

# PM<sub>2.5</sub> Concentration Prediction By Data Mining Method



Hung Thuan Nguyen, Chi Quynh Nguyen

**Abstracts:** *The global air pollution is constantly increasing and causing negative effects on human health such as respiratory, cardiovascular diseases and cancers. Recently, pollution in Hanoi has become increasingly worse, especially when PM<sub>2.5</sub> concentration is always at high level. Thus, PM<sub>2.5</sub> prediction is of more urgency to issue early forecasts. Depending on air data including meteorological indicators and air pollution indicators collected in Hanoi, we have proposed a new characteristic extraction method that gave better results when using the same algorithm compared to those of old methods. XGBoost algorithm was applied to predict the concentration of PM<sub>2.5</sub> and the test showed that the accuracy of this algorithm is higher than that of other data mining algorithms while the training time is significantly lower.*

**Keywords:** *air quality forecasting, data mining, PM<sub>2.5</sub> prediction, XGBoost.*

## I. INTRODUCTION

Increasing air pollution is raising many problems concerning human health. According to World Health Organization, air pollution has impact on everybody in all countries. This resulted in 4.2 million premature deaths globally in 2016, 91% of which is from South East Asian and West Pacific countries. The main cause derives from particulate matter sized 2.5 μm or smaller in the air pollution, which are responsible for cardiovascular, respiratory and cancer diseases. The problem of air pollution is more serious in big cities due to the high density of the population causing the increase in air emissions. Besides, the construction of buildings, roads also increases the amount of dust in the air in big cities. Hanoi is facing the increase in air pollution. In September 2019, Hanoi was ranked as one of the most air-polluted cities in the world. The key factor to this situation is the surge of PM<sub>2.5</sub> concentration in the air. This dust type has a negative impact on human health, therefore, predicting the PM<sub>2.5</sub> dust pollution level is increasingly necessary. PM<sub>2.5</sub> prediction methods have been studied in developed countries over the years. The algorithms applied included hybrid system combined with fuzzy inference, Random Forest (RF), Support Vector Machine (SVM) and neural network.

These algorithms gave positive results in terms of predictive accuracy. However, these methods were conducted in datasets collected at different times and locations, so it is difficult to choose a predictive method from the above studies that is suitable for the data about the air in Hanoi. Therefore, we have surveyed various studies related to pollution level of PM<sub>2.5</sub> indicators to get an overview on prediction methods in part 2. On that basis, in part 3, we analyze the collected data, propose a new way to extract features and choose a suitable model training method for PM<sub>2.5</sub> prediction in Hanoi in one hour later. Meteorological indicators are necessary for the prediction, in addition to other pollution indicators (particulate matter with a diameter of 10 μm - PM<sub>10</sub>, concentration of CO<sub>2</sub>, total volatile organic compounds - TVOC) and time factor is also considered to influence the predictions. By this extraction method, we make a comparison between the old extraction method and tests on different prediction models: SVM, RF, MLP (Multi-layer Perceptron) and XGBoost (Extreme Gradient Boosting) in part 4. Finally, we draw the conclusions and discuss future development in part 5.

## II. RELATED WORKS

In this part, we conducted a survey of related studies. First of all, some of the studies applied Adaptive Neuro Fuzzy Inference System (ANFIS) for prediction. The use of ANFIS made progress only when Fuzzy Inductive Reasoning (FIR) was applied; however, it did not make much difference. This inference derived from the study about PM<sub>2.5</sub> concentration in the center of Mexico city. However, this study did not exploit many meteorological factors to predict the PM<sub>2.5</sub> concentration. Another study on PM<sub>10</sub> concentration prediction in Konya city also employed ANFIS, including some meteorological factors: temperature, humidity, pressure and wind speed. In terms of data processing, Output-dependent data scaling (ODDS) was recommended, and it gave a promising result. However, the historical values of PM<sub>10</sub> concentration were not taken into prediction [2][3]. In addition, there are studies that applied other algorithms such as SVM, RF in air quality forecasting. These studies all used meteorological factors and historical values of pollutants as inputs for their algorithms. As for studies using SVM method, the results are positive, but each pollutant was only consistent with a certain kernel function. According to the test results, RBF kernel function produced the best result for SO<sub>2</sub> indicator, while linear function worked the best on NO<sub>2</sub> indicator. Besides SVM, RF is also an algorithm applied by a number of studies in air quality forecasting. A study conducted in Shenyang city (China) built an RF-based RAQ algorithm for air quality forecasting in the city.

Manuscript received on November 29, 2021.

Revised Manuscript received on November 27, 2021.

Manuscript published on November 30, 2021.

\* Correspondence Author

**Hung Thuan Nguyen**, Department of Bachelor of Science, Posts and Telecommunications Institute of Technology, Hanoi, Vietnam. Email: [hungnt61h@gmail.com](mailto:hungnt61h@gmail.com)

**Chi Quynh Nguyen\***, Department of Computer Science, Posts and Telecommunications Institute of Technology, Hanoi, Vietnam. Email: [ching@ptit.edu.vn](mailto:ching@ptit.edu.vn)

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an [open access](https://creativecommons.org/licenses/by-nc-nd/4.0/) article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>)

They built and tested on data sets collected from 10 monitoring stations that included many factors: meteorological data, data of air pollution indicators, traffic data and geography. The prediction method with RAQ algorithm gave outstanding results, with accuracy up to 81.5%, while those of artificial neural network (ANN) and Decision Tree was respectively only 71.8% and 77.4%. Another study applying RF in their prediction method was conducted with data set collected in Warsaw city to predict the average level of contamination of substances over the next day. The method performed consisted of 2 main phases: feature selection and predictive application. In the phase of feature selection, they implemented two methods: Genetic Algorithm (GA) and Stepwise Fit (SF) to remove the features from the original feature set. In predictive application phase, they built two models, one with features through neural networks and other machine learning algorithms (MLP, RBF, SVM) and then RF to synthesize the predicted results of mentioned networks, the other model had the feature as a direct input to the RF. The pollution indicators tested for prediction in this study included: PM<sub>10</sub>, SO<sub>2</sub>, NO<sub>2</sub>, O<sub>3</sub>. The results showed that the selection of features influenced the predicted results, the SF method often gave results higher than GA up to 2.88%. [4] [5][4][6] [7]

In addition to SVM and RF, neural networks are also applied in PM<sub>2.5</sub> prediction. The study on the data set collected in Hefei (China) gave high accuracy when conducting PM<sub>2.5</sub> concentration prediction in the next day by artificial neural network (ANN). Their data included PM<sub>2.5</sub> concentration and meteorological data. The design model had vector input including: PM<sub>2.5</sub> concentration and meteorological factors (temperature, wind speed, wind direction, humidity). The study gave predictions high accuracy with the following measurements: Mean Absolute Error (MAE) [ $\mu\text{g}/\text{m}^3$ ] 0.92472; Root-mean-square Error (RMSE) [ $\mu\text{g}/\text{m}^3$ ]: 1.2756; Coefficient of Determination (R<sup>2</sup>-score): 0.9188; R: 0.9315 [8]. In recent years, Extreme Gradient Boosting (XGBoost) algorithm has emerged in solving this problem. Some studies applying this algorithm have superior accuracy compared to RF, MLP with shorter training time [9] [10].

### III. METHOD

In this part, we present the implementation method including the steps: analyzing the collected data set, proposing feature selection and building predictive model.

#### A. Data description

Our data set was collected at an observation station in Hanoi from August 17, 2018 to July 22, 2019. Each record in the data set contains columns: time, SO<sub>2</sub>, NH<sub>3</sub>, O<sub>3</sub>, PM<sub>2.5</sub>, PM<sub>10</sub>, CO<sub>2</sub>, PM<sub>0.1</sub>, TVOC, CO, temperature, humidity, light. The sampling period was approximately 40 seconds/time. However, the data set had several nulls and noisy records. The distribution of feature values (columns) is shown in Figure 1. The existence of noisy records drove the chart of value distribution of most indicators to incline greatly towards the left. Next we filtered out noisy records and executed feature extraction.

First, we removed the defective and noisy records that are outside the permissible range (for example, PM<sub>0.1</sub> indicator exists negative value or measured temperature is greater than 50 C degrees). Through surveying many studies, meteorological factors: we stored the data of temperature,

humidity, and light, as these are indicators reflecting weather and environmental conditions. They are also key factors in PM<sub>2.5</sub> prediction model.

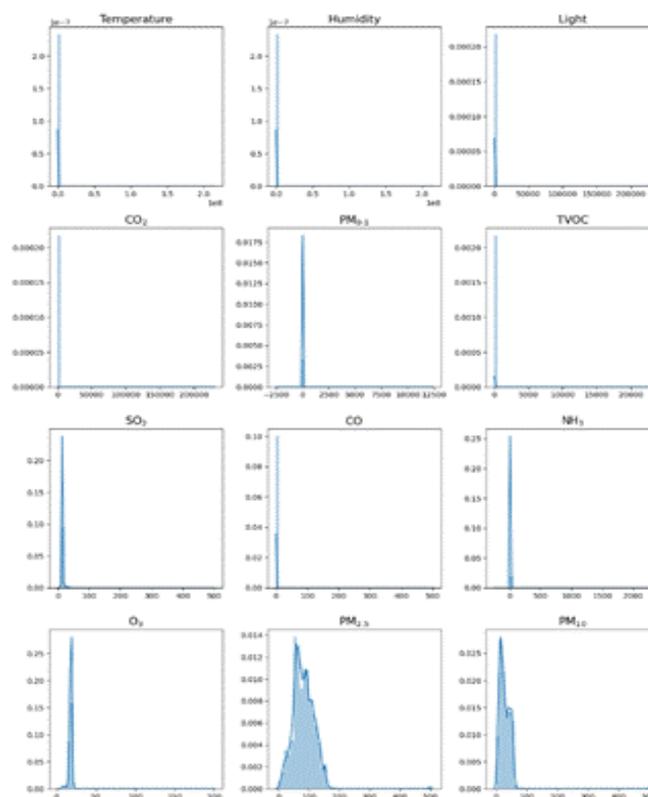


Figure I – The data distribution of indicators in the data set

Next, we removed the other unnecessary indicators by evaluating the correlation level with the PM<sub>2.5</sub> indicator and its value. From

and Figure I and Figure II, it can be inferred that the presence of CO imposed no impact on the prediction, as its all values are equal to zero. Besides, the indicators TVOC, SO<sub>2</sub>, NH<sub>3</sub>, O<sub>3</sub> were also omitted because they showed no correlation with PM<sub>2.5</sub> indicator as shown in Figure 2.

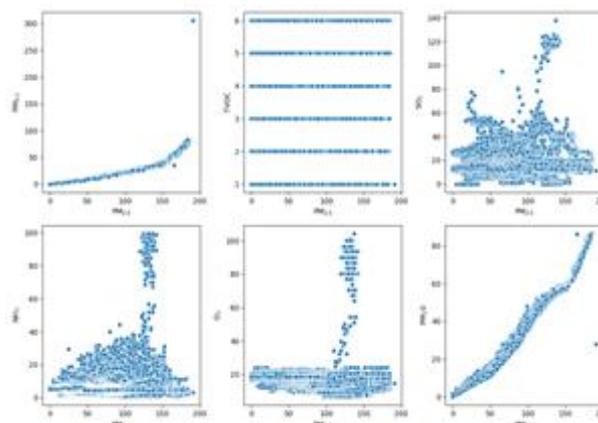


Figure 2 – Correlation chart of pollution indicators with PM<sub>2.5</sub> indicator

**Table 1 – Value description of CO, NH<sub>3</sub>, O<sub>3</sub>, PM<sub>10</sub> indicators**

	CO	NH <sub>3</sub>	O <sub>3</sub>	PM <sub>10</sub>
Record No.	329455	329455	329455	329455
Average	0.000	5.052	17.991	28.401
Standard deviation	0.000	2.374	2.456	15.641
Global minimum	0	0.5	6.5	0
25%	0	4.6	17.6	15.7
50%	0	5	18.5	25
75%	0	5.3	18.5	41.7
Global maximum	0	99.6	104.5	86.6

**Table 2 – Value description of PM<sub>0.1</sub>, TVOC, SO<sub>2</sub> indicators**

	PM <sub>0.1</sub>	TVOC	SO <sub>2</sub>
Record No.	329455	329455	329455
Average	17.927	3.535	14.405
Standard bias	8.972	1.605	4.604
Global minimum	0	1	0
25%	11	2	12.9
50%	17	4	14.3
75%	24	5	14.3
Global maximum	306	6	137.9

It can be seen that PM<sub>10</sub> và PM<sub>2.5</sub> indicators have the strongest relationship among the above indicators, so they are kept. Finally, the indicators needed for prediction are temperature, humidity, light, CO<sub>2</sub>, PM<sub>10</sub> and distribution value are described in **Error! Reference source not found.** In the next part, we present the method to extract feature from the remaining indicators after data pre-processing.

**B. Feature extraction method**

The features we extracted are based on the selection of features of the studies we have investigated previously, among which, SF and GA methods are applied to detect the best from the initial set of features. In order to solve PM<sub>2.5</sub> prediction problem in Hanoi, we took the potential features from the results of the study in Warsaw city [7]. Those recommended features include:[7] [8]

- Current features:  $f_1$  - current PM<sub>2.5</sub> indicator value;  $f_2$  - current PM<sub>10</sub> indicator value;  $f_3$  - present temperature value;  $f_4$  - current moisture value;  $f_5$  - current light value;  $f_6$  - current CO<sub>2</sub> indicator value. These are the values that describes the current air quality and aims at supporting prediction in the next hour.

- Time-based features:  $f_{7-8}$  - season (2-bit representation: 00 - spring, 01 - summer, 10 - autumn, 11 - winter);  $f_9$  - day off (1 - days off, 0 – working days);  $f_{10}$  - hour. Seasonal features are essential because the climate in Hanoi is tropical monsoon; as a result, there are four seasons in a year despite Hanoi’s tropical location. In addition, the day off and daytime features are also considered because air pollution is mainly triggered by human activities.

- Features of the previous 24 hours:  $f_{11-35}$  - PM<sub>2.5</sub> indicator values for the previous 1 to 24 hours. These features provide hourly variation tracking to predict the next hour.

- Meteorological features in the previous 24 hours:  $f_{36-38}$  - maximum, minimum and average PM<sub>2.5</sub> indicator values in the previous 24 hours;  $f_{39-41}$  - maximum, minimum and average temperature values in the previous 24 hours;  $f_{42-44}$  - maximum, minimum and average values of humidity in the previous 24 hours. These features aim at showing the degree of environmental volatility within 24 hours, which affects the change in the PM<sub>2.5</sub> indicator in the next hour.

The predicted value is the average value of the PM<sub>2.5</sub> indicator in the next hour. After feature extraction process, we executed data normalization by z-score normalization formula:

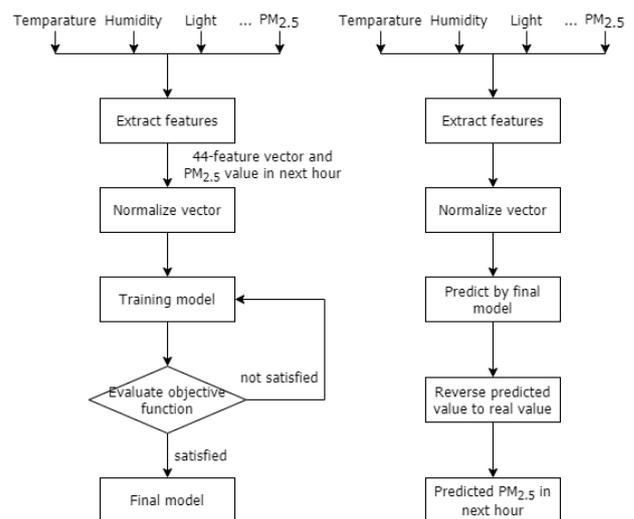
$$z = \frac{x - \mu}{\sigma} \tag{1}$$

$\mu$  of which is the average of the elements,  $\sigma$  is the standard deviation,  $x$  is the value to be normalized.

**C. Predictive model description**

The predictive model we proposed includes the training process and the predictive process shown in *Figure 3*.

In the training process, from the input data which are meteorological indicators and pollution indicators, we extracted the 44-dimensional feature vector as described in the previous section. This vector was normalized, and the algorithm we applied was XGBoost built on Gradient Boost [11]. This algorithm has won numerous contests on Kaggle (the Machine Learning and Data Science community that regularly hosts competitions in the field) in recent years. Unlike RF, this algorithm uses boosting method to solve the problem. More specifically, new trees are generated sequentially with the aim of minimizing the error from the previous tree by partially re-learning the error from the previous tree, updating the bug to get a better tree. From that on, in the previous step, the wrongly assigned points will have more chances of being corrected in the future [12].



**Figure 3 – Predictive model**

The data set consists of  $(x_i, y_i)$  pairs of which are the 44-dimensional feature vector and the corresponding predictive value. Learning model is described as follow:

$$\hat{y}_i = \phi(\mathbf{x}_i) = \sum_{k=1}^K f_k(\mathbf{x}_i), f_k \in \mathcal{F} \quad (2)$$

In this formula,  $\mathcal{F} = \{f(\mathbf{x}) = w_{q(x)}\}(q: \mathbb{R}^m) \rightarrow T, w \in \mathbb{R}^T$ ,  $q$  is the tree to map the vector to the predictive value at leaf node,  $T$  is the number of leaf node,  $K$  is the number of tree,  $f_k$  is the independent  $k^{\text{th}}$  tree in the model,  $w_i$  is the weight of  $i^{\text{th}}$  leaf node and  $\hat{y}_i$  is the predictive value of the target function:

$$\mathcal{L}(\phi) = \sum_{i=1}^n l(y_i, \hat{y}_i) + \sum_k \Omega(f_k) \quad (3)$$

in which,  $n$  is the number of data point,  $\Omega(f) = \gamma T + \frac{1}{2} \lambda \|w\|^2$  is the regularization. As the target function can not be optimized by Stochastic Gradient Descent (SGD) method, learning process is described as follow:

With  $\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(\mathbf{x}_i)$  and  $\hat{y}_i^{(0)} = 0, \hat{y}_i^{(t)}$  as the predictive value of  $i^{\text{th}}$  instance at  $t^{\text{th}}$  iteration, the current target function is:

$$\mathcal{L}^{(t)} = \sum_{i=1}^n l(y_i, \hat{y}_i^{(t-1)} + f_t(\mathbf{x}_i)) + \Omega(f_t) \quad (4)$$

Approximation formula is:

$$\mathcal{L}^{(t)} \approx \sum_{i=1}^n \left[ l(y_i, \hat{y}_i^{(t-1)}) + g_i f_t(\mathbf{x}_i) + \frac{1}{2} h_i f_t^2(\mathbf{x}_i) \right] + \Omega(f_t) \quad (5)$$

With  $g_i = \partial_{\hat{y}_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)})$ ,  $h_i = \partial_{\hat{y}_i^{(t-1)}}^2 l(y_i, \hat{y}_i^{(t-1)})$ , if the constant is omitted, the target function can be simplified as below:

$$\tilde{\mathcal{L}}^{(t)} = \sum_{i=1}^n \left[ g_i f_t(\mathbf{x}_i) + \frac{1}{2} h_i f_t^2(\mathbf{x}_i) \right] + \Omega(f_t) \quad (6)$$

Set  $G_j = \sum_{i \in I_j} g_i, H_j = \sum_{i \in I_j} h_i$ , với  $I_j = \{i | q(\mathbf{x}_i) = j\}$  to be the value set at leaf node  $j$ .

Optimized weight at each leaf node:

$$w_j^* = -\frac{G_j}{H_j + \lambda} \quad (7)$$

Error function in the entire tree:

$$\tilde{\mathcal{L}}^{(t)} = -\frac{1}{2} \sum_{j=1}^T \frac{G_j^2}{H_j + \lambda} + \gamma T \quad (8)$$

The training process ends after some iterations or when the target function value is less than a certain threshold. After the training process, the model is used to predict the mean value of the PM<sub>2.5</sub> indicator over the next hour. With the data on meteorological and pollution indicators within 24 hours as input, the data was extracted into a 44-dimensional vector and then normalized. This vector is fed into the training model to provide a predictive value. In the next section, we test the extraction method and the predictive model that has been presented.

#### IV. EXPERIMENTS AND RESULTS

Since the sampling period is about 40 seconds, therefore, in order to conduct the test, they averaged those records by the hour. The results obtained were 6433 records of the hourly atmospheric readings. Next, we performed the preprocessing, extraction and normalization of this data. To perform the training and evaluation process, the records were taken

randomly and divided into 2 sets: the training set accounts for 75% of the initial data and the remaining 25% of the data is the test set.

The measures we used for evaluation include formulas: R<sup>2</sup> - score (9), MAE (10) and RMSE (11) as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (9)$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (10)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (11)$$

In which,  $n$  is the number of elements,  $y_i$  is the real value,  $\hat{y}_i$  is the predicted value, and  $\bar{y}$  is the mean of the number of elements. These metrics are employed because they clearly represent the difference between the real value and the predicted value. This is suitable for regression problems because the predicted value is on the continuous domain instead of labels like the classification problem. In terms of R<sup>2</sup> - score, the higher the value, the stronger the model (which represents relevance to the data set) and preferably 1.00, with MAE and RMSE as small as possible (these two measures represent difference between predicted value and real value)

Next, we compared the results our extraction method presented in part 3 (Method 1) and another extraction method that includes only features extracted from meteorological factors (Method 2) [8]. Specifically, our method takes into account the time of day and year, in companion with the PM<sub>10</sub> indicator and the input data of the indicators in the previous 24 hours. Method 2 only concerned the meteorological factors in the current range. Let's compare the performance results with the measurements presented in Table III. The predicted results of the two methods are shown in Figure 5: the left side of which is the comparison between the real value with the predicted value when method 1 was applied, on the right is that of method 2.

Table 3 – The comparison results between two methods

	R <sup>2</sup> - score	MAE	RMSE
Method 1	0.9508	0.1387	0.2266
Method 2	0.9368	0.1515	0.2521

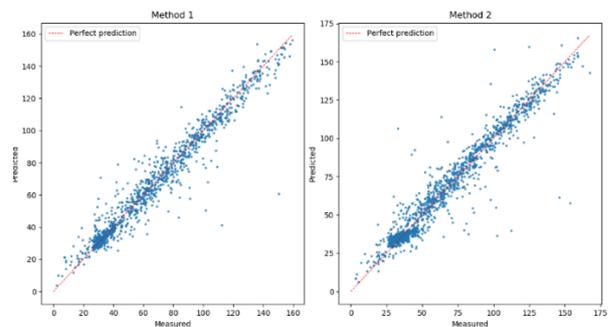


Figure 4 – The predictive results of two methods

It can be seen that our extraction method gives ~ 2% higher results than the old method when we conduct the test on the same model. Therefore, PM<sub>10</sub> indicator and time factor are proved to have an impact on the PM<sub>2.5</sub> prediction in the next hour besides the basic meteorological factors such as temperature, humidity, light.

Next we compared the predictive model with other models: SVM, RF, MLP and XGBoost. The hyper-parameter of each algorithm is shown in Table IV.

**Table I – Hyper-parameter of each algorithm**

	Hyper-parameter
SVM	gamma='auto' kernel='rbf' C=100 epsilon=0.0001
Random Forest	n_estimators=150 max_features='auto'
MLP	hidden_layer_sizes=(192,128,96) max_iter=1000 learning_rate_init=0.01 tol=1e-6 batch_size=192
XGBoost	n_estimators=200 max_depth=8 gamma=0.7 objective='reg:squarederror'

The criteria for similar comparison include the measures: R<sup>2</sup> - score, MAE, RMSE and training time (in second). The results are presented in Table V.

**Table V – Result comparison among algorithms**

	R <sup>2</sup> - score	MAE	RMSE	Training time
SVM	0.9553	0.1154	0.2101	27.0608
Random Forest	0.9587	0.1115	0.2020	35.5577
MLP	0.9562	0.1276	0.2078	8.2011
<b>XGBoost</b>	<b>0.9595</b>	<b>0.1126</b>	<b>0.1999</b>	<b>4.8872</b>

Through the measure R<sup>2</sup> - score, it can be seen that the XGBoost algorithm gave the highest fitting rate with data set (95.95%). Looking at the results of RMSE, we can see that the difference between the predicted value and the actual value is the smallest, or in other words, the accuracy of the prediction is the highest compared to the remaining values. Although the difference in accuracy among the algorithms is not too much, in terms of training time, XGBoost has the shortest. This shows the potential of this model in training and its predictive accuracy over time.

**V. CONCLUSION AND DEVELOPMENT TREND**

By looking at the data we collected in Hanoi including meteorological factors and pollution indicators, we found that the PM<sub>10</sub> indicator in Hanoi is correlated with the PM<sub>2.5</sub> indicator. Since then, along with other surveys, we proposed a new feature extraction method. The new extraction method includes not only current meteorological and pollution factors, but also those of the past (many hours earlier). This provides better prediction because historical values help show the changing trend of the PM<sub>2.5</sub> indicator in the next hour. In addition, time also plays a role in influencing the predicted results due to climate change, seasonal environmental changes in the year in Hanoi and different

human activities in different time framework in a day or in a week. The experiments proved that our extraction method gave better results of PM<sub>2.5</sub> prediction in Hanoi than the old method, which only concerned the meteorological factors.

The study also shows that the XGBoost algorithm is a good algorithm with high accuracy and short training time compared to other machine learning algorithms. Turning to our problem, this algorithm is suitable because of its accurate prediction ability and low model training cost. However, this algorithm's nature of trying to best match the data makes it susceptible to overfitting. So in the future, we will look at some methods to limit overfitting and do experiments on other deep learning algorithms to predict time series data problems.

In terms of the current data, we also lack some meteorological factors such as wind direction, wind speed. These are also factors that can affect the prediction of air pollution because wind can diffuse or concentrate dust density in an area. With the climate in Hanoi, wind also has different seasonal features such as wind direction, speed, and humidity. In addition, traffic data also needs to be considered due to the large number of personal vehicles in Hanoi. In the future, we will collect more data to observe the correlation between them and the level of air pollution in Hanoi and experiment with another model to improve the accuracy and predictive range of space and time.

**REFERENCES**

1. R. Yu, Y. Yang, L. Yang and G. Han, "RAQ-A Random Forest Approach for Predicting Air Quality in Urban Sensing Systems," Sensors, vol. 16, p. 86, 11 January 2016.
2. WHO, "Air pollution," 2 May 2018. [Online]. Available: [https://www.who.int/en/news-room/fact-sheets/detail/ambient-\(outdoor\)-air-quality-and-health](https://www.who.int/en/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health).
3. C.-M. Vong, W.-F. Ip, P.-k. Wong and J.-y. Yang, "Short-Term Prediction of Air Pollution in Macau Using Support Vector Machines," Journal of Control Science and Engineering, vol. 2012, 2012.
4. K. Siwek and S. Osowski, "DATA MINING METHODS FOR PREDICTION OF AIR POLLUTION," Int. J. Appl. Math. Comput. Sci, vol. 26, 2016.
5. K. Polat and S. S. Durduran, "Usage of output-dependent data scaling in modeling and prediction of air pollution daily concentration values (PM10) in the city of Konya," Neural Computing and Applications, p. 21, 2011.
6. A. Nebot and F. Mugica, "Small-particle pollution modeling using fuzzy approaches," Advances in Intelligent Systems and Computing, pp. 239-252, 2014.
7. NandigalaVenkatAnurag, YagnavalkBurra and S.Sharanya, "Air Quality Index Prediction with Meteorological Data Using Feature Based Weighted Xgboost," International Journal of Recent Technology and Engineering (IJRTE), vol. 8, no. 1, pp. 1355-1358, May 2019.
8. A. Li và X. Xu, "A New PM2.5 Air Pollution Forecasting Model Based on Data Mining and BP Neural Network Model," Advances in Computer Science Rese, tập 65, 2018.
9. M. Z. Joharestani, C. Cao, X. Ni, B. Bashir and S. Talebiesfandarani, "PM2.5 Prediction Based on Random Forest, XGBoost, and Deep Learning Using Multisource Remote Sensing Data," Atmosphere, 2019.
10. W.-F. Ip, C.-M. Vong, J. Y. Yang and P. K. Wong, "Least squares support vector prediction for daily atmospheric pollutant level," Proc. 2010 IEEE/ACIS 9th International Conference on Computer and Information Science (ICIS), pp. 23-28, August 2010.
11. T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," 2016.
12. L. Breiman, "Random Forests," Machine Learning, vol. 45, pp. 5-32, 2001.



## AUTHOR PROFILE



**Hung Thuan Nguyen**, graduated B.Sc in Information System at Posts and Telecommunications Institute of Technology, Hanoi, Vietnam in 2020, main research in data warehouse, mining time-series data and detect fraud in logistic industry.



**Chi Quynh Nguyen**, graduated B.Sc in Information Technology at Hanoi University of Technology, Vietnam in 1999, M.Sc and Ph.D Candidate in Computer Science at University of California, Davis, USQ in 2004 and 2006. Her main research focuses on datawarehousing, data mining and bioinformatics.