Leveraging Big Data for $PM_{2.5}$ Prediction: A Case Study in Selangor, Malaysia

En Xin Neo[1], Khairunnisa Hasikin[1*], Khin Wee Lai[1], Mohd Istajib Mokhtar[2*],
Muhammad Mokhzaini Azizan[3], Sarah Abdul Razak[4], Hanee Farzana Hizaddin[5]

[1]Department of Biomedical Engineering, Faculty of Engineering, Universiti Malaya, Malaysia
[2]Department of Science & Technology Studies, Faculty of Science, Universiti Malaya, Malaysia
[3]Department of Electrical and Electronix Engineering, Faculty of Engineering and Built Environment, Universiti Sains Islam Malaysia, Malaysia
[4]Institute of Biological Sciences, Faculty of Science, Universiti Malaya, Malaysia
[5]Department of Chemical Engineering, Faculty of Engineering, Universiti Malaya, Malaysia

[*]Corresponding Author

doi: https://doi.org/10.21467/proceedings.141.16

ABSTRACT

Air pollution has become a serious issue and has continually increased since the half-decade ago due to globalization. Activities such as urbanization, industrialization, power plants, agricultural open burning and natural disasters such as wildfires are the key factors in air pollution. The air pollutants produced include particulate matter (PM_{10} and $PM_{2.5}$), ozone (O_3), carbon monoxide (CO), sulfur dioxide(SO_2), nitrogen dioxide (NO_2) and heavy metals such as lead (Pb) and cadmium (Cd). According to the most recent revision of the Global Burden of Diseases (GBD), PM_{10} and $PM_{2.5}$ were listed as the fourth most common killer out of 85 risk factors. Hence, it is important to assess air pollution, especially the particulate matter concentration in the air. In this study, we emphasize the development of $PM_{2.5}$ prediction models using machine learning for air pollution evaluation in Selangor, Malaysia. This is because Selangor contributed most pollutants due to its highest population distribution in the country. The machine learning models involved are Random Forest, Naïve Bayes, KNN, SVM, and Gradient Boosting. Gradient boosting and Random Forest contributed comparable prediction results. However, gradient boosting was chosen as the best model for the prediction in this study due to the accuracy and precision in predicting the Classes of $PM_{2.5}$ without misclassification. The accuracy, precision, and recall of the model are 99.9% and 99.94% for F1 score respectively.

Keywords: Air Pollution, Machine Learning, $PM_{2.5}$ prediction

1 Introduction

Since the half-decade ago, pollution is continually increasing due to globalization. Activities such as urbanization, industrialization, power plants, agricultural open burning, and natural disasters such as volcanic eruptions and wildfires are the major factors in air contamination [1]. Nowadays, air pollution has become a huge problem for the planet, and it is also a major cause of death. According to the World Health Organization (WHO), it is estimated that air pollution worldwide caused around 6.9 million deaths. Pollutant markers that often contributed to air pollution are particulate matter (PM_{10} and $PM_{2.5}$), ozone (O_3), carbon monoxide (CO), sulphur dioxide(SO_2), nitrogen dioxides (NO_2) and heavy metals such as lead (Pb) and cadmium (Cd) [2]. Contemporary environmental issues such as global warming, acid rain, reduces visibility, and climate change are contributed by poor air conditions [3]. Increasing air pollutants concentration in the air lead to a multitude of medical conditions in human, from bronchitis to heart diseases, from pneumonia to lung cancer, as well as developmental defects during pregnancy and premature deaths. Children and the elderly are often the most vulnerable group to be affected by air pollution.

Proceedings DOI: 10.21467/proceedings.141; Series: AIJR Proceedings; ISSN: 2582-3922; ISBN: 978-81-957605-4-1

To indicate the quality of air, Air Quality Index (AQI) is implemented by showing the current or forecasted pollution level of the air in the areas. The air quality indices vary by country and air quality standards [4]. According to Liang, et al. [4], US Environmental Protection Agency (US EPA) monitors pollutants such as O_3, PM_{10}, $PM_{2.5}$, NO_2, SO_2, and Pb at more than 4000 locations across the country in the United States. However, in Malaysia, the Department of Environment (DoE) monitors pollutants such as CO, O_3, NO_2, SO_2, and PM_{10}, $PM_{2.5}$ throughout the whole country. There are 6 levels of AQI in Malaysia, as the AQI increases, a higher percentage of the population is exposed to the pollution as the air is more polluted.

Air quality evaluation is crucial in monitoring and controlling air pollution. To predict and evaluate air quality, many researchers have employed machine learning algorithms to predict the air quality index (AQI) in determining pollution levels. Recently, machine learning-based prediction techniques are widely used and becoming more and more common as big data and artificial intelligence are evolving. Several prediction studies in various fields involving big data and machine learning techniques have been carried out by Jamaludin, et al. [5], Wong, et al. [6], Zamzam, et al. [7]. In addition, there are also several researches that place a greater emphasis on advanced statistical learning algorithms for assessing air quality and forecasting air pollution. Authors such as Veljanovska and Dimoski [8], Yi, et al. [9], Yu, et al. [10] are performing AQI prediction using machine learning techniques such as random forest, and neural networks. From the studies, their artificial neural networks outperform all the algorithms such as KNN and decision tree with an accuracy of 92.3%. The discussion above shows the feasibility and reliability of machine learning in the prediction of air quality.

In this paper, we emphasize the development of $PM_{2.5}$ prediction models using machine learning for air pollution evaluation in Selangor, Malaysia. This is crucial as Selangor has the highest population distribution in Malaysia where it contributed most to pollutants formation. Machine learning techniques algorithms involved are random forest, Naïve Bayes, K-Nearest Neighbour (KNN), support vector machine (SVM), and gradient boosting. In addition, we observe the models' performances by evaluating the matrices such as accuracy, precision, recall, and F1 Score.

2 Materials and Methods

2.1 Study area

This study is conducted in Selangor, Malaysia. The retrospective data are collected from the Department of Environment (DoE) for 4 air monitoring stations in Selangor. The locations of the monitoring station include Petaling Jaya, Shah Alam, Klang and Bantingas shown in Figure 1. The data are accessed from January 2010 to December 2016.

Among 13 states and 3 federal territories, Selangor was chosen due to the highest population distribution in Malaysia. Besides, economic activities such as industrialization, construction and heavy traffic are the factors in choosing Selangor as the site for the research. These activities cause the high emission of pollutants such as CO, $PM_{2.5}$, PM_{10}, O_3, and SO_2 which worsen the air quality and polluted the ambient air in the city. Consequently, it contributed to more medical consequences due to air pollution.

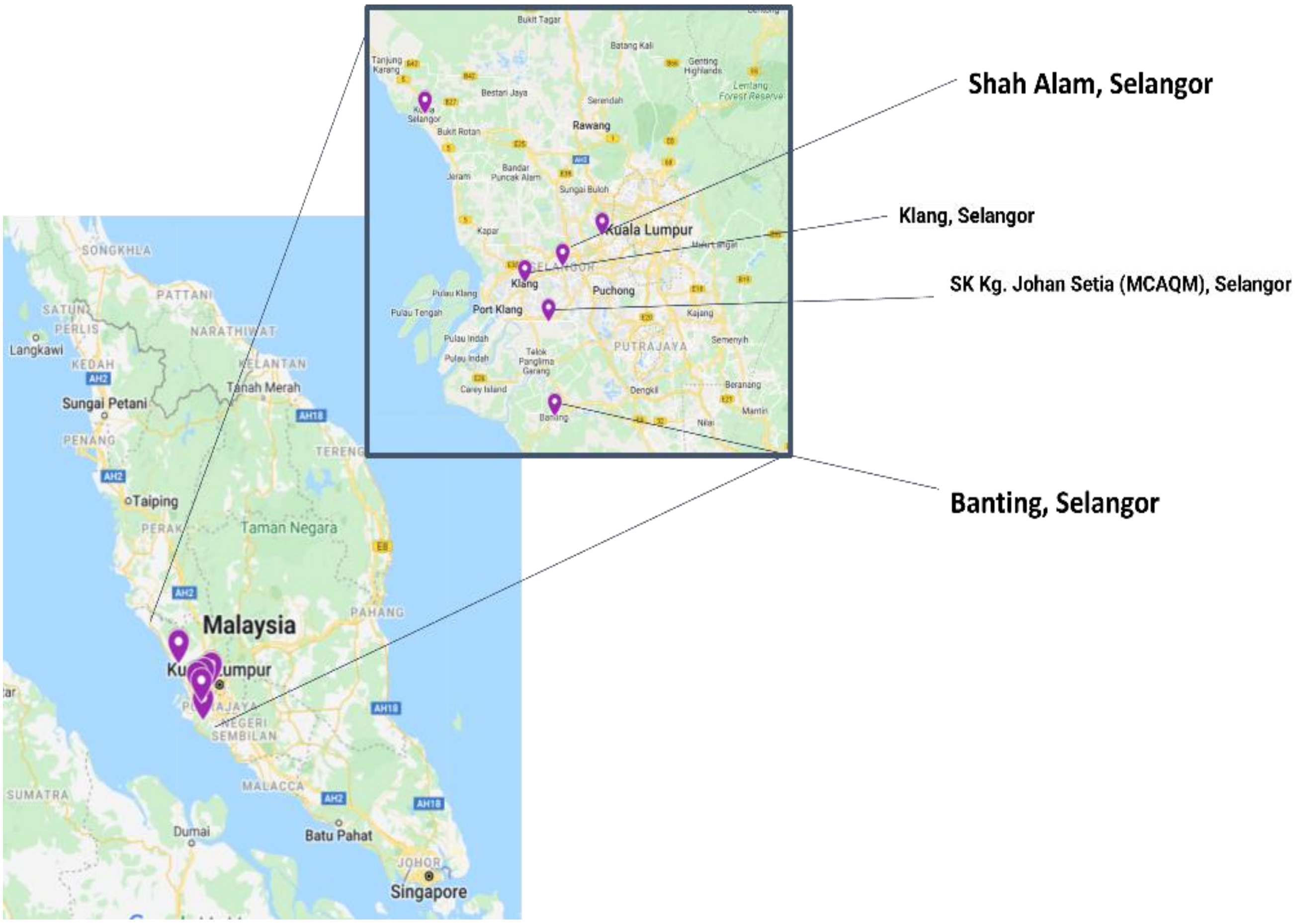


Figure 1: *Air monitoring stations in Selangor*

2.2 Data cleaning and preparation

The data from January 2010 to December 2016 on an hourly basis is collected from DoE. The data consists of environmental pollution markers and meteorological data. The environmental data include PM_{10}, $PM_{2.5}$, SO_2, NO_2, O_3, and CO while meteorological data includes wind speed, wind direction, temperature, and humidity. In this study, $PM_{2.5}$ is labelled as the targeted parameter in the prediction of $PM_{2.5}$.

Feature selection is completed by analyzing the raw data set. The features selected are environmental pollution markers and meteorological data, $PM_{2.5}$ is labelled as the output parameter. The undesired features such as time, date, and location ID were removed to minimize redundancy and maximize relevance. In data labelling, both environmental and meteorological data are classified. Environmental data are categorized according to the Air Pollution Index (API) sub-index recommended by DoE accordingly, meanwhile, in meteorological data, the data are being classified statistically by quartile as there are no certain guidelines in classifying the meteorological data. Therefore, in this study, there are 10 input parameters, namely i) O_3, ii) CO, iii) SO_2, iv) NO_2, v) PM_{10}, vi) wind speed, vii) wind direction, viii) temperature and ix) humidity and 1 targeted response which is $PM_{2.5}$.

In this study, we focused on the pollutant $PM_{2.5}$. Table 1 shows the classification of $PM_{2.5}$ recommended by the DoE.

Table 1: *$PM_{2.5}$ Concentration recommendations by Department of Environment (DoE), Malaysia [11]*

$PM_{2.5}$ Concentration Breakpoint [ug/m^3]	Air Pollution Index (API)	API Status	API Description
0-12	0-50	Good	There is little pollution and there are no harmful health effects.
12.1-35.4	51-100	Moderate	No harmful effects on health.
35.5-55.4	101-150	Unhealthy	Sensitive to folks and should avoid. Health conditions for the elderly, pregnant women, children, and persons with heart and lung issues deteriorate
55.5-150.4	151-200		
150.5-250.4	201-300	Very unhealthy	Unhealthy for the public. Worsening health and a reduced tolerance for the activity might lead to lung and heart issues.
250.5-350.4	301-400	Hazardous	Emergency
350.5-500.4	401-500		

3 Theory and Calculation

3.1 Machine Learning Techniques

3.1.1 Support Vector Machine (SVM)

Support Vector Machine (SVM) is a supervised machine learning algorithm which is commonly used in classification, regression, and other purposes such as detection of an outlier. In SVM, a boundary is constructed which is known as the hyperplane, which separates the distinct data points and subsequently deduces the output [12]. The hyperplane is important as it determines the number of classes of the data and predicts the output based on the similarity of the new data holds. Meanwhile, in regression, the hyperplane is constructed at the maximum margin with linear regression, with an additional parameter ε-an insensitive loss which is crucial for deviation toleration that lies inside the region of ε. In this study, we focused on the classification of SVM [13].

3.1.2 Random Forest

Random forest, another supervised learning algorithm is also implemented in this study. The random forest has a comparable trait to SVM which it can handle classification and regression cases. Random forest combines different decision trees to create a forest and bagging idea and introduces unpredictability into the model construction. The individual tree is divided using a random selection of features, and each decision tree's training data subset is made using a random selection of instances. The variable from the random number of features is taken into consideration for the optimal split at each decision node in every tree. In this algorithm, every test data point is run through the forest's decision trees in order to make a forecast where the result of the prediction is made based on the results of votes among the trees. Hence, a stronger and more reliable single learner is formed.

3.1.3 Gradient Boosting

Gradient boosting is known as an algorithm that creates a stronger prediction ability by combining weak learning models together. In gradient boosting, decision trees are commonly implemented when comes to gradient boosting. In constructing the gradient boosting, a custom loss function is used, and standardised loss functions are supported by a gradient boosting classifier. In this case, the loss function in the algorithm has to be differentiable. This is important as gradient boosting performs prediction by identifying shortcoming using high-weight data points using gradients in the loss function which measure the performance of models' coefficient fitting underlying data.

3.1.4 Naïve Bayes

Naïve Bayes, a supervised learning algorithm which applies Bayes's statistical theorem with a "naïve" assumption of conditional independence. A Bayesian statistical classifier is built on the Bayesian hypothesis where when it comes to categorising the unknown tuple X, it was assumed that the classifier would conditionally predict the element vector belongs to the class with the highest posterior probability. This algorithm places the unknown tuple X in class Ci when the $P(C_iX)>P(C_jX)$. The description of Bayes Theorem is described in the formula below.

$$P(y_j|x_i) = \frac{P(x_i|y_j)P(y_j)}{P(x_i)} \tag{1}$$

$P(x_i)$= the prior probability of class

$P(y_i|x_i)$= the posterior probability of class given predictor

$P(x_i|y_i)$= the probability of predictor given class

3.1.5 K-Nearest Neighbour (KNN)

Aside from the machine learning algorithm mentioned above, k-nearest neighbour (KNN), a simple and well-known supervised machine algorithm is also implemented in this study. Similar to the algorithms mentioned, KNN can be used in classification and regression problems. KNN gathers data points by distance from the arrival data point, where the distance can be measured in a variety of ways, however, the most recommended way suggested by experts is the Euclidian distance. KNN classifier takes the top K neighbours and applies a simple majority by voting.

3.2 Performance Evaluations

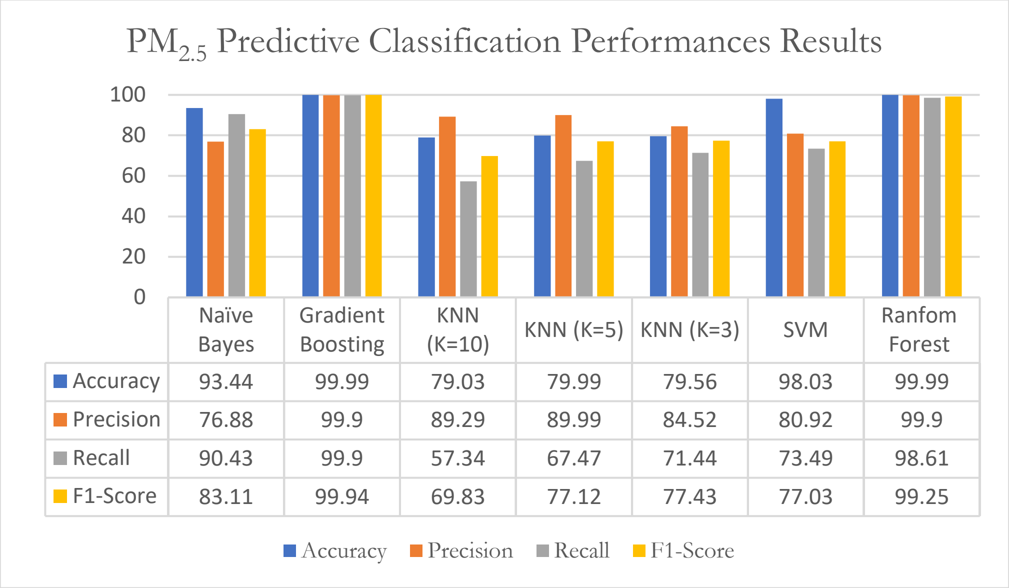
In accessing the performances of the machine learning model, there are several metrics are computed in determining the model outputs. In research, the machine learning algorithms were evaluated through qualitative performance characteristics namely i) accuracy, ii) precision, iii) recall, and iv) F-score. The description of these metrics is discussed in Table 2.

Table 2: *Machine learning performance metrics description*.

Metric	Description
Accuracy	The degree to which quantity is derived by measuring and the real value of measurand agree.
Precision	Degree of independent test results under specified conditions that agree with one another
Recall	The ratio of numbers of true positives to the number of true negatives.
F-score	Measurand of the accuracy of algorithm based on the recall and precision obtained.

4 Results and Discussion

In this research, $PM_{2.5}$ is desired to be predicted in Selangor, Malaysia. In this section, the results of the prediction using various machine learning techniques will be discussed and presented in Figure 2.



$PM_{2.5}$ Predictive Classification Performances Results

	Naïve Bayes	Gradient Boosting	KNN (K=10)	KNN (K=5)	KNN (K=3)	SVM	Ranfom Forest
🟦 Accuracy	93.44	99.99	79.03	79.99	79.56	98.03	99.99
🟧 Precision	76.88	99.9	89.29	89.99	84.52	80.92	99.9
⬜ Recall	90.43	99.9	57.34	67.47	71.44	73.49	98.61
🟨 F1-Score	83.11	99.94	69.83	77.12	77.43	77.03	99.25

Figure 2: *Graph of accuracy, precision, recall and F1-score for $PM_{2.5}$ prediction using different machine learning techniques*

Based on the results presented in Figure 2, Gradient Boosting and Random Forest presented the same accuracy and precision values which is 99.90%. However, Gradient Boosting has the best performance in predicting $PM_{2.5}$. This can be supported by referring to the recall, and F1-score which are 99.90% and 99.4% respectively, relatively higher than Random Forest which is 98.61% and 99.25% respectively. Meanwhile, in k-Nearest Neighbour had the worst performance in $PM_{2.5}$ prediction as compared to Naïve Bayes, Gradient Boosting, SVM, and Random Forest. KNN presented only 79% of accuracy in $PM_{2.5}$. and precision between 85% to 90% for the numbers of neighbours of 3, 5, and 10. However, among these models, KNN with numbers of neighbours of 3 presented an overall better result as compared to K=5 and K=10. The model presented an accuracy of 79.56%, precision of 84.52%, recall of 71.44% and F1 score of 77.43. On the other hand, SVM presented a comparable accuracy in $PM_{2.5}$ prediction with gradient boosting and Random Forest. Unfortunately, the model has a lower recall value which is only 73.49%. As referring to the results, Naïve Bayes presented a higher recall value when compared to SVM which is 90.43%, even though the accuracy is only 93.44%, around 6% less accurate than SVM. Hence, from the comparison, Naïve Bayes can be ranked as a moderate performed model after Gradient Boosting and Random Forest as the accuracy is comparable to SVM and the recall value is higher than SVM.

5 Conclusions

$PM_{2.5}$ prediction is important as a part of the air quality assessment. $PM_{2.5}$ concentration has been classified into 6 classes by DoE, Malaysia. In this study, we identified the supervised machine learning techniques and successfully identified the best model for $PM_{2.5}$ prediction which is the gradient boosting algorithm. It possessed 99.9% accuracy, precision, recall and F1 score. From the confusion matrix, Gradient Boosting also presented an excellent performance in Class 1, Class 4, Class 5, and Class 6 classification prediction with no misclassification. Hence, it can be concluded that gradient boosting is the best model in this study. However, the results of the prediction seem to be overfitting, as this can be explained by the nature of the data, which is an imbalanced distribution of the data classes.

6 Declarations

6.1 Study Limitations

Limitation in this study includes the nature of the data which is imbalanced distributed presented in the dataset. This could be a factor where the result presented is overfitting.

6.2 Acknowledgements

The authors would like to thank Department of Environment, Malaysia for the data provided in this study. This research is funded by Ministry of Higher Education through MRUN Young Research Grant (MY-RGS), MR001-2019, entitled “Climate Change Mitigation: Artificial Intelligence-based Integrated Environmental System for Mangrove Forest Conservation" and Universiti Malaya for the continuous support through Universiti Malaya Research Universiti Grant (SATU) by Universiti Malaya.

6.3 Funding source

Ministry of Higher Education on Malaysian Young Research Grant (MYRGS), MR001-2019.
Universiti Malaya Research Universiti Grant (SATU), ST065-2021.

6.4 Competing Interests

The authors declare there is no conflict of interest in presenting this study.

6.5 Publisher’s Note

AIJR remains neutral with regard to jurisdiction claims in published maps and institutional affiliations.

References

[1] Doreswamy, H. K S, Y. Km, and I. Gad, "Forecasting Air Pollution Particulate Matter (PM2.5) Using Machine Learning Regression Models," *Procedia Computer Science,* vol. 171, pp. 2057-2066, 2020/01/01/ 2020, doi: https://doi.org/10.1016/j.procs.2020.04.221.

[2] E. X. Neo *et al.*, "Towards Integrated Air Pollution Monitoring and Health Impact Assessment Using Federated Learning: A Systematic Review," *Frontiers in Public Health,* Systematic Review vol. 10, 2022. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fpubh.2022.851553.

[3] K. Kumar and B. P. Pande, "Air pollution prediction with machine learning: a case study of Indian cities," *International Journal of Environmental Science and Technology,* 2022/05/15 2022, doi: 10.1007/s13762-022-04241-5.

[4] Y.-C. Liang, Y. Maimury, A. H. Chen, and J. R. Juarez, "Machine Learning-Based Prediction of Air Quality," *Applied Sciences,* vol. 10, no. 24, 2020, doi: 10.3390/app10249151.

[5] M. R. Jamaludin *et al.*, "Machine Learning Application of Transcranial Motor-Evoked Potential to Predict Positive Functional Outcomes of Patients," *Computational Intelligence and Neuroscience,* vol. 2022, p. 2801663, 2022/05/20 2022, doi: 10.1155/2022/2801663.

[6] W. Y. Wong *et al.*, "Water Quality Index Using Modified Random Forest Technique: Assessing Novel Input Features," *Computer Modeling in Engineering \& Sciences,* vol. 132, no. 3, 2022, doi: 10.32604/cmes.2022.019244.

[7] A. H. Zamzam *et al.*, "Prioritisation Assessment and Robust Predictive System for Medical Equipment: A Comprehensive Strategic Maintenance Management," *Frontiers in Public Health,* Original Research vol. 9, 2021. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fpubh.2021.782203.

[8] K. Veljanovska and A. Dimoski, "Air quality index prediction using simple machine learning algorithms," *International Journal of Emerging Trends & Technology in Computer Science (IJETTCS),* vol. 7, no. 1, pp. 025-030, 2018.

[9] X. Yi, J. Zhang, Z. Wang, T. Li, and Y. Zheng, "Deep distributed fusion network for air quality prediction," 2018 2018, pp. 965-973.

[10] R. Yu, Y. Yang, L. Yang, G. Han, and O. A. Move, "RAQ–a random forest approach for predicting air quality in urban sensing systems," *Sensors,* vol. 16, no. 1, p. 86, 2016.

[11] D. o. Environment. "Air Pollution Index (API)." https://www.doe.gov.my/portalv1/en/info-umum/english-air-pollutant-index-api/100 (accessed 26.7.2022.

[12] I. Muhammad and Z. Yan, "SUPERVISED MACHINE LEARNING APPROACHES: A SURVEY," *ICTACT Journal on Soft Computing,* vol. 5, no. 3, 2015.

[13] M. Awad and R. Khanna, "Support vector regression," in *Efficient learning machines*: Springer, 2015, pp. 67-80.