ISSN 2347 - 3983
Volume 8. No. 6, June 2020
International Journal of Emerging Trends in Engineering Research
Available Online at http://www.warse.org/IJETER/static/pdf/file/ijeter55862020.pdf
https://doi.org/10.30534/ijeter/2020/55862020

Agarwood Oil Quality Classification using Support Vector Classifier and Grid Search Cross Validation Hyperparameter Tuning

Mohamad Aqib Haqmi Abas[1], Nurlaila Ismail[1*], Nor AzahMohd Ali[2], SaifulNizam Tajuddin[3], Nooritawati Md. Tahir[1]
[1]Faculty of Electrical Engineering, UniversitiTeknologi MARA (UiTM), Selangor, Malaysia,
*nrk_my@yahoo.com
[2]Herbal Product Development, Natural Product Division, Forest Research Institute Malaysia (FRIM), Kepong, Malaysia, norazah@frim.gov.my
[3]Bioaromatic Research Centre of Excellence, Universiti Malaysia Pahang, Gambang, Pahang, Malaysia, saifulnizam@ump.edu.my

ABSTRACT

This study investigates the performance of using grid search cross validation as a hyperparameter tuning method for support vector classifier in classifying the quality of agarwood oil. The data of agarwood oil sample were obtained from Forest Research Institute Malaysia (FRIM) and Universiti Malaysia Pahang, Malaysia. The chemical compound abundances of agarwood oil sample are used as its input whiles the quality of agarwood oil of high quality and low quality as the output. The parameter used to train the support vector machine classifier by using the grid search cross validation is the parameter C, gamma and kernel. Based on the results of the study, it shows that by using combination of C with value of 1, gamma value of 10 and radial basis function kernel gives the best classification accuracy of 100% and performance measure scores of 1.0.

Key words: agarwood oil, cross validation, grid search, quality.

1. INTRODUCTION

Agarwood oil is a type of essential oil that can be produce from Aquilaria species plant. The agarwood oil has a very high market demand; however, the production has been decreasing in the past few years [1]. In countries where agarwood oil demands are high such as in the Middle East countries, Japan and China, the agarwood oil is widely used as incenses, and in traditional and religious ceremony. To solve the problem of decreasing production, agarwood plant has started to be plant commercially. The grading method used to grade the quality of agarwood oil is different according to country and buyer. The traditional grading method is by hiring human expert to classify the agarwood oil quality. However, studies show that it has a questionable grading accuracy when grading with a large production oil sample. This happens because human nose has sensory limitation as it will get fatigue when used to smell for a long period and high volume of sample [2-3]. In the past few years, few studies have been carried out to classify the agarwood oil using machine learning classifier model based on its chemical compound rather than using the colour and odour properties of the oil itself [4-6].

Support Vector Machine (SVM) is one of supervised machine learning model that can be used as classification and regression problems [7-8]. During training phase, SVM will learn on each training data and uses the training data point that is on the border between classes to construct and represent for the decision boundary with largest margin. This is because the basic idea of SVM is that the only data points that matters to construct for the decision boundary are the ones that lie close to the boundary.

SVM usually is used for linear however by using kernel tricks it can be used as a nonlinear method. Kernelized support vector is widely used nowadays as a lot of dataset used in real world are not even close to being linearly separable[9]. Some of the kernel tricks available to be used in support vector classifier are the polynomials, radial basis function (RBF) and sigmoid (saturating) [9-10]. In machine learning, grid search is one of the techniques used to tune the hyperparameter to find the best possible parameters for machine learning model[9,11]. The grid-search method uses brute-force searching technique where it will use all possible combinations of the parameter values of interest and try to find the combination of those parameters that gives the best result scores. Cross-validation is one of data splitting method used to evaluate generalization performance of model. In cross-validation data splitting technique, the data is being split in multiple fold and a model will be trained for each fold available. Thus, allowing the dataset to be fully utilized and used effectively by the model for training and testing. Grid search and cross-validation technique can be used together in finding the best parameter for machine learning model with a much more stable and accurate prediction [12]. This paper proposes on using the support vector classifier to classify the quality of agarwood oil and using grid search cross-validation technique to find the optimal parameters for the classifier model to achieve the best performance results.

2. METHOD

This experiment uses anaconda software with python programming language to build, train and evaluate the machine learning model. The agarwood oil dataset used in this experiment was obtained from Forest Research Institute Malaysia (FRIM), Kepong, Selangor, Malaysia and Faculty of Industrial Sciences & Technology, Universiti Malaysia Pahang (UMP). The data is based on the chemical compound abundances of agarwood oil samples. The quality of the agarwood oil used as an output for this experiment is high and low quality.

The first step is of the experiment is data acquisition. The dataset used in the experiment consist of 156 agarwood oil sample with 7 input features (independent variables) and 1 target feature (dependent variable). The input features are represented by the chemical compounds of agarwood oil which are β-agarofuran, α-agarofuran, 10-epi-γ-eudesmol, γ-eudesmol, longifolol, hexadecanol and eudesmol respectively. The target feature used represents the quality of agarwood oil sample which are ‘high quality’ and ‘low quality’. There are equal number of high and low quality sample which is 78 samples of ‘high quality’ and 78 samples of ‘low quality’ agarwood oil. The dataset is balanced with 50% class 0 (low quality) and 50% class 1 (high quality).

Next, the data acquired will undergo pre-processing phase. In this experiment, min-max scaling technique is used as the feature scaling technique. Min-max scaling is used to scale the continuous variable for input features in the dataset to range of 0 to 1. Feature scaling is crucial for data that will be used to train support vector classifier.

After data pre-processing phase, the data will be used to train the support vector classifier with grid search cross-validation technique. The parameters value grid of support vector used is C, gamma and kernel. Parameter C value used to test is 0.01, 0.1, 1, 10 and 100. Lower value of C tends to make the classifier model have a larger gap of margin but more margin violations while higher value of C will cause the model to have smaller gap of margin but lower margin violations. In other words, if the value of C too high it will make the model to overfit and not able to generalize well.
Parameter gamma values used in the experiment is 0.01, 0.1, 1, 10 and 100. The parameter gamma used acts like a regularization hyperparameter for kernelized support vector. If the value of gamma parameter is too high, it will make the model overfit and if the value is too low, the model will underfit. Parameter kernel represent the kernel trick that will be used for the support vector classifier model whether linear (non-kernel), polynomial, radial basis function or sigmoid. The formula for the kernel tricks used in support vector classifiers are [13] as in Eqn. (1) to (4);

$$Linear: K(a,b) = a^T.b \tag{1}$$

$$Polynomial: K(a,b) = (\gamma a^T.b + r)^d \tag{2}$$

$$Gaussian\ RBF: K(a,b) = \exp(-\gamma||a-b||^2) \tag{3}$$

$$Sigmoid: K(a,b) = \tanh(\gamma a^T.b + r) \tag{4}$$

The general idea of using grid search is to determine the best parameter based on the given parameter grid value to build the support vector classifier model. With a proper tuning of parameter, a good balanced and well generalized model can be obtained. The number of cross-validation fold used for the grid search is 5. Overall, there are 100 possible combinations of parameter values to be searched for. After the optimal parameters for support vector classifier has been identified, a new support vector classifier model will be train using hold-out test set with ratio of 80% training set and 20% testing set to match with the cross-validation split of 5. This process is to evaluate and analyze the model performance by using the optimal parameters found from the grid search result earlier to a new support vector classifier model.

Next, the support vector classifier model built will be tested, evaluated and analyzed in the model evaluation process. The performance measures used to evaluate the model performance are classification accuracy, confusion-matrix based performance measure and precision, recall and F1 measure score. The precision, recall and F1 measure score is very useful as it would give more accurate measurement information on the support vector classifier model built than only using the classification accuracy. The formula for precision and recall are can be found in [9,14] Precision measures number of samples that is predicted positive by the classifier model are actually positive. Recall measures the numbers of samples that are actually positive has successfully been predicted as positive by the model. The formula of precision and recall are given in [14]. After the evaluation phase of using the parameter values found using grid search, a new support vector classifier model was built using the varying values for parameter C and gamma to analyze and evaluate the effect when using different values of parameters. During this analysis phase, the other parameter will be kept at constant value [15, 16].

The last step of the experiment is to assess whether the model is viable based on the results given in the previous evaluation phase. If any error or problems arises with the score of performance measures, the support vector classifier model will be rebuilt with the data, and both the data and model will be checked for the errors or problems [15, 16].

3. RESULTS AND DISCUSSION

The Table 1 tabulates the result of the data after undergoing min-max scaling technique during the data pre-processing phase. The value of each input feature (chemical compound abundances of agarwood oil) has been scaled to range of 0 to 1.

Table 2 tabulates the results of grid search cross-validation method on the agarwood oil dataset for support vector classifier after running through 100 possible combinations of parameter grid values. From the result obtained the best and optimal combination parameter value is when using C = 1, gamma =10 and kernel = RBF.
Figure 1(a), Figure 1(b) and Figure1(c) shows the heatmap for classification accuracy of the model on training set. To obtain the value in percentage (%), multiply the values to 100. From Figure 1(a) and Figure 1(b), it shows that only the polynomial and rbf kernel gives a 100% classification accuracy on the

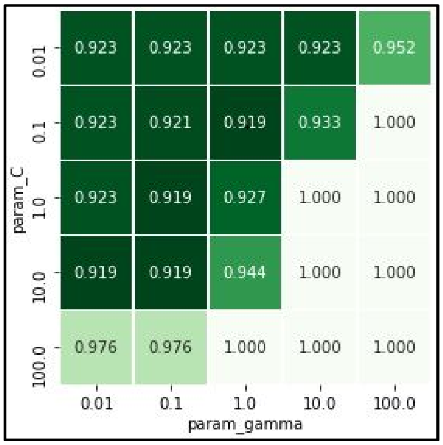
train set data. From the figures, there are other combination between C, gamma and kernel that produce 100% classification accuracy, however, the grid search method works sequentially. To get the best parameter from the grid search method, the model will then be tested on the testing set data and the highest classification accuracy of both training set and testing set sequentially will be used.

Table 1: Per-feature min and max value information before and after scaling

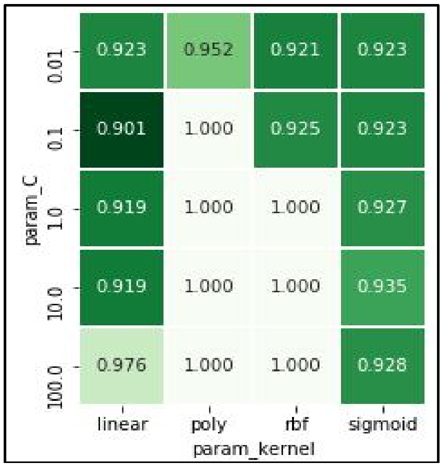
Chemical compound	Per-feature (before)		Per-feature (after)	
	Min	Max	Min	Max
β-agarofuran	0.00	7.09	0.00	1.00
α-agarofuran	0.00	3.56	0.00	1.00
10-epi-γ-eudesmol	0.00	21.47	0.00	1.00
γ-eudesmol	0.00	15.78	0.00	1.00
longifolol	0.00	16.77	0.00	1.00
hexadecanol	0.00	3.48	0.00	1.00
eudesmol	0.00	10.75	0.00	1.00

Table 2: Result of grid search method.

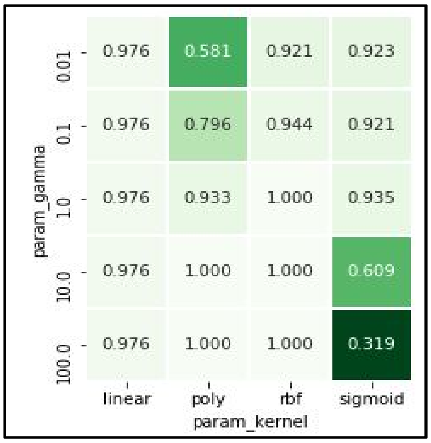
Parameter grid	Optimal value
C	1
gamma	10
kernel	RBF



(a)



(b)



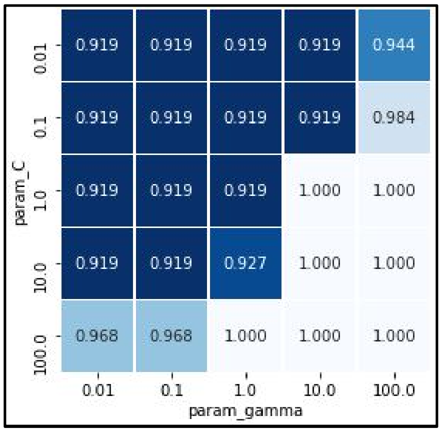
(c)

Figure 1: Heatmap of classification accuracy on training set for different values of parameter.

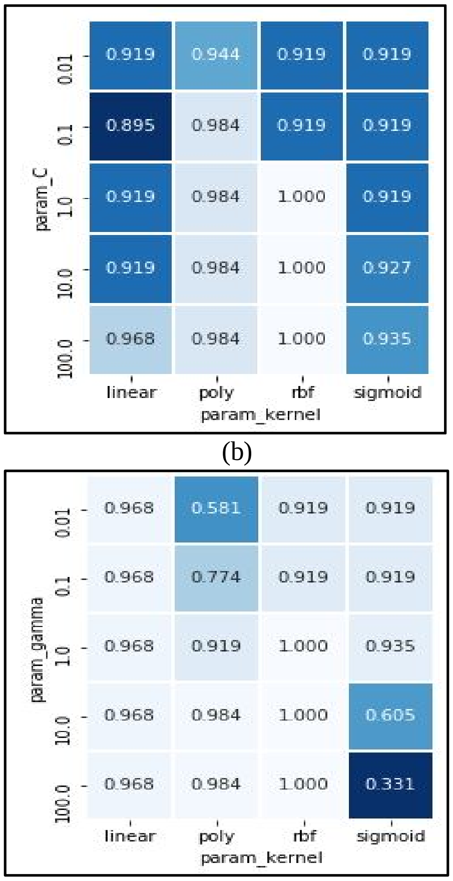
Figure 2(a), Figure 2(b) and Figure 2(c) shows the heatmap for classification accuracy of the model on testing set. To obtain the value in percentage (%), multiply the values to 100. From Figure 2(b) and Figure 2(c), it shows that the best kernel type is the radial basis function (rbf) as only the rbf kernel can achieve 100% classification accuracy with combination of C and gamma higher than 1.0. Even though there are other combination between C, gamma and kernel that produce 100% classification accuracy, the grid search method works sequentially. Thus, the best parameter grid value given in Table 2 is the first combination it meets first during searching that achieve 100% classification accuracy.

The experiment proceeds by using the parameter values based on the results in Table 2 to build a new support vector classifier model for testing and evaluation of the model. A hold-out test set with ratio of 80% training set and 20% testing set is used to ensure the same size of data is used with grid search cross-validation technique previously.

Figure 3(a) shows the confusion matrix of the model on training set data while Figure 3(b) shows the confusion matrix of the model on testing set data. From the figures, it can be seen that the model has successfully predicted all the sample based on its quality correctly inside both the training and testing set data.



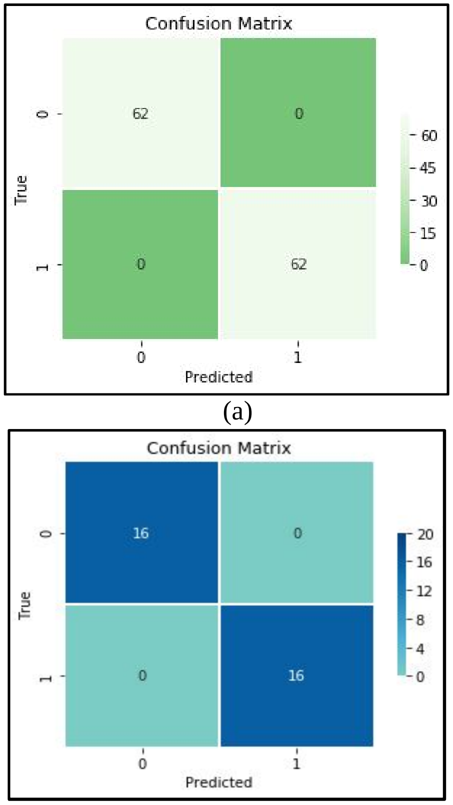
(a)



(b)

(c)

Figure 2: Heatmap of classification accuracy on testing set for different values of parameter.



(a)

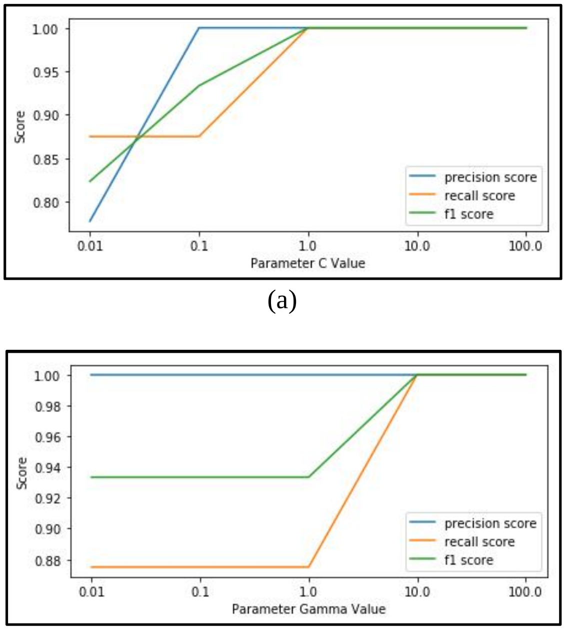
(b)

Figure3: Confusion matrix of model used

Table 3 tabulates the performance measure score and classification accuracy score achieved by the model. Since the model can classify the quality of agarwood oil perfectly as shown in figures above, the performance measure score of precision, recall and F1 measure has achieved score of 1.0. Figure 4(a) shows the performance measure score of support vector classifier model when using different value for parameter C and constant value for parameter gamma and kernel while Figure 4(b) shows the performance measure score of support vector classifier model when using different value for parameter gamma and constant value for parameter C and kernel. It can be seen from both performance measure scores figures that the best parameter values tend converge to be as predicted in optimal parameter values in Table 2 as the value increase. Thus, proving the optimal parameter value to be used for support vector classifier on the agarwood oil dataset is as in Table 2.

Table 3: Performance measure and accuracy score of the model.

Performance measure	Score
Accuracy	100%
Precision	1.0
Recall	1.0
F1 measure	1.0

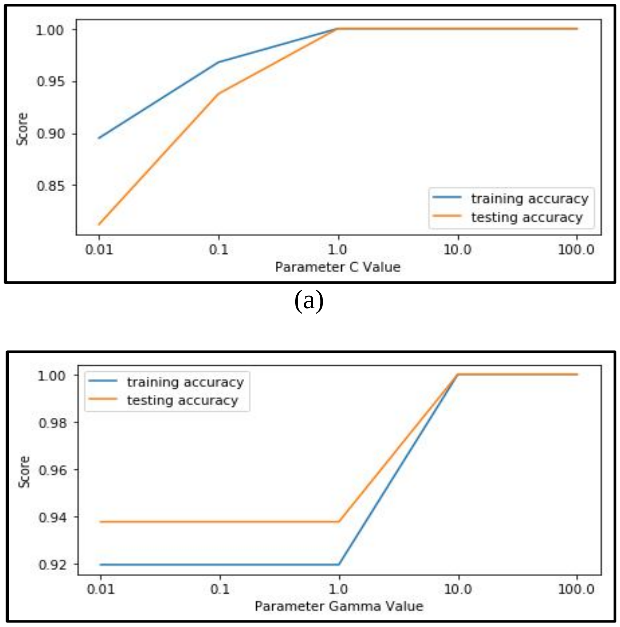


(a)

(b)

Figure 4: Performance measure score graph for variable parameters value

Figure 5(a) shows the classification accuracy score of support vector classifier model when using different value for parameter C and constant value for parameter gamma and kernel while Figure 5(b) shows the classification accuracy score of support vector classifier model when using different value for parameter gamma and constant value for parameter C and kernel. Both Figure 5(a) and Figure 5(b) has the same problem in reproducing the results achieved in grid search due to the software random generator differs and cannot be set consistent for both functions. However, the pattern is still the same and the score achieve 100% follows the grid search cross validation results in Figure 1(a) and Figure 2(a).



(a)

(b)

Figure 5: Classification accuracy score graph for variable parameters value.

4. CONCLUSION

The agarwood oil quality classifier using support vector classifier model with grid search cross validation technique has been successfully built by using combination of value 1 for parameter C, 10 for parameter gamma and using radial basis function (rbf) kernel achieve the highest performance score. The finding is sufficient and very significant especially for agarwood oil quality grading and its related research area.

ACKNOWLEDGEMENT

This research is funded by Research Management Institute (RMC), UniversitiTeknologi MARA (UiTM) Selangor, Malaysia under Grant No: 600-IRMI/MyRA 5/3/ BESTARI (041/2017). The authors would like to thank Prof IrDrMohd Nasir Taib, Associate Professor IrDrMohdHezriFazalul Rahman and Associate Professor IrDr Ahmad IhsanMohd Yassin for their valuable knowledge contribution in this article.

REFERENCES

[1] N. S. A. Zubir, M. A. Abas, N. Ismail, N. A. M. Ali, M. H. F. Rahiman, and N. K. Mun, “Analysis of Algorithms Variation in Multilayer Perceptron Neural Network for Agarwood Oil Qualities Classification,” no. August, pp. 4–5, 2017. https://doi.org/10.1109/ICSGRC.2017.8070580

[2] R. Naef, “The volatile and semi-volatile constituents of agarwood, the infected heartwood of Aquilaria species: A review.,” FlavourFragr. J., vol. 26, no. September 2010, pp. 73–89, 2011. https://doi.org/10.1002/ffj.2034

[3] N. A. M. Ali et al., “Comparison Of Chemical Profiles Of Selected Gaharu Oils From Peninsular Malaysia,” vol. 12, no. 2, pp. 338–340, 2008.

[4] N. Ismail, M. H. F. Rahiman, M. N. Taib, N. A. M. Ali, M. Jamil, and S. N. Tajuddin, “Application of ANN in agarwood oil grade classification,” Proc. - 2014 IEEE 10th Int. Colloq. Signal Process. Its Appl. CSPA 2014, pp. 216–220, 2014. https://doi.org/10.1109/CSPA.2014.6805751

[5] M. A. A. Aziz, N. Ismail, I. M. Yassin, A. Zabidi, and M. S. A. M. Ali, “Agarwood Oil Quality Classification using Cascade- Forward Neural Network,” pp. 112–115, 2015. https://doi.org/10.1109/ICSGRC.2015.7412475

[6] S. Lias, N. A. M. Ali, M. Jamil, M. H. Zainal, and S. H. A. Ghani, “Classification of Pure and Mixture Agarwood oils by Electronic Nose and Discriminant Factorial Analysis (DFA),” in 2015 International Conference on Smart Sensors and Application, 2015, pp. 7–10. https://doi.org/10.1109/ICSSA.2015.7322500

[7] Y. Es-Saady, I. El Massi, M. El Yassa, D. Mammass, and A. Benazoun, “Automatic recognition of plant leaves diseases based on serial combination of two SVM classifiers,” Proc. 2016 Int. Conf. Electr. Inf. Technol. ICEIT 2016, pp. 561–566, 2016. https://doi.org/10.1109/EITech.2016.7519661

[8] K. Elangovan and S. Nalini, “Plant Disease Classification Using Image Segmentation and SVM Techniques,” Int. J. Comput. Intell. Res., vol. 13, no. 7, pp. 1821–1828, 2017.

[9] Z. Cao, S. Wang, and L. Guo, “Using Kernel SVM for Predicting Membrane Protein Types by Fusing PseAAC and DipC,” Int. Conf. Comput. Sci. Netw. Technol., pp. 143–147, 2017.

[10] Pedregosa, “Scikit-learn: Machine Learning in Python,” JMLR, vol. 12, pp. 2825–2830, 2011.

[11] C. X. S. A. R. Images, “Paddy-Rice Phenology Classification Based on Machine-Learning Methods Using Multitemporal Co-Polar X-Band SAR Images,” vol. 9, no. 6, pp. 2509–2519, 2016.

[12] A. C. Müller and S. Guido, Introduction to machine learning with Python, 1st ed. O’Reilly Media, 2016.

[13] A. Géron, Hands-On Machine Learning with Scikit-Learn and TensorFlow, 1st ed. O’Reilly Media, 2017.

[14] J. D. Kelleher, B. Mac Namee, and A. D’Arcy, Fundamentals of machine learning for predictive data analytics, 1st ed. Massachusetts: The MIT Press, 2015.

[15] Pavithra G, Abirami P, Bhuvaneshwari S, Dharani S and Haridharani B, “A Survey on Intrusion Detectin Mechanism using Machine Learning Algorithms”, Int. J. of Emerging Trends in Engineering Research (IJETER), Vol. 8, No. 4, 2020.
https://doi.org/10.30534/ijeter/2020/01842020

[16] N. Chandra Sekhar Reddy, Purna Chandra Rao Vemuri and A. Govardhan, “An Empirical Study on Support Vector Machine for Intrusion Detection”, Int. J. of Emerging Trends in Engineering Research (IJETER), Vol. 7, No. 10, 2019.
https://doi.org/10.30534/ijeter/2019/037102019