Machine Learning Approach for Predicting Air Quality Index

K.M.O.V.K. Kekulanadara
Department of Computing and Information Systems, Sabaragamuwa University of Sri Lanka
Belihuloya, Sri Lanka
oshadhikekulandara96@gmail.com

B.T.G.S Kumara
Department of Computing and Information Systems, Sabaragamuwa University of Sri Lanka
Belihuloya, Sri Lanka
btgsk2000@gmail.com

Banujan Kuhaneswaran
Department of Computing and Information Systems, Sabaragamuwa University of Sri Lanka
Belihuloya, Sri Lanka
bhakuha@appsc.sab.ac.lk

2021 International Conference on Decision Aid Sciences and Application (DASA) | 978-1-6654-1634-4/21/$31.00 ©2021 IEEE | DOI: 10.1109/DASA53625.2021.9682221

***Abstract*— Air pollution has become one of the most detrimental environmental issues in the modern world. Urbanization, industrialization, and development are the foremost factors to escalate air pollution. Polluted air can create negative impacts on human health and environmental well-being. Therefore, many countries around the world are interested in assessing air quality in their living areas. Air Quality Index (AQI) values are used as metrics to evaluate daily air quality. Various machine learning algorithms are now widely used to research forecasting, predicting, and classification tasks. This paper addresses the comparative analysis of various Machine Learning Algorithms like Decision Tree, Random Forest, Support Vector Machine (SVM), and Artificial Neural Network (ANN) for predicting AQI using major pollutants: NO, NO_2, CO, SO_2, O_3, NH_3, NO_X, $PM_{2.5}$, PM_{10}, Benzene, Toluene, and Xylene. The results prove that machine-learning algorithms can be utilized appropriately to predict the AQI.**

Keywords— ***Air Quality Index (AQI), Air Pollution Predicting, Machine Learning Algorithms, Decision Tree, SVM, ANN***

I. Introduction

Air pollution is the presence of harmful substances such as gases, chemicals, dust, and other particles above the level in the atmosphere. These particles can create adverse environmental issues and cause negative repercussions on human health. Long and short-term exposure to air pollution may have effects on different organs and systems [1] like the respiratory system, cardiovascular system, nervous system, urinary system, etc. According to the American Lung Association, air pollution-related diseases cost the United States about 37 billion dollars per year [2]. Furthermore, Air pollution contributes to a high mortality rate. Nearly 4.2 million people have died from exposure to harmful air composition, according to an article in the State of Global Air 2017 from the Institute of Health Metrics and Assessment [3]. Air pollution is badly affected not only the health of living organisms but also the environment. Some of the adverse environmental effects that have arisen due to polluted air include global warming, climate change, ozone depletion, and acid rain.

Air pollution has increased dramatically in recent years due to industrialization, rapid development, and overpopulation in urban areas. World Health Organization reports state that 92% of the world's population resides in regions where the poor air quality and the air quality is less than WHO acceptable levels [4]. Hence, it seems essential to formulate new regulations and policies to control air pollution levels. Many countries have focused on monitoring and measuring the air quality around their living areas. The Air Quality Index (AQI) can be used as a global indicator to measure daily air quality. In 1976, the U.S. Environmental Protection Agency (EPA) introduced a daily air pollution index. It was known as Pollution Standards Index (PSI) before being renamed as AQI [5]. The most important aspect of AQI value is to provide correct data about regional air quality, how polluted air may impact people, and how can safeguard human health. [6]. The AQI reviews pollutants including O_3, CO, SO_2, $PM_{2.5}$, PM_{10}, etc. Air quality is divided into six categories according to AQI value. These six categories are good, satisfactory, moderate, poor, very poor, and severe. AQI values, corresponding air quality levels, and health concerns are shown in Table I.

TABLE I. AQI Categories and Level of Health Concern

AQI Value	Air Quality Level	Level of Health Concern
0 to 50	Good	Good
51 to 100	Satisfactory	Satisfactory
101 to 150	Moderate	Unhealthy for sensitive people
151 to 200	Poor	Unhealthy
201 to 300	Very Poor	Very Unhealthy
301 to 500	Severe	Hazardous

AQI can measure the air quality in a defined area at a given time. The results may help to reveal air quality trends, identify highly polluted areas, make national regulations, policies and take steps to control high pollutant activities.

In recent times, machine learning techniques have been used in many environmental science-related pieces of research as well as for predicting purposes. Machine learning is a type of Artificial Intelligence (AI) and a field that is constantly evolving. The focus of the work presented in this paper is predicting the AQI using various machine learning algorithms, namely, Decision Tree, Random Forest, Support Vector Machine (SVM), and Artificial Neural Network (ANN). Finally, proposing a more accurate machine learning algorithm for predicting AQI.

The rest of the paper is arranged as follows. Section II: An overview of the relevant works, Section III: Description of the systematic approach for the implementation, Section IV: Step-by-step discussion of the results obtained through implementation, Section V: Description of future works and conclusion of this paper.

II. Literature Review

Air pollution is a vital environmental issue that has attracted the attention of many researchers around the world. Since the 1960s, there has been a greater focus on predicting and analyzing air quality in Western developed countries [7]. Various methods and approaches were used to research air pollution. The most commonly used approaches to predict air quality were deterministic and statistical methods [8, 9]. Both of these methods need more data to make more accurate predictions [10]. The deterministic method was able to estimate pollutant emissions using meteorological principles and chemical models. This approach employed mathematical equations to create numeric models for transferring and diffusion processes of air concentration and predicting air quality. Community Multi-Scale Air Quality (CMAQ) model, Nested Air Quality Prediction Modeling System (NAQPMS) are some examples of widely applied numerical models. However, the output results of these models were not very accurate due to some factors. These factors include more detailed source data requirements, incomplete theoretical basis, needfulness of various parameters for calculations, incorrect and suspicious emission data, confusing surface conditions, etc. Over time, statistical methods got more attention as compared with deterministic methods. Because statistical methods were not based on more theoretical hypotheses as well as they were used past data and experiences for future air quality predictions. Multiple Linear Regression (MLR) model, Auto Regression Moving Average (ARMA) model were some instances for commonly used statistical models. However, statistical models were unable to present non-linear patterns. Therefore, these models also could not generate more accurate outputs.

Machine learning approaches have become more and more famous due to technological advancements, especially in the field of forecasting. They were not like pure statistical models. Because machine learning algorithms have a strong ability to deal with several parameters, analyze diverse data from various sources and solve complex nonlinear problems. Hence, the prediction results were more accurate than previous linear models as the most of air quality data relies on non-linear patterns than linear ones [8] Machine learning approaches [2, 9] include the ANN method, Support Vector Regression (SVM) method, Genetic Programming (GP), hybrid methods, etc.

The ANN has become one of the most widely used methods for predicting air quality [11], and due to its structure, it can perform excellently. To predict air quality, various ANN models have been applied. Neuro-fuzzy neural network, Multi-Layer Perception (MLP), BP neural network (BPNN), and Recurrent neural network are a few of them (RNN).

The following are some of the studies related to this area. In [12], a hybrid ANN model was proposed based on wavelet transform, Multi-Layer Perceptron (MLP) neural network, Weather Research and Forecasting (WRF) meteorological model, and Monte Carlo simulation. In this study, daily PM2.5 concentrations were predicted using the ANN model and the above-proposed hybrid model. The comparison results prove that the hybrid ANN model was more accurate than pure ANN models. Djebbri and Rouainia [13] developed a Nonlinear Auto-Regressive (NARX) model based on ANN to monitor NOX and CO pollutant concentration in industrial sites. The results showed that neural networks can be used for predicting pollutants concentrations efficiently.

In [10], a hybrid model was created to simulate and predict hourly pollution concentration in different stations in Tehran. Meteorological parameters like wind speed, wind direction, pressure, temperature, relative humidity are also considered to build this model. The wavelet transformation and neuro-fuzzy network with fuzzy clustering algorithms were employed to build it. In [14], authors have introduced a prediction model using ANFIS (Adaptive Neuro-Fuzzy Inference System) techniques to predict particulate matter concentration.

In [15], the authors conducted a study using logistic regression and autoregression algorithms to detect and predict the $PM_{2.5}$ level in a specific city. In the first step, the $PM_{2.5}$ level was detected based on considered factors such as temperature, wind speed, dew point, pressure, etc. In this phase, a logistic regression algorithm was applied to classify whether the air sample was polluted or not. In the second step, using an autoregression model to predict the $PM_{2.5}$ level for a particular date. The proposed system was able to detect air quality and predict $PM_{2.5}$ concentrations in the future.

In [16], the SVM model has been introduced for Air Pollution Index (API) predictions. Only the kernel functions model parameters were considered to conduct this study. Reddy, Yedavalli, Mohanty, and Nakhat [17] found an LSTM framework to forecast air pollution in Beijing based on time series data and meteorological data. This framework supported extending the air pollution prediction time step to 5 - 10 hours in the future.

III. Methodology

The specific objective of this research is to predict the AQI by considering major pollutants in the air. Decision Tree, Random Forest, SVM, and ANN algorithms were used to create prediction models.

The experiments were carried out using a Windows 10 installed computer with an Intel Core i7-7500U, 2.70 CPU, and 8GB RAM. Weka software, Anaconda Navigator GUI, Jupyter Notebook IDE, and python were used for implementations.

Fig 1 depicts the framework of the suggested solution. The proposed method consists of a five-step process. They are 1) data collection, 2) data preprocessing, 3) feature selection, 4) training models 5) evaluation. Each step of the process is outlined in detail below.

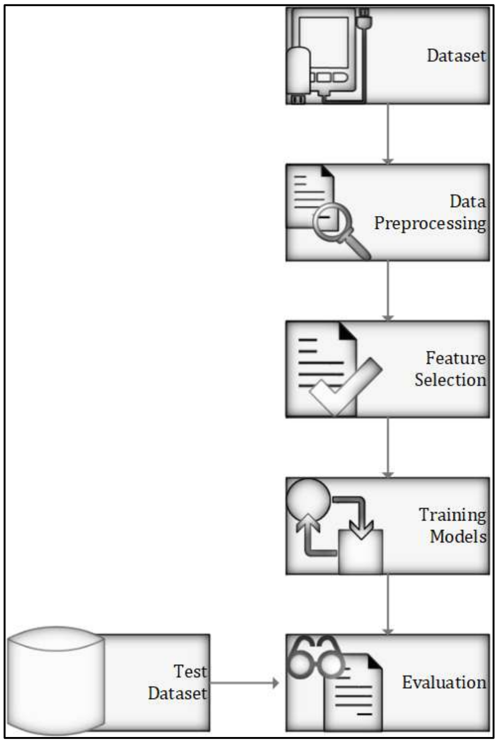

Fig. 1. The framework of the AQI prediction task

A. Data Collection

The dataset that is used in this research was taken from Kaggle online dataset library. Data is collected from 15 January 2015 to 1 July 2020. It is contained hourly and daily levels of AQI and air quality at different stations across various cities in India. We conducted this study by using hourly AQI data. In this dataset, there were 16 attributes with 100,000 rows of instances. The dataset contains both internal and external attributes to this study. These features include station id, date, time, the concentration of $PM_{2.5}$, PM_{10}, NO, NO_2, NO_x, NH_3, SO_2, O_3, CO, Benzene, Toluene, Xylene, and AQI. The dataset was used to train the prediction model. Particulate Matter (PM) [18] is also known as particle pollution. There are two major features, including $PM_{2.5}$ and PM_{10}. These particles also cause serious health problems. Therefore, many studies have been conducted to predict the PM levels in the air [19]. In the dataset, PM values and other pollutant data were measured in units of micrograms per cubic meter ($\mu g/m^3$) except for NO_x, which was measured in ppb, and CO, which was measured in mg/m^3. The following Table II displays a snapshot of the dataset.

TABLE II. AIR POLLUTION DATASET

Parameter	**Station Id**	
	AP005	**DL019**
DateTime	4/5/2020 7:00:00 PM	6/29/2018 7:00:00 PM
$PM_{2.5}$	20.25	26.25
PM_{10}	43.5	41
NO	11.75	2.37
NO_2	34	49.35
NO_x	27.6	28.15
NH_3	7.85	41.4
CO	0.11	085
SO_2	7.28	9.37
O_3	7.72	3.2
Benzene	1.07	3.48
Toluene	2.67	16.17
Xylene	0.33	2.95
AQI	50	62
AQI_Bucket	Good	Satisfactory

B. Data Preprocessing

After obtaining the dataset, it was necessary to identify and eliminate unnecessary attributes and incomplete data from it. Because some of the pollutant data from specific periods were not included in the dataset under consideration. Therefore, data preprocessing is required to enhance data quality, obtain more accurate information and make the right predictions. So then Missing values, tuples, and noisy values of input instances were all excluded from the dataset. The dataset was comprised of 61,285 instances after the preprocessing step.

C. Feature Selection

Feature selection is a method of recognizing key attributes and features from the observed dataset. These selected features display only related information for producing the expected outcome. Feature selection is a core step in degrading overfitting and advancing accuracy. The features that came up with the most prediction accuracy were selected through feature selection. Unnecessary or somewhat relevant features might have a negative impact on prediction performance. In this study, thirteen of the sixteen features were selected. They were $PM_{2.5}$, PM_{10}, NO, NO_2, NO_x, NH_3, CO, SO_2, O_3, Benzene, Toluene, Xylene, and AQI_Bucket. The excluded attributes involved Station_Id, Date_Time, and AQI values. The AQI_Bucket was set as the target variable.

D. Training models

Here, different machine learning algorithms such as Decision Tree, Random Forest, SVM, and ANN algorithms were used to train the prediction models. As the first part of the implementation, Weka software was used to execute the Decision Tree, Random Forest, and SVM algorithms. Weka is open-source software and it can implement a large number of algorithms. The model accuracy is dependent on the percentage split value.

The rest of the implementation was carried out using Python programming, Keras deep learning library, and Softmax activation which is commonly used in deep learning. Anaconda distribution was installed to Windows to conduct python implementations. After installing the Anaconda distribution, the Python environment and Jupyter Notebook were set up to be ready for implementation. All of the required modules like Keras, pandas, and sklearn were needed to install before starting the implementation. Dataset was loaded using pandas and split into input and output variables like x and y. The output variable was encoded because it consists of six different string classes: good, satisfactory, moderate, poor, very poor, and severe. The different neural network architectures were applied by changing the number of hidden layers, batch size, and epochs in each model to select the comparatively accurate model. After model creation, the created model was evaluated using the K-Fold Cross Validation method. The value of K was substituted by 10 for evaluating the model. Because it produces a more accurate output than using a lower value for K.

IV. RESULTS AND DISCUSSION

As mentioned in Section III-B and III-C, data preprocessing and feature selection phases were aided to clean the dataset by avoiding missing data and redundant features. Machine learning algorithms were able to interpret data features very well after preprocessing the dataset. The following Table III shows the dataset summary after preprocessing it.

TABLE III. SUMMARY DETAILS OF THE DATASET

	Before	**After**
# of Instance	100,000	61,285
Dataset Features	Station_ID, DateTime, $PM_{2.5}$, PM_{10}, NO, NO_2, NO_x, NH_3, SO_2, O_3, CO, Benzene, Toluene, Xylene, AQI, AQI_Bucket	$PM_{2.5}$, PM_{10}, NO, NO_2, NO_x, NH_3, SO_2, O_3, CO, Benzene, Toluene, Xylene, AQI_Bucket
Number of Features	16	13

After the implementation, the results show that 80% data splitting percentage generates the highest accuracy among all algorithms. Therefore, 80% of instances are better for training the model and the rest of 20% data is used for testing. AQI_Bucket was selected as the target variable for executing these implementations. Comparison results are shown in Table IV, Table V, Table VI, and Table VII for Decision Tree, SVM, Random Forest, and ANN respectively.

Accuracy, Loss rate, and Root Mean Square Error (RMSE) were used to compare the results obtained for various percentage splits.

TABLE IV. SUMMARY OF DECISION TREE ANALYSIS

Percentage split (%)	70	75	80	85	90
Accuracy	63.323	63.749	64.943	64.125	63.708
Lost rate	36.677	36.251	35.058	35.875	36.292
RMSE	0.319	0.317	0.314	0.316	0.318

TABLE V. SUMMARY OF SVM ANALYSIS

Percentage split (%)	70	75	80	85	90
Accuracy	61.414	61.419	61.345	60.622	60.311
Lost rate	38.586	38.581	38.656	39.378	39.687
RMSE	0.328	0.328	0.328	0.328	0.328

TABLE VI. SUMMARY OF RANDOM FOREST ANALYSIS

Percentage split (%)	70	75	80	85	90
Accuracy	73.310	73.520	74.039	73.338	74.037
Lost rate	26.690	26.480	25.961	26.662	25.963
RMSE	0.250	0.249	0.248	0.249	0.247

Various neural network models were defined and evaluated using the K-fold cross-validation method to find out the most accurate model. Each model had 12 input variables including $PM_{2.5}$, PM_{10}, NO, NO_2, NO_x, NH_3, SO_2, O_3, CO, Benzene, Toluene, and Xylene. And AQI_Bucket was defined as the target variable. The number of hidden layers, neurons in the hidden layers, epochs, and batch size were changed and compared the model accuracy. Table VII shows the summary results of neural implementation.

TABLE VII. SUMMARY OF NEURAL IMPLEMENTATION

# of layers	# of hidden layers	Nodes in the hidden layer	Epochs	Batch size	Accuracy (%)	RMSE
3	1	7	100	5	63.63	0.48
3	1	7	1000	200	64.57	0.99
3	1	100	1000	400	69.32	0.63
3	1	300	1000	500	70.82	0.48
4	2	9,7	1000	500	65.25	0.77
4	2	24,10	1000	500	66.99	0.60
4	2	100,50	1000	500	69.11	0.54
4	2	120,70	1000	500	69.31	0.37
4	2	240,100	1000	500	71.2	0.59
5	3	24,15,10	400	100	66.89	0.32
5	3	24,15,10	4000	1000	67.41	0.62
5	3	100,150,200	400	100	69.42	0.79
5	3	250,150,100	400	10	71.02	0.5
5	3	250,150,100	500	300	71.90	0.75
6	4	260,250,150,100	500	300	72.44	0.55
6	4	270,250,150,100	500	300	72.96	0.71
6	4	280,250,150,100	500	300	72.35	0.38
6	4	290,250,150,100	500	300	72.50	0.40
6	4	300,250,150,100	500	300	72.84	0.61
6	4	400,300,200,100	500	300	73.16	0.49
6	4	400,300,200,100	1000	10000	73.52	0.46
6	4	400,350,250,150	1000	500	73.78	0.53
6	4	400,350,250,150	1000	1000	73.72	0.55
6	4	450,400,350,250	1000	1000	73.85	0.57

According to Table VIII, the most accurate classification was done by a Random Forest classification algorithm. It performed a maximum 74.039% of accuracy value. ANN also performed a relatively approximate accuracy value. But it had a comparatively high RMSE value. It is better to have as much as a lower value to RMSE for proper predictions. And the neural implementation takes too much time compared with others. Because the dataset was very large and needed to define more models by altering their architecture to choose a more accurate model. The Decision Tree algorithm integrates some decisions and the Random Forest integrates various Decision Trees to get output. When the dataset is getting larger, Random Forest algorithms generate more complex output and take much time. This study concludes that machine learning techniques can be used properly to predict the AQI because all of these selected algorithms perform a low-level RMSE value. And also Random Forest Algorithm was the best because it generated the highest accuracy and the lowest RMSE value.

TABLE VIII. COMPARISON OF ACCURACY LEVEL AMONG CONCERNED ALGORITHMS

	Decision Tree	SVM	Random Forest	Neural Network
Accuracy	64.943	61.345	74.039	73.82
RMSE	0.314	0.328	0.248	0.57

V. CONCLUSION

Air pollution has rapidly emerged as one of the decisive environmental issues that are gaining global concern. People must be aware of the level of air pollution around them. Because polluted air may cause many health issues. AQI is used as a global metric to assess air quality and how it affects human health. That's why predicting AQI is so important.

As a result of technological advancement, machine learning has become more famous for prediction tasks in different fields. In this study, we used four machine learning algorithms. The findings of this study show that machine learning algorithms can be used to more effectively detect air quality. And also the results proved that Random Forest and neural network algorithms are the best for AQI predictions.

The results of this research will be significant to people, government, environmental and meteorological departments as well. Because they will help to take actions to reduce adverse air pollution, formulate new rules and urge individuals to manage their everyday activities without creating hazardous air pollution.

REFERENCES

[1] Kampa, M., and Castanas, E.: 'Human health effects of air pollution', Environmental pollution, 2008, 151, (2), pp. 362-367

[2] Castelli, M., Clemente, F.M., Popovič, A., Silva, S., and Vanneschi, L.: 'A machine learning approach to predict air quality in California', Complexity, 2020, 2020

[3] Macatangay, L.H., and Hernandez, R.M.: 'A Deep Learning-Based Prediction and Simulator of Harmful Air Pollutants: A Case from the Philippines', 'Book A Deep Learning-Based Prediction and Simulator of Harmful Air Pollutants: A Case from the Philippines' (IEEE, 2020, edn.), pp. 381-386

[4] Chaudhary, V., Deshbhratar, A., Kumar, V., and Paul, D.: 'Time series based LSTM model to predict air pollutant's concentration for prominent cities in India', UDM, Aug, 2018

[5] Moscoso-López, J.A., Urda, D., González-Enrique, J., Ruiz-Aguilar, J.J., and Turias, I.J.: 'Hourly air quality index (AQI) forecasting using machine learning methods', 'Book Hourly air quality index (AQI) forecasting using machine learning methods' (Springer, 2020, edn.), pp. 123-132

[6] EPA, A.Q.I.: 'A guide to air quality and your health', USA: EPA, 2009

[7] Chen, S., Kan, G., Li, J., Liang, K., and Hong, Y.: 'Investigating China's Urban Air Quality Using Big Data, Information Theory, and

Machine Learning', Polish Journal of Environmental Studies, 2018, 27, (2)

[8] Li, X., Peng, L., Hu, Y., Shao, J., and Chi, T.: 'Deep learning architecture for air quality predictions', Environmental Science and Pollution Research, 2016, 23, (22), pp. 22408-22417

[9] Wen, C., Liu, S., Yao, X., Peng, L., Li, X., Hu, Y., and Chi, T.: 'A novel spatiotemporal convolutional long short-term neural network for air pollution prediction', Science of the total environment, 2019, 654, pp. 1091-1099

[10] Fotouhi, S., Shirali-Shahreza, M.H., and Mohammadpour, A.: 'Concentration prediction of air pollutants in tehran', 'Book Concentration prediction of air pollutants in tehran' (2018, edn.), pp. 1-7

[11] Srivastava, C., Singh, S., and Singh, A.P.: 'Estimation of air pollution in Delhi using machine learning techniques', 'Book Estimation of air pollution in Delhi using machine learning techniques' (IEEE, 2018, edn.), pp. 304-309

[12] Sadabadi, Y., Salari, M., and Esmaili, R.: 'A hybrid model for online prediction of PM2: 5 concentration: A case study', Scientia Iranica, 2021, 28, (3), pp. 1699-1710

[13] Djebbri, N., and Rouainia, M.: 'Artificial neural networks based air pollution monitoring in industrial sites', 'Book Artificial neural networks based air pollution monitoring in industrial sites' (IEEE, 2017, edn.), pp. 1-5

[14] Mihalache, S.F., Popescu, M., and Oprea, M.: 'Particulate matter prediction using ANFIS modelling techniques', 'Book Particulate matter prediction using ANFIS modelling techniques' (IEEE, 2015, edn.), pp. 895-900

[15] Aditya, C., Deshmukh, C.R., Nayana, D., and Vidyavastu, P.G.: 'Detection and prediction of air pollution using machine learning models', 'Book Detection and prediction of air pollution using machine learning models' (2018, edn.), pp. 204-207

[16] Leong, W., Kelani, R., and Ahmad, Z.: 'Prediction of air pollution index (API) using support vector machine (SVM)', Journal of Environmental Chemical Engineering, 2020, 8, (3), pp. 103208

[17] Reddy, V., Yedavalli, P., Mohanty, S., and Nakhat, U.: 'Deep air: forecasting air pollution in Beijing, China', Environmental Science, 2018

[18] Harrison, R.M., and Yin, J.: 'Particulate matter in the atmosphere: which particle properties are important for its effects on health?', Science of the total environment, 2000, 249, (1-3), pp. 85-101

[19] Wu, D., Lary, D.J., Zewdie, G.K., and Liu, X.: 'Using machine learning to understand the temporal morphology of the PM 2.5 annual cycle in East Asia', Environmental monitoring and assessment, 2019, 191, (2), pp. 1-14