Application of Machine Learning Algorithms for Calculating Air Quality Index

Nebojša Bogdanović, Mladen Koprivica, *Member, IEEE*, Goran Marković, *Member, IEEE*

***Abstract*—Air pollution is an ever-growing issue, especially severe in urban and industrial areas. Air Quality Index (AQI) is a unit of measuring the level of air pollution, which takes into account the concentrations of all relevant air pollutants. There are two main problems that must be addressed in AQI calculations, i.e. regression and classification. The regression problem consists of calculating (approximating) the AQI index based on the concentrations of different air pollutants. In classification problem, the measurements of air pollutants' concentrations are classified into different Air Quality Classes. In this paper a number of Machine Learning (ML) and Deep Learning (DL) algorithms were designed and used in order to solve both the regression and classification problems for AQI. The main goal was to present performance comparison for wide set of ML and DL algorithms based on the values of Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Coefficient of Determination (R squared) in regression tasks, and Accuracy in classification tasks. Also, the percentage of algorithms' convergence and the time needed to perform these regression and classification tasks are also measured.**

***Index Terms*—Air Quality Index (AQI); Machine Learning; Deep Learning; Regression; Classification**

I. Introduction

Air pollution presents growing issue, which is especially severe in urban and industrial areas, and occurs whenever excessive quantities of pollutants such as gases, particulates, and bio-molecules are introduced into the atmosphere. It has harmful consequences on human population and other living organisms (i.e. it can cause diseases and/or even death, and impairing crops). Air pollutants can be solid particles, liquid droplets, or gases, and are classified as primary (i.e. directly emitted from the source) or secondary (i.e. formed in the atmosphere) pollutants. National environmental agencies set the standards and air quality guidelines regarding acceptable levels for air pollutants, while the air quality index (AQI) is used as an indicator in order to report the measuring of the air pollution and how unhealthy the air is (i.e. reports on possible associated health effects, above all for risk groups). AQI is calculated based on the maximum individual AQI measured for the observed criteria (air) pollutants, and this calculation is rather complex and thus its implementation is not suitable for applications with low-cost sensor platforms employed in the form of dense IoT-based sensor network. In fact the process of calculating AQI by formulas consists of two steps: (1) calculation of air quality index for every pollutant in each of the measurements separately, and (2) observing values of the all indexes for every measurement in order to find the maximum. On the other hand, in a case of application of machine learning (ML), a whole process of training and testing the algorithms takes longer than the use of formulas, but these algorithms only need to be trained once, before its application in real-time systems. Thus, when compared to formulas which needs to be used every time we have a different measurement, the time needed to perform this task by ML algorithms is shorter. It should be noticed, that formulas used for these calculations are not complex, but since the implementation of these formulas requires using multiple loops and case functions, the process takes longer when compared to the testing part of the ML and DL algorithms.

On the other hand, the more useful, flexible and scalable usage of AQI in terms of influence on the human population health, would be to deploy air quality forecasting system based on the measured levels of concentration of individual air pollutants, which would be able to predict AQI (i.e. air quality) locally and in short-term manner (hourly). This demands the use of dense network of low-cost sensors and thus requires simple solution for the determination of AQI based on the local low-quality air pollutant measurements.

So far, research community and environmental agencies have developed different methods for calculation of AQI, [1][2], but still no universally accepted method exist that is appropriate in all scenarios, [3]. The machine learning (ML) based methods are proposed as an obvious and natural solution for AQI determination and prediction, such as fuzzy lattices decision support system, [1], the support vector regression (SVR), [2], or different ML algorithms (linear regression, random forest, decision tree, SVR, and K-Nearest neighbor).

In this paper, the broad set of ML algorithms, including deep learning (DL), are observed as possible solutions for determination of AQI based on the measured levels of six criteria pollutants. Also, we here addressed two main issues in AQI calculation: regression problem that represents AQI calculation based on criteria pollutants concentrations, and classification problem in which the measurements of air pollutants' concentrations are classified into the Air Quality Classes. The output of the ML models is the approximation of the current values of AQI, while the prediction of the future values of AQI is something we are considering for the future works. In total, 8 different ML algorithms and 5 DL models were analyzed for the regression task, while 9 different ML algorithms and 3 DL models were observed for the classification task. We here observed much broader set of ML algorithms than in previous work, i.e. in [3]. ML algorithms and DL models were designed, optimized and tested based on dataset consisting of real-time measurements

Nebojša Bogdanović is a student at the School of Electrical Engineering, University of Belgrade, Bulevar kralja Aleksandra 73, 11020 Belgrade, Serbia (e-mail: bn150118d@student.etf.bg.ac.rs).

Mladen Koprivica is with the University of Belgrade, School of Electrical Engineering, Bulevar kralja Aleksandra 73, 11120 Belgrade, Serbia (e-mail: kopra@etf.bg.ac.rs).

Goran Marković is with the University of Belgrade, School of Electrical Engineering, Bulevar kralja Aleksandra 73, 11120 Belgrade, Serbia (e-mail: gmarkovic@etf.bg.ac.rs).

gathered from 5 countries. The performance metrics are defined, and performance analysis and comparison of the observed ML algorithms and DL models is performed for regression and classification tasks.

The paper is organized as follows. In the section II the basic concepts related to air quality pollutants, monitoring, and scale are given. Section III gives short description of the observed machine learning algorithms and the deep learning models observed in this paper, as well as a short description of AQI regression and classification tasks, while a dataset used in ML and DL algorithm training and performance analysis is described in section IV. The main results and conclusion are presented in section IV, followed by the final concluding remarks.

II. Air Pollutants and Air Quality Scale

Air pollution is most frequently man-made. It usually comes from factories, powerplants and heating plants which use unrenewable energy sources, cars and public transport.

According to International Energy Agency (IEA) [4], from the year 2018 to 2040 the projected energy demand should rise annually by 1.3%. This projected growth can be seen on Fig. 1, where in the year 2020 around 60% of all the energy should be generated using non-renewable energy sources. While the use of renewable energy sources is projected to increase by the year 2040, because of the growth in energy demand, the amount of energy generated by coal, gas, oil and nuclear energy will not decrease.

This is a growing problem, in both the developed and in countries in development, because a usage of fossil fuels results in high concentrations of air pollutants released into the atmosphere. Many of developed countries are fighting this problem by imposing laws which are restricting the amounts of fossil fuels burned each year. Also, powerplants and factories are required to use filters in order to reduce the emission of air pollution into the atmosphere.

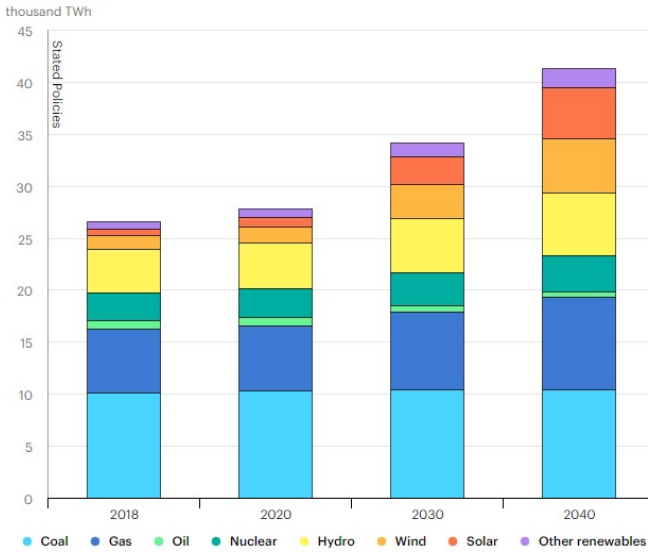

Fig. 1. Projected growth of energy demand from 2018 to 2040, [4]

The most common types of air pollution according to New South Wales Ministry of Health (NSW Health) [5], are listed below:

- Carbon Monoxide (CO), mostly generated by motor vehicles and industry plants;
- Ozone (O_3), the main component of smog, a product of interaction between sunlight and emissions from motor vehicles and industry plants;
- Particulate Matter (PM 2.5, PM 10), the small solid particles and liquid droplets suspended in air, made up of variety of components including nitrates, sulfates, organic chemicals, metals, soil or dust particles and allergens, which mostly comes from motor vehicles and industry plants;
- Nitrogen Dioxide (NO_2), generated by motor vehicles, industry plants, and unflued gas-heaters; and
- Sulphur Dioxide (SO_2), generated by the fossil fuel combustion at power plants and industrial facilities.

A. Air Quality Scale

Air Quality Index (AQI) can be calculated in a number of different ways, and depending on which formulas are used for calculations, there are different AQI scales. In this paper, the formulas and scales used are created by the Central Pollution Control Board, Ministry of Environment, Forests and Climate Change in India. This corresponding air quality scale is presented in Table I.

TABLE I
Air Quality Scale (India)

Category	AQI Index	Possible Health Impacts
Good	0-50	Minimal health impacts
Satisfactory	51-100	Minor breathing discomfort to sensitive people
Moderate	101-200	Breathing discomfort to the people with lung, asthma and heart diseases
Poor	201-300	Breathing discomfort to most people on prolonged exposure
Very Poor	301-400	Respiratory illness on prolonged exposure
Severe	401-	Affects healthy people and seriously impacts those with existing diseases

III. Algorithms and Problems

As defined in the introduction, there are two types of challenges in calculating AQI index which are addressed in this paper, the regression and the classification tasks. Both of these issues hold valuable information when calculating levels of air pollution. Some of the reasons for using ML algorithms in this area are:

- Provision of real-time decision support for air quality sensors, especially in a case of wide usage of low-cost sensors (i.e. for IoT-based environmental monitoring networks). Specifically, a verification that sensors for monitoring concentrations of various pollutants are working well, to predict the missing values in a case of sensor malfunction, and to evaluate inputs and decide whether and alarm should be triggered or not [1];
- Improvement of sensor performance for lower-cost air quality monitoring [6]; and
- Forecasting (prediction) of future values of pollution concentrations and AQI index [2] [3].

In this paper, we have performed comparison of the wide set of various machine learning algorithms for regression and classification tasks, such as: Multiple Linear Regression (MLR), Stochastic Gradient Descent (SGD) Classifier based on Linear Regression, Support Vector Machine (SVM), K-

Nearest Neighbors (KNN), Random Forest (RF), Decision Tree, Extra Trees Regression, Adaptive Boosting based on Decision Trees (AdaBoost), and Gradient Tree Boosting (GradBoost). Also, we designed and estimated performance for several Deep Learning algorithms in both tasks

Regression and classification tasks are rather similar, and thus the classification task can be realized by classifying the results achieved by regression algorithms into the respective categories in Table I. In this case, we would achieve 100% accuracy for classification task for the all algorithms, except for Multiple Linear Regression, but a time needed to execute would be even higher than for the regression algorithms. This is why in this paper we proposed a different method. In our method, the input data used for classification algorithms is the same as for the regression algorithms, and that is just the concentrations of the pollutants of measurements, while the labels used for training of the ML algorithms and DL models are final categories in Table I. By using this method we expected slightly lower classification accuracy (which will be discussed in section VI), when compared to the first method (based on regression), but the time needed to execute such algorithms would be, depending on an algorithm, from two to ten times smaller than for their respective regression algorithms.

The performance metrics for AQI regression algorithms were the values of Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Coefficient of Determination (R^2), while for the classification tasks we used the algorithm accuracy as the main performance metric. As the additional performance metrics for both tasks, we measured percentage of convergence and the elapsed time needed to perform these tasks for all observed algorithms.

IV. Air Pollutant Measurement Dataset

The dataset used in the analysis was created from data gathered from websites data.world [7], and openaq.org [8], and it consists of 35440 independent measurements from 5 countries (Serbia, India, USA, Australia and Taiwan). The measurements data were gathered from 2016 to April 2021. Each measurement consists of measured concentrations of Carbon-monoxide (CO), Ozone (O_3), particulate matter PM 2.5 and PM 10, Nitrogen-dioxide (NO_2) and Sulphur-dioxide (SO_2). The concentrations of all of the pollutants are measured in $\mu g/m^3$, except for CO, which is measured in mg/m^3. The mean values (mean) and standard deviations (std) of the measurements in dataset, are given in Table II.

TABLE II
Dataset description

	Count	Mean	Std
CO	35440	1.39088	1.34341
O_3	35440	51.9751	61.5998
PM 2.5	35440	76.7299	112.159
PM 10	35440	152.993	185.914
NO_2	35440	51.9291	56.3290
SO_2	35440	8.51487	8.71780
AQI	35440	202.835	195.820

Based on these concentrations, the reference AQI index and the air quality class were calculated for each of the measurements, by using formulas implemented in Python scripts. The dataset was divided into training and test sets in the ratio 90%:10%, and the training set was further divided into training and validation sets in the same ratio.

V. Results of Performance Analysis

The performance analysis of observed machine learning algorithms is performed by using Scikit-learn library for the Python programming language, while the deep learning algorithms were implemented using Tensorflow and Keras libraries for Python. The implementations were executed on Google Colaboratory cloud computing platform, by using Intel(R) Xeon(R) CPU @ 2.30 GHz processing unit with 16 GB of available RAM memory.

Both the machine learning and deep learning algorithms were trained and tested independently 42 times, with the values for *random_state* parameter ranging from 0 to 41, in order to guarantee different train/test splits of the dataset for the each iteration of the observed algorithm.

A. Regression algorithms

The analysis showed that all of the regression algorithms have the convergence rate of 90.47% (38/42). In 4 executions where the algorithms diverge, the corresponding MAE and RMSE values were not taken into account in the calculation of the mean values and standard deviations of these errors.

The five Deep Learning models, marked DL#1 to DL#5, were designed, optimized and used. These neural networks models are defined as: DL#1 model with 3 hidden layers comprising of with 128 neurons in each layer, DL#2 model with 3 hidden layers with 256 neurons in each layer, DL#3 model with 3 hidden layers with 512 neurons in each layer, DL#4 model with 3 hidden layers with 128 neurons in the first hidden layer, 1024 neurons in the second hidden layer, and 128 neurons in the third hidden layer (DL#4), and DL#5 model with 3 hidden layers with 256 neurons in the first hidden layer, 1024 neurons in the second hidden layer, and 256 neurons in the third hidden layer. Data is normalized in input layer of each DL model. The activation function for all layers was a ReLU function, while the loss function used was Mean Squared Error (MSE). The Adam optimization function was used with the learning rate of 0.001, and every neural network model is trained over 100 epochs. Different numbers of epochs for training DL models were considered during the design of these models, for both regression and classification tasks. In this process, it is observed that even if for some of the lower numbers of epochs the algorithms performed similarly as for 100 epochs, the results were not consistent enough, e.g. the convergence rate was lower (i.e. for 80 epochs the algorithms converged in 28/42 cases). Thus, we choose the number of 100 epochs, since the further rise in the number of epochs did not give better results.

The mean values and standard deviations of MAE and RMSE for all of the algorithms used in AQI regression task are shown in Table III, while in Table IV the times needed to train (single execution) of all these algorithms are given. The time needed for execution of all trained algorithms for one test measurement was similar and very short (in ms).

From the MAE and RMSE values, shown in Table III, it can be inferred that the best overall performance in a case of regression was achieved by using Adaptive Boosting based on Decision Trees. Also, by analyzing tables III and IV it can be inferred that the simpler algorithms, such as Multiple Linear Regression and Decision Tree take shortest time to train (and execute). On the other hand, the algorithms that

consist of a large number of decision trees (i.e. Random Forest, AdaBoost or Extra Trees), SVM and deep learning algorithms, take the longest time to train (and execute) due to the complexity.

TABLE III
REGRESSION ALGORITHMS - MAE AND RMSE VALUES

Algorithm	MAE		RMSE	
	Mean	Std	Mean	Std
Random Forest	0.515066	0.065481	4.202714	1.167030
Decision Tree	0.652400	0.094673	5.911044	1.294427
AdaBoost	0.102419	0.02371	1.056525	0.477244
GradBoost	0.492849	0.049863	3.154737	0.952986
Extra Trees	0.469917	0.043560	2.800131	0.763884
KNN	7.586408	0.243762	17.655297	1.181481
SVM	6.811546	0.204737	12.891746	1.510237
MLR	35.95306	0.51608	52.938264	1.547197
DL#1	1.857936	0.407002	3.732621	0.413279
DL#2	1.552141	0.340354	3.413873	0.465218
DL#3	1.819313	0.501331	3.687509	0.576523
DL#4	1.616337	0.356425	3.417748	0.458047
DL#5	1.719751	0.539735	3.565049	0.655192

TABLE IV
REGRESSION ALGORITHMS - DURATION OF TRAINING (SINGLE EXECUTION)

Algorithm	Time [s]		Alg.	Time [s]	
	mean	std		mean	std
Random Forest	142.762	8.4116	MLR	0.007	0.0112
Decision Tree	0.215	0.0043	DL#1	170.780	27.4982
AdaBoost	69.558	2.3812	DL#2	250.169	16.4083
GradBoost	42.799	1.1223	DL#3	494.753	20.3062
Extra Trees	103.272	2.2578	DL#4	354.726	24.7223
KNN	0.499	0.0156	DL#5	627.936	30.0591

Furthermore, a more detailed statistical and error analysis (i.e. the minimum and maximum error values, the threshold values corresponding to 25%, 50% and 75% of instances), as well as time needed for training of AdaBoost algorithm are shown in Table V.

TABLE V
DETAILED ANALYSIS OF ADABOOST ALGORITHM FOR REGRESSION TASK

	MAE	MSE	RMSE	R^2	Time [s]
Mean	0.10242	1.338014	1.056525	0.999965	69.558
Std	0.02371	1.122726	0.477244	0.000029	2.3812
Min	0.06659	0.178894	0.422959	0.999895	64.468
25%	0.08444	0.40364	0.635299	0.999943	67.793
50%	0.09975	0.807562	0.897779	0.999978	69.868
75%	0.11428	2.176002	1.474417	0.999989	71.291
Max	0.15632	3.988713	1.997176	0.999995	75.529

As obvious in Table V, 50% of the MAE values for AdaBoost algorithm are under 0.1, with its mean value being just over 0.1. These are by far the best values of MAE for all of the observed regression algorithms that were compared in this paper.

B. Classification algorithms

The analysis showed that all the observed classification algorithms have the convergence rate of 100%, which means that the algorithms manage to converge around the mean values of accuracy in all of the 42 independent executions. Besides machine learning algorithms, three Deep Learning models, marked DL#6 to DL#8, were designed, optimized and used. These neural networks models are defined as: DL#6 model with 2 hidden layers with 128 neurons in each layer, DL#7 model with 2 hidden layers with 256 neurons in each layer, and DL#8 model with 2 hidden layers with 512 neurons in each layer. Data is normalized in the input layer of every neural network model. The activation function of each hidden layer is the ReLU function and the activation function of the output layer is the Softmax function. The loss function is Binary Cross-entropy, and the metrics of the loss function is the binary accuracy function with the 0.5 threshold value. The Adam optimization function was used with the learning rate of 0.001, and every neural network model is trained over 100 epochs.

The mean values and corresponding standard deviations (std) of classification accuracy for all observed classification algorithms, as well as the times needed for the training are given in Table VI.

Based on accuracy values for different algorithms, shown in Table VI, it can be inferred that the best algorithm for the classification task is the Gradient Tree Boosting. Also, the difference in time needed to train (and execute) more and less complex classification ML algorithms is not as big as it is in case of regression algorithms. This can be explained by the fact that the classification problem is easier to solve, and it does not require as much time as the regression one. Deep learning algorithms for classification take longer to train, since these were trained over 100 epochs.

TABLE VI
CLASSIFICATION ALGORITHMS - ACCURACY AND DURATION OF TRAINING

Algorithm	Accuracy		Time [s]	
	mean	std	mean	std
Random Forest	0.998683	0.00052	26.04913	0.303142
Random Forest Hybrid	0.998388	0.000612	18.30846	0.115094
Decision Tree	0.99822	0.000775	0.096139	0.003632
AdaBoost	0.998233	0.000753	0.104626	0.003506
GradBoost	0.999422	0.000432	24.62098	3.891909
Extra Trees	0.990682	0.001651	12.96546	0.247514
KNN	0.92313	0.004718	0.232663	0.008677
SVM	0.976849	0.00237	67.03444	6.434671
SGD	0.695058	0.009976	0.444438	0.019513
DL#6	0.994421	0.001118	118.5808	22.9412
DL#7	0.994591	0.001186	139.3796	2.512952
DL#8	0.994536	0.001218	440.6514	24.03057

The more detailed analysis of the Gradient Tree Boosting algorithm is shown in Table VII (the minimum and the maximum accuracy values are given, as well as threshold values corresponding to 25%, 50% and 75% of instances, and time needed for training), while estimated confusion matrix for this algorithm is shown on Fig. 2. It can be seen that in a case of Gradient Tree Boosting algorithm, only 3 of 7088 independent measurements of the test set used were misclassified (see confusion matrix in Fig. 2).

When compared to the classification results achieved in [3], we here achieved slightly better results for classification accuracy for the same algorithms that were used in both papers. However, in this paper the number of epochs for training the DL algorithms was higher than in [3], which can be one of the reasons for better accuracy results. Yet, the novelty of our paper, when compared to the work in [3], is that we included a number of algorithms that were not

implemented in [3], for which we here achieved even better results in classification accuracy.

TABLE VII
DETAILED ANALYSIS OF GRADBOOST ALGORITHM

	Accuracy	Time [s]
mean	0.999422	24.62098
std	0.000432	3.891909
min	0.998025	22.1305
25%	0.999154	22.70442
50%	0.999436	22.99523
75%	0.999718	23.20253
max	1	35.08803

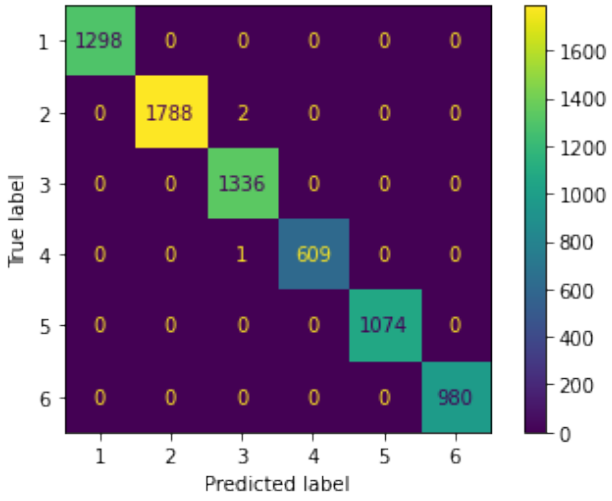

Fig. 2. Confusion matrix for the GradBoost algorithm

VI. CONCLUSION

Alongside global warming, the air pollution is one of the most alarming global ecological problems. Thus, developed countries, international health organizations, as well as some international companies are investing money in to reduce the impact air pollution have on global health. Also, some of air pollution aware companies try to motivate people to contribute to the cause, by giving them a chance to connect their air quality sensors with the global network of sensors, created by these companies. I.e., one of the most famous companies and websites that does this is called IQ Air [8].

The main topics covered in this paper are calculating the AQI (regression task), and the classification of air pollutant measurements into different air quality classes. We observed a wide set of machine learning and deep learning regression and classification algorithms for these tasks, and presented the performance comparison of these algorithms, based on the values of MAE, RMSE and accuracy, as well as the time needed to execute these algorithms. In total, 8 and 9 ML algorithms, as well as 5 and 3 DL models, were observed for regression and classification tasks, respectfully. It is shown that the AdaBoost algorithm presents best choice in the case of regression task, while the GradBoost algorithm presents the best choice in the case of classification tasks.

The presented results, as well as the designed and trained algorithms, present a foundation of a forecasting model, for predicting the missing and future pollution measurements and values of air quality index. This forecasting model could be used as a part of mobile application, which would inform users about the daily and weekly predictions of the pollution levels. This is one of the ideas for the future works. Another possible way of using the designed and trained algorithms, would be implementing in industrial plants.

ACKNOWLEDGMENTS

This work has been partly supported by the Ministry of Education, Science and Technological Development of the Republic of Serbia.

REFERENCES

[1] I. N. Athanasiadis, V. G. Kaburlasos, P. A. Mitkas, and V. Petridis, “Applying Machine Learning Techniques on Air Quality Data for Real Time Decision Support”, 1st Intl. Symposium on Information Technologies in Environmental Engineering (ITEE 2003), pp. 51, 2003, ICSC/NAISO Academic Press.

[2] M. Castelli, F. M. Clemente, A. Popovič, S. Silva, and L Vanneschi, "A Machine Learning Approach to Predict Air Quality in California", Complexity, vol. 2020, Article ID 8049504, 23 pages, 2020.

[3] M. Sharma, J. Samyak, S. Mittal, and T. Sheakh, “Forecasting and Prediction of Air Pollutants Concentrates Using Machine Learning Techniques: The Case of India”, IOP Conference Series: Materials Science and Engineering, ICCRDA 2020, Vol. 1022, 012123, 2021.

[4] Electricity generation by fuel and scenario, 2018-2040, IEA, Paris https://www.iea.org/dataand-statistics/charts/electricity-generation-by-fuel-and-scenario-2018-2040 (last time accessed on 10.06.2021.)

[5] https://www.health.nsw.gov.au/environment/air/Pages/common-air-pollutants.aspx (last time accessed on 10.06.2021.)

[6] N. Zimmerman, A. A. Presto, S. P. N. Kumar, J. Gu, A. Hauryliuk, E. S. Robinson, A. L. Robinson and R. Subramanian, "A machine learning calibration model using random forests to improve sensor performance for lower-cost air quality monitoring", Atmos. Meas. Tech., vol. 11, no. 1, 2018, pp. 291–313.

[7] https://data.world/ (last time accessed on 10.06.2021.)

[8] https://openaq.org/ (last time accessed on 10.06.2021.)