. Journal of Reliability and Statistical Studies, 229–242. Retrieved January 16, 2024, from <https://doi.org/10.13052/jrss0974-8024.1621>
- [5] Sci-Hub | Prediction of Air Pollutants Using Supervised Machine Learning. 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS) | 10.1109/iciccs51141.2021.9432078. (n.d.). Sci-Hub.se. Retrieved January 19, 2024, from <https://sci-hub.se/https://ieeexplore.ieee.org/document/9432078>
- [6] Smart City Air Quality Prediction using Machine Learning | IEEE Conference Publication | IEEE Xplore. (n.d.). Ieeexplore.ieee.org. Retrieved January 16, 2024, from <https://ieeexplore.ieee.org/document/9432074>
- [7] The. (2024). Hongkou Liangcheng, Shanghai Air Pollution: Real-time Air Quality Index. Retrieved January 16, 2024, from aqicn.org website: <https://aqicn.org/city/shanghai/hongkouliangcheng/>
- [8] World. (2019, July 30). Air pollution. Retrieved January 16, 2024, from Who.int website: [https://www.who.int/health-topics/air-pollution#tab=tab\\_1](https://www.who.int/health-topics/air-pollution#tab=tab_1)