Proceedings of the Seventh International Conference on Research in Intelligent and Computing in Engineering pp. 89–94
DOI: 10.15439/2022R03
ISSN 2300-5963 ACSIS, Vol. 33

Analysis and Prediction for Air Quality Using Various Machine Learning Models

To-Hieu Dao[1,2]
[1]EEE, Phenikaa University
[2]GU of Science & Technology
Ha Noi, Viet Nam
hieu.daoto@phenikaa-uni.edu.vn

Hoang Van Nhat
Electrical & Electronic Engineering
Phenikaa University
Ha Noi, Viet Nam
van.nhathoang@aalto.fi

Hoang Quang Trung
Electrical & Electronic Engineering
Phenikaa University
Ha Noi, Viet Nam
trung.hoangquang@phenikaa-uni.edu.vn

Vu Hoang Dieu[1,2]
[1]EEE, Phenikaa University
[2]GU of Science & Technology
Ha Noi, Viet Nam
dieu.vuhoang@phenikaa-uni.edu.vn

Nguyen Thi Thu
Electrical & Electronic Engineering
Ha Noi University of Industry
Ha Noi, Viet Nam
thunt@haiu.edu.vn

Duc-Nghia Tran
Institute of Information Technology
Viet Nam Academy of Science & Technology
Ha Noi, Viet Nam
nghiatd@ioit.ac.vn

Duc-Tan Tran*
Electrical & Electronic Engineering
Phenikaa University
Ha Noi, Viet Nam
tan.tranduc@phenikaa-uni.edu.vn

***Abstract*—Air pollution has been a concern in recent years. Measuring the extent of pollution is important to know about the air quality. Previous research has used machine learning algorithms to forecast the Air Quality Index (AQI) in specific locations. Even though that research achieved quite reliable results, they still have some drawbacks that need to be taken into consideration, such as low accuracy or lack of data analysis.On a public dataset, we used Random Forest, XGBoost, and Neural Network to build a machine learning model for the purpose of making predictions about the air quality index (AQI) in a number of cities located in India. The performances of these models were evaluated by using their score errors, Root Mean Square Error (RMSE), and Coefficient Of Determination (R^2). This paper demonstrates the analysis of air pollutants from the dataset, which is an effective way to enhance the model's performance.**

***Index Terms*—Machine Learning, AQI, Data Analysis**

I. Introduction

Recent economic and social developments have had an effect on various environmental variables, including the land, water resources, and air. Because of this, wireless sensor network-based air quality monitoring is a popular research topic [1], [2]. According to WHO [3], seven million deaths were related to air pollution each year.

Based on the computation of pollutants that are harmful to human health, the air pollution level index can be created [4]. The Air Quality Index (AQI) is the name of this index, which ranges from 0 to 500. A high AQI is not good for people. There are distinct ways to calculate the AQI, such as using the formula or using machine learning techniques. In 2018, study led by Samir Lemes and colleagues demonstrated the disparity between several approaches to estimate the AQI by calculating and ranking AQI values according to certain criteria [5]. They then used these parameters to calculate the levels of air pollution in two different parts of Bosnia and Herzegovina. The final result of their work illustrates the comparison of AQI values on the same dataset, which was obtained by using different methods of US AQI, EU AQI, and SAQI_11 standards.

Machine learning, while on the other hand, has demonstrated its superior effectiveness by combining knowledge from various fields, such as statistics, artificial intelligence, and computer science [6]–[8]. In recent years, the use of machine learning to predict AQI values has become common and has piqued the interest of researchers [9]–[11]. Although the models constructed in the experiments performed well, they still have some limitations, such as filling in missing values, feature significance analysis, and feature creation to fully exploit the dataset. The final results were then evaluated using three machine algorithms with different validation criteria.

Corresponding(*): Duc-Tan Tran. Email: tan.tranduc@phenikaa-uni.edu.vn
Address: Phenikaa University, Ha Noi, Viet Nam

II. AQI Calculation And Dataset

A. AQI Calculation

In order to get the AQI value, there are several different methods to calculate it worldwide. For example, the AQI formula for China, which is based on the National Ambient

Air Quality Standard of China (NAAQS-1996), differs from the AQI calculation method defined by the US Environmental Protection Agency (1994) and from the method developed by India (NAAQS Dependent Air Quality Index) [4]. Therefore, in this work, we used the estimating formula for China as the reference in comparison with using the Machine Learning approach.

Based on the method proposed by NAAQS-2012, the components used in the formula include 6 pollutants ($PM10, PM2.5, SO_2, Ozone, NO_2$, and CO) and 7 indexes, including the maximum 8-hour Ozone concentration (mg/m^3), the maximum 1-hour Ozone concentration, and the daily average concentration of $SO_2, NO_2, CO, PM10$, and $PM2.5$. The calculation by Eqa. (1) for each individual pollutant is:

$$AQI = \frac{AQI_h - AQI_l}{BP_h - BP_l} \times (C_Q - BP_l) + AQI_l \qquad (1)$$

Where, C_Q is the pollutant Q's daily mean value;The pollution levels for substance Q are, respectively, AQI_h and AQI_l, with the corresponding estimated highs and lows being BP_h and BP_l. The final AQI value is the largest value in the AQI series by Eqa. (2), obtained after completing each AQI math operation:

$$AQI = \max(AQI_0, AQI_1, ...AQI_n) \qquad (2)$$

Where n is the number of pollutants considered. In this experiment, we decided to use the Machine Learning approach to predict the AQI value because the first method is quite time consuming, and complicated. Most importantly, the dataset is not always available to be calculated by the formula, which requires information on pollutant concentrations both daily and hourly.

B. Dataset

This research employed a publicly accessible dataset containing 29531 instances of Indian air quality. This data set was collected over a six-year period (January 2015 to June 2020), allowing us to evaluate proposed air quality calculation methods. Each instance has had the average daily AQI and some other pollutants from different stations in cities across India. The Central Pollution Control Board [12], the official website of the Government of India, provides the dataset. There are 12 features that have been recorded, including some significant air pollution contaminants like particulate matter ($PM2.5$ and $PM10$), ozone (O_3), nitrogen oxides (NO), NO_x, nitrous dioxide (NO_2), sulfur dioxide (SO_2), carbon monoxide (CO) emissions, amoniac NH_3 and other chemical occurrences (benzen, toluene, xylene). However, there are some important features that contribute mostly to the value of the AQI, which are particulate matter ($PM2.5$ and $PM10$), CO, NO_2, and SO_2. On top of that, NO_x, which is associated with acid rain, photochemical smog, and tropospheric ozone destruction, is another indicator for AQI prediction [13].

In the dataset, time plots are significant for some analysis related to changes in AQI over months and years, which helps us choose an effective method to predict the AQI value along with time series. We did some analysis regarding the changes in all data features and the AQI value according to year in Fig. 1. The total value is the sum of all pollutants recorded in all cities at different times throughout the given period. It is evident to note that there is an upward trend when it comes to the pollutants and AQI values throughout the 6-year period. The last 3 years from 2018 to 2020 witnessed the highest figures of these pollutants. According to this, 2019 and 2020 are the most polluted years recorded, in which the AQI value and particulate matter peaked in October, November, and December.

According to [4], there are some main pollutants that lead to high degrees of air pollution. Thus, we used the total value recorded in five main indexes, including AQI, $PM10$, $PM2.5$, CO, and NO_2 to rank the most polluted cities. The visualization is displayed in Fig. 2.

Among the five most polluted cities above, they all recorded high levels of five pollutants, which are AQI, $PM10$, $PM2.5$, CO, and NO_2. On average, Ahmedabad is the most polluted city when it comes to the AQI value, at almost 450 on average. Second in terms of pollution are Delhi, Patna, Gurugram, Lucknow, and so on, which had a high degree of pollutants including AQI, $PM10$, $PM2.5$, and NO_2. However, the CO value was record-high in only Ahmedabad, with more than 20, whereas in other cities this substance only ranged from 0 to approximately 2.

III. Data Preprocessing And Method

The data preprocessing is the first and most important step, which not only results in a good validation result but also improves the predictive performance of the model later on. This stage often includes missing data imputation, removing strange datapoints, feature engineering techniques, and feature selection. The two first steps help us have a full set of data, improving the accuracy of the models. Meanwhile, selecting useful features can reduce running time, minimize overfitting while running the model.

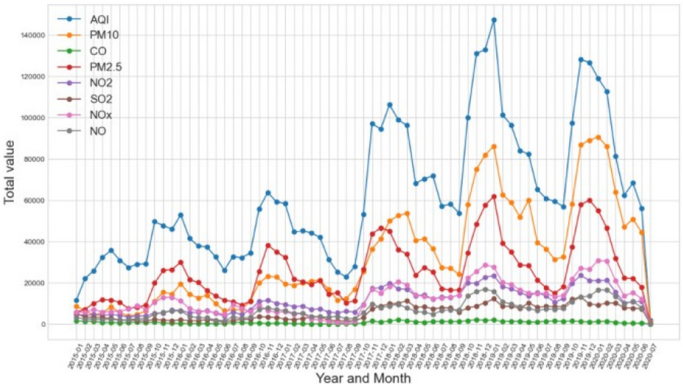

Fig 1. Air pollutants depicted by time.

TAB I
STATISTICS OF NULL VALUES

	Pollutants	Value	Percentage (%)
1	Xylene	18109	61.3
2	$PM10$	11140	37.7
3	NH_3	10328	35.0
4	Toluene	8041	27.2
5	Benzene	5623	19.0
6	AQI	4681	15.9
7	$PM2.5$	4598	15.6
8	NO_x	4185	14.2
9	O_3	4022	13.6
10	SO_2	3854	13.1
11	NO_2	3585	12.1
12	NO	3582	12.1
13	CO	2059	7.0

A. Missing Data Imputation

Tab. I shows the missing value percentage for each column in the dataset. From the given result, the missing data proportion is distributed mostly in xylene, $PM10$, NH_3, and toluene, which are 61.3%, 37.7%, 35.0%, and 27.2%, respectively. Data loss rates in other situations range from 12% to 19%. This issue can be resolved in a number of ways, including by eliminating dropped data points or by adding the most common value from each case to the missing data. In this study, the K-Nearest Neighbors Imputer (KNNImputer) technique is used to overcome losing information [14]. At this stage, each sample's missing values are imputed by the mean value, calculated from 3-neighbors nearest data points in the dataset. This technique was used because it is easy to use and works well. It is also more accurate than simple imputation.

B. Feature Engineering

New features that are created based on the features in the dataset can be very helpful to improve the performance of the model. In this step, we used the mathematical transform, which groups some existing features into a new one that has a good association with the target. In the dataset, we came up with 3 new features by using this method. The first one is "$ParticulateMatters_i$", which was made by adding the value of $PM10$ to $PM2.5$ together. The second new feature is "$Nito_i$", obtained by the sum of NO_2, NO_x, and NO. Finally, the attribute $average\_N_i$ is the average value of N previous AQI data points. In addition, year and month are two time features extracted from data information. Three new numerical features are depicted by Eqa. (3)-(5) below:

$$ParticulateMatters_i = PM10_i + PM2.5_i \tag{3}$$

$$Nito_i = (NO_2)_i + NO_i + (NO_x)_i \tag{4}$$

$$average\_N_i = \frac{\sum_{k=i-N}^{i-1} AQI_k}{N} \tag{5}$$

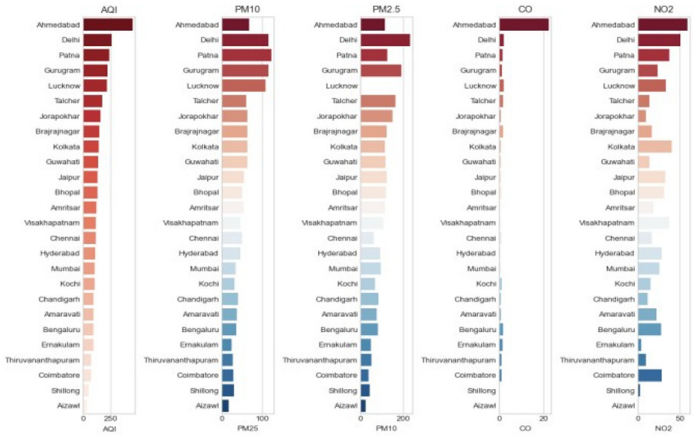

Fig 2. Most polluted cities towards five pollutants.

C. Method

Our method to estimate the AQI value used nine main features, which were reached in the previous stages. We implemented steps in order to get the AQI prediction. Firstly, after being preprocessed as well as experienced data selection and data engineering, selected features were divided into 2 subsets called Training set and Test set. In which Training set accounted for 80% of the total dataset while Test set held the remaining volume, at 20%. The aim of the division is to validate the model's performance later on. In this work, we used three machine learning algorithms, namely Random Forest Regression, Gradient Boosted Regression, and Neural Network Regression, to train three models on the training set. Then, the trained models could be applied to the test set to deal with the unknown data, and get the target prediction. Finally, we used some criteria to assess its effectiveness with regard to model prediction. Fig 3 shows the steps we undertook in our study:

1) Random Forest Regression (RFR): This algorithm is a synthetic prediction algorithm that integrates many different models to create more efficient models. Random Forest (RF) consists of many decision trees, each of which predicts a certain object well and is different from the others [15]. By averaging the results, we were able to significantly reduce the number of overfitting while maintaining the model's good predictive score. The steps are as follows:

Step 1. From the initial dataset, we need to build several subsets of data. The technique used to do this task is called the bootstrapping method, in which, from $n\_samples$ data points, we repeatedly choose random data points with replacement. The result is n datasets called bootstrap samples, which have the same size as the original dataset but in which some data points will be absent or some will be repeated [16]. This method guarantees that each bootstrap sample is modestly different from the others.

Step 2. For each new dataset, build a decision tree with a slight modification: instead of choosing the best test for a specified node, in each node we randomly choose k features

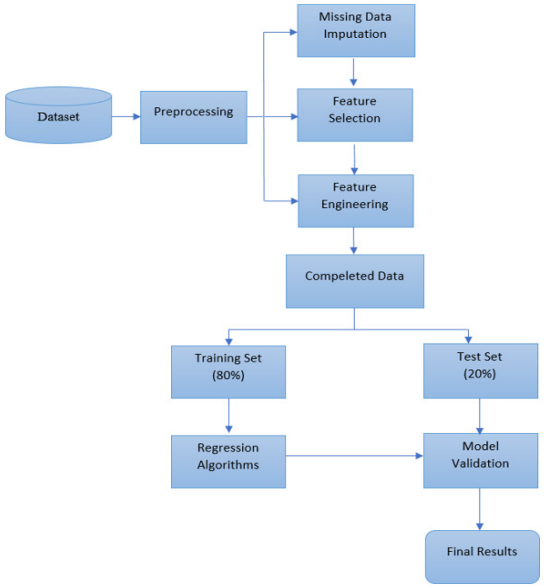

Fig 3. Machine Learning prediction steps.

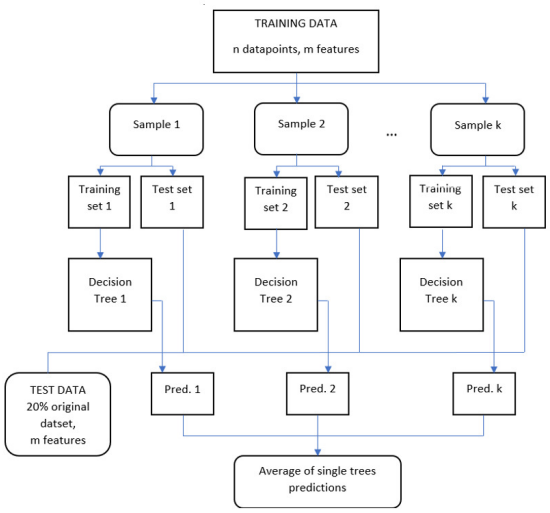

Fig 4. Flow chart of Random Forest Regression.

(out of the total $n\_features$, where $k < n$), and choose the best split among these chosen attributes. By doing this, each tree will perform differently on k distinct selected features, leading to different performances each time.

Step 3. To make a prediction on the unknown dataset, the algorithm uses the predictions obtained in **step 2** and averages the results to get the final prediction.

Fig. 4 and Fig. 5 demonstrate the flowchart of the Random Forest Model based on the research of Lingjian Yang in 2017 [17], and the flowchart of Decision Tree Model which is a part of the Random Forest Algorithm based on the work done in 2017 by Ibrahim A Ibrahim [18].

2) Gradient Boosted Regression (XGBoost): In this study, Gradient Boosted Regression (XGBoost) [19] was implemented as a variant of Gradient Boosted Regression Trees. Similar to RF, Gradient Boosted (GB) models are built by many simple decision trees (weak learners), with a depth of one to five. However, the idea behind this model is that each tree can better predict and correct the mistakes of the previous ones. This results in the overall performance of the GB model being improved by adding more trees, and it can make more accurate predictions than the RF model if the parameters are set up meticulously. Therefore, XGBoost requires high accuracy and reliability from datasets. However, it requires careful tuning of the parameters and takes a long time to run.

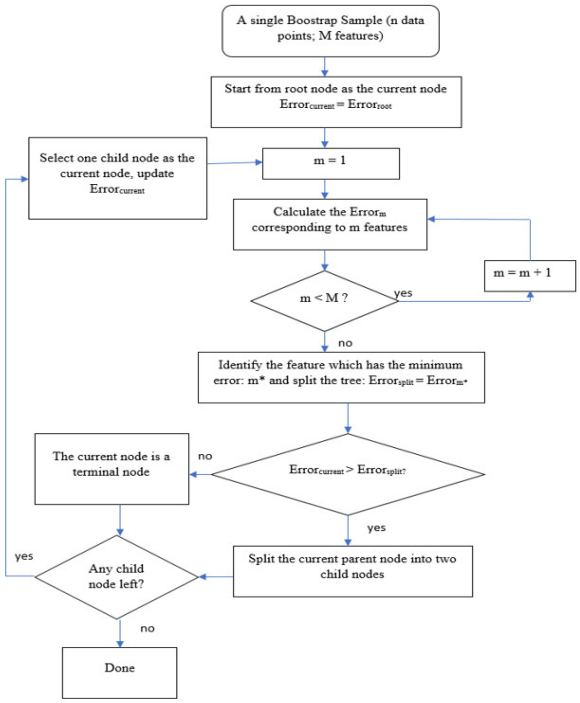

Fig 5. Flowchart of each bootstrap Sample in Random Forest.

3) Neural Network: Neural Network or Multilayer Perceptrons is a type of linear model that uses various stages of processing to get the final output. A multilayer model can perform efficiently with a large dataset, constructing a very complex model [20]. However, the models require quite a bit of running time as well as meticulous fine-tuning of the parameters. There are some parameters in this model that we have to take into account while implementing. First, hidden layer sizes (HLS), which is the number of hidden layers in the models; the number of units in each hidden layer; and the regularization, which is used to control the model's complexity.

D. Validation

In this study, overall performance was assessed using three indexes: 1) Mean Absolute Error (MAE), which is the absolute difference between the observed value (y_i) and the predicted result ($\hat{y}_i$). The lower the MAE, the closer the predicted result is to the actual value, and $MAE = 0$

is the ideal value; 2) $RMSE$ [21] is the average of the difference between $\hat{y}_i$ and y_i. The lower the $RMSE$, similar to MAE, the closer $\hat{y}_i$ is to y_i. The higher the $RMSE$, the more dispersed the $\hat{y}_i$ values are over a wider range; 3) The coefficient of determination (R_2) [22] has a value range of 0 to 1, indicating how close the predictions $\hat{y}_i$ are to the true value y_i of the model. When $R_2 = 1$, the ideal prediction is understood because it perfectly fits the real data and maximizes performance. In contrast, as R_2 approaches zero, the model becomes less reliable. In summary, a good model is satisfied when the $RMSE$, MAE, and R_2 are low. Eqa. (6)-(8) determines the above three indexes.

$$MAE = \frac{\sum_{i=1}^{n} |\hat{y}_i - y_i|}{n} \tag{6}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (\hat{y}_i - y_i)^2} \tag{7}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{n} (\bar{y}_i - y_i)^2} \tag{8}$$

Where, $\bar{y}_i$ is the average AQI value at data point i.

IV. RESULTS AND DISCUSSION

In the study, the target of the model is the AQI in various Indian cities. In order to get the most precise prediction, we split the dataset into 2 parts: a training set containing 80% of the total data, which was used to train the model; and a test set holding the rest of the data points, which was used in the validation step. In order to choose the parameters for each model, we used a method called Grid Search [23], which tried different values of parameters in the model and then chose the optimized set containing the best ones.

A. Prediction of the AQI Applying Random Forest Regression

In this research, we employed two values: $n\_estimators$ and $n\_features$. The number of decision trees included in the model is $n\_estimators$, and the subset of features in each decision tree is $n\_features$. We applied Grid Search to locate the $n\_stimators$ parameter. Meanwhile, $n\_features$ were obtained by taking the square of the total features. According to that, the best combination of parameters used in the work was $n\_estimators = 500$ and $n\_features = 3$. The criteria for validating the model were: $MAE = 19.18$, $R2 = 0.94$, and $RMSE = 33.22$.

B. Performance Of XGBoost Model In Predicting AQI

For the XGBoost model, we also used Grid Search to choose two parameters, $n\_estimators$ and $learning\_rate$, which are the number of trees and the rate at which a tree can fix the mistakes of the previous ones. Meanwhile, the third parameter n_jobs, the selected set of parameters was $n\_estimators = 300$, $learning\_rate = 0.02$, and $n\_jobs = 4$. The model's statistical criteria were: $MAE = 18.98$, $R^2 = 0.942$, and $RMSE = 32.6$.

TAB II
PERFORMANCE OF MODEL BASED ON 3 CRITERIA.

Methods	MAE	R^2	$RMSE$
Random Forest (RFR)	19.18	0.94	33.22
XGBoost	18.98	0.942	32.6
Neural Network	22.36	0.928	36.39

C. AQI Prediction Of Neural Network Model

According to the previous section, we had two parameters in Neural Network models: hidden layer sizes (HLS) and the number of units in each hidden layer (α). Using the Grid Search method, we obtained the optimal values for the two parameters with $HLS = 50$, and $\alpha = 0.5$. The results of the validation criteria were $MAE = 22.36$, $R^2 = 0.928$, and $RMSE = 36.39$. Tab. II displays the comparison of the three models' performances over the three corresponding criteria.

As can be seen from the result, while Random Forest and XGBoost got performances that are approximately the same in both three criteria, Neural Network, however, performed less efficiently with the same conditions. The MAE and $RMSE$ of this model are much higher, at 23.36 and 36.39, respectively, yet $R_2 = 0.928$ is lower than that of the two other algorithms. As a result, XGBoost is the most effective among the three models when it comes to the statistical criteria, with $MAE = 18.98$, $R_2 = 0.942$, and $RMSE = 32.6$. Fig. 6 shows the comparison between the result of the XGBoost model's prediction and actual AQI values with 500 samples. Look at this diagram. The predicted value line (orange) closely follows the actual value line (blue). The distinction is insignificant. This is consistent with the $MAE = 18.98$ value that we measured.

Huixiang Liu *et al.* [10] used two indices to examine the difference in AQI prediction in Beijing, China: correlation (R_2) and mean of difference ($RMSE$). They used Support Vector Regression (SVR) and Random Forest Regression (RFR) in their study and obtained two sets of indices ($R2 = 0.9760, RMSE = 94.4918$) and ($R2 = 0.8401, RMSE = 83.6716$). These two indexes are also used by Chao Song and Xiaoshuang Fu in their paper [[24]. They integrated a set of algorithms into the one called Combination Forecasting Model (CFM) to get the predictions of AQI in Zengzhou and Shanghai, China. Their results finally reached $RMSE = 36.89$ and $R_2 = 0.86$ for the dataset collected in Zhengzhou, and ($RMSE = 35.32, R_2 = 0.72$) for the other location – Shanghai. Even though these works are different from our research because of the dataset, Machine Learning algorithms, and some other criteria used to evaluate the models, it is suggested that our models achieved quite good results compared to those of other research when assessed using the same criteria.

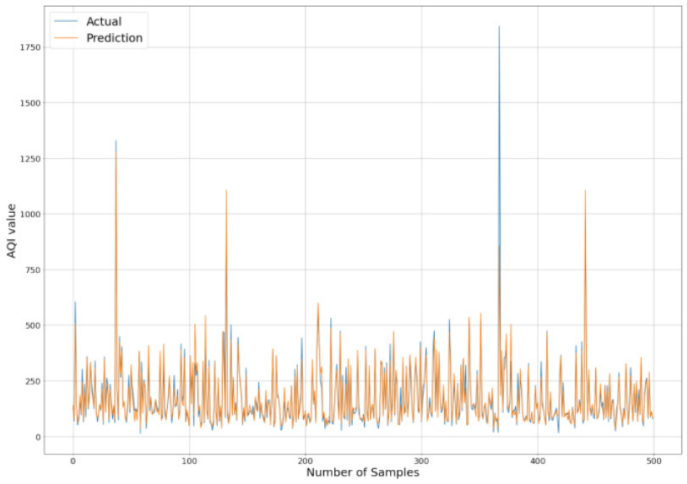

Fig 6. Correlation between AQI values predicted and measured using the XGboost algorithm.

V. CONCLUSIONS

Air quality has a direct impact on human life and society as a whole. As a result, not only the government, but also individuals and organizations, must work together to prevent environmental pollution, particularly air pollution. As a result, the AQI index is needed to evaluate air quality, and it can also be used to design and produce intelligent meteorological monitoring devices. In this paper, the air pollution indicators in many Indian cities were analyzed and predicted in this study using real data on pollutants provided by the Indian government. The study's findings demonstrated that, while all three models provide good predictive results, the XGB model outperforms the others. Meanwhile, the Neural Network model, which requires careful tuning parameters, and much operating time, was not as effective as XGB and RFR. To conclude, the study showed the data analysis and transformation, then built three models for AQI predictions, which attempted to improve the performance in terms of each model's accuracy. In the future, we expect to develop algorithms on devices that use low-power microcontrollers to predict air quality remotely [6], [8].

REFERENCES

[1] D.-C. Nguyen, T. Duc-Tan, and D.-N. Tran, "Application of compressed sensing in effective power consumption of WSN for landslide scenario," in *2015 Asia Pacific Conference on Multimedia and Broadcasting*, 2015, pp. 1–5.

[2] D. T. Pham, D. C. Nguyen, V. V. Pham, B. C. Doan, and D. T. Tran, "Development of a Wireless Sensor Network for Indoor Air Quality Monitoring," in *The 2015 International Conference on Integrated Circuits, Design, and Verification*, Vietnam, 2015, pp. 178–183.

[3] H. Gu, W. Yan, E. Elahi, and Y. Cao, "Air pollution risks human mental health: an implication of two-stages least squares estimation of interaction effects," *Environmental Science and Pollution Research*, vol. 27, no. 2, pp. 2036–2043, 2020.

[4] S. Kumari and M. K. Jain, "A Critical Review on Air Quality Index," *Environmental Pollution*, vol. 77, pp. 87–102, 2018.

[5] S. Lemeš, "Air Quality Index (AQI)—comparative study and assesment of an appropriate model For B&H," in *2th Scientific/Research Symposium with International Participation 'Metallic And Nonmetallic Materials*. MNM, 2018, pp. 282–291.

[6] N. H. Van, P. Van Thanh, D. N. Tran, and D.-T. Tran, "A new model of air quality prediction using lightweight machine learning," *International Journal of Environmental Science and Technology*, 2022. [Online]. Available: https://doi.org/10.1007/s13762-022-04185-w

[7] N. C. Minh, T. H. Dao, D. N. Tran, Q. H. Nguyen, T. T. Nguyen, and D. T. Tran, "Evaluation of Smartphone and Smartwatch Accelerometer Data in Activity Classification," in *2021 8th NAFOSTED Conference on Information and Computer Science (NICS)*, 2021, pp. 33–38.

[8] N. T. Thu, T.-h. Dao, B. Q. Bao, D.-n. Tran, P. V. Thanh, and D.-T. Tran, "Real-Time Wearable-Device Based Activity recognition Using Machine Learning Methods," *International Journal of Computing and Digital Systems*, vol. 12, no. 1, pp. 321–333, 2022. [Online]. Available: https://dx.doi.org/10.12785/ijcds/120126

[9] J. K. Sethi and M. Mittal, "A new feature selection method based on machine learning technique for air quality dataset," *Journal of Statistics and Management Systems*, vol. 22, no. 4, pp. 697–705, 2019. [Online]. Available: https://doi.org/10.1080/09720510.2019.1609726

[10] H. Liu, Q. Li, D. Yu, and Y. Gu, "Air quality index and air pollutant concentration prediction based on machine learning algorithms," *Applied Sciences (Switzerland)*, vol. 9, no. 19, 2019.

[11] 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, pp. 1–23, 2020. [Online]. Available: https://doi.org/10.1155/2020/8049504

[12] P. Bhawan and E. A. Nagar, "Central Pollution Control Board," pp. 1–93, 2019.

[13] R. R. Dickerson, D. C. Anderson, and X. Ren, "On the use of data from commercial NOx analyzers for air pollution studies," *Atmospheric Environment*, vol. 214, no. June, p. 116873, 2019. [Online]. Available: https://doi.org/10.1016/j.atmosenv.2019.116873

[14] S. Zhang, "Nearest neighbor selection for iteratively kNN imputation," *Journal of Systems and Software*, vol. 85, no. 11, pp. 2541–2552, 2012. [Online]. Available: http://dx.doi.org/10.1016/j.jss.2012.05.073

[15] Y. Liu, Y. Wang, and J. Zhang, "New machine learning algorithm: Random forest," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 7473 LNCS, pp. 246–252, 2012.

[16] Y. Fang and J. Wang, "Selection of the number of clusters via the bootstrap method," *Computational Statistics and Data Analysis*, vol. 56, no. 3, pp. 468–477, 2012. [Online]. Available: http://dx.doi.org/10.1016/j.csda.2011.09.003

[17] L. Yang, S. Liu, S. Tsoka, and L. G. Papageorgiou, "A regression tree approach using mathematical programming," *Expert Systems with Applications*, vol. 78, pp. 347–357, 2017. [Online]. Available: http://dx.doi.org/10.1016/j.eswa.2017.02.013

[18] I. A. Ibrahim, T. Khatib, A. Mohamed, and W. Elmenreich, "Modeling of the output current of a photovoltaic grid-connected system using random forests technique," *Energy Exploration and Exploitation*, vol. 36, no. 1, pp. 132–148, 2018.

[19] Y. Wang, Z. Pan, J. Zheng, L. Qian, and M. Li, "A hybrid ensemble method for pulsar candidate classification," *Astrophysics and Space Science*, vol. 364, no. 8, 2019.

[20] D. Faraggi and R. Simon, "A neural network model for survival data," *Statistics in Medicine*, vol. 14, no. 1, pp. 73–82, 1995.

[21] C. J. Willmott and K. Matsuura, "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance," *Climate Research*, vol. 30, no. 1, pp. 79–82, 2005.

[22] A. H. Murphy, "The coefficients of correlation and determination as measures of performance in forecast verification," *Weather and Forecasting*, vol. 10, no. 4, pp. 681–688, 1995.

[23] A. C. Müller and S. Guido, *Introduction To Machine Learning With Python: A Guide For Data Scientists*. O'Reilly Media, Inc., 2016.

[24] C. Song and X. Fu, "Research on different weight combination in air quality forecasting models," *Journal of Cleaner Production*, vol. 261, p. 121169, 2020. [Online]. Available: https://doi.org/10.1016/j.jclepro.2020.121169