Classification of Agarwood Oil Quality Using Random Forest And Grid Search Crossvalidation

Mohamad Aqib Haqmi Abas[1], Nurul Syakila Ahmad Zubair[1], Nurlaila Ismail[1], Ahmad Ihsan Mohd Yassin[1], Saiful Nizam Tajuddin[2] and Mohd Nasir Taib[1]

***Abstract*— This paper presents a machine learning technique to classify the agarwood oil quality. The random forest classifier model is used with the grid search cross validation technique to classify the quality of agarwood oil. The data of agarwood oil sample were obtained from Forest Research Institute Malaysia (FRIM) and Universiti Malaysia Pahang, Malaysia. In this experiment, the chemical compound abundances information of the agarwood oil that has been extracted from GC-MS machine is used as the input feature and the quality of the sample oil which is high quality and low quality is used as the output feature. Based on the result obtained from this study, using Gini impurity measure as criterion combined with 3 level maximum depth of decision trees and 3 number of maximum features for each tree provides the best classification accuracy of the agarwood oil quality sample at 100% and performance measure scores of 1.0.**

***Index Terms*— random forest, agarwood oil quality, machine learning, grid search, cross validation**

I. INTRODUCTION

Agarwood oil is an essential oil which that is produced by agarwood plant. Essential oil is well known to be more expensive than normal fragrance oils and perfume oils because it contains the true essence of the plant and it is known to gives therapeutic benefits [1]. Agarwood oil has been known to be used widely in religious ceremony, perfumery industry and as a traditional medicine. It is being traded internationally with a high market demand from countries in the Middle East.

Traditional grading method of the agarwood oil is by hiring trained experts to manually grade the agarwood by examining the physical properties of the agarwood oil such as its odour and colour. However, this method has a lot of disadvantages mainly being costly to hire experts, consumes a lot of time to grade the quality in large production samples and has poor reproducibility. This is because human nose cannot be used to grade a large samples of agarwood oil as it will easily get fatigued [2], [3]. Therefore, researchers have begun to use the chemical compound properties of the agarwood oil to grade it to high quality or low quality. Numerous number of studies show that grading agarwood oil using its chemical properties works better when compared to using traditional grading method. Since the past few years, there are a few studies of grading of agarwood oil using machine learning classifier model [4], [5].

This manuscript is submitted on 6th December 2017 and accepted on 26th April 2018. [1]Faculty of Electrical Engineering, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia[2] Bio Aromatic Research Centre of Excellent, University Malaysia Pahang, 26300 Gambang, Pahang, MalaysiaMohdaqib93@yahoo.com

Random forest is a type of supervised machine learning classifier model. It is currently among the most widely used machine learning models [6]. Among machine learning model, random forest and neural networks are known as black box method as it is usually hard to explain the predictions made. Random forest is one of the ensemble method of decision tree. It is used to address the problem in decision trees, where in decision trees classifiers the model tends to overfit easily. This is because the top levels of the tree have great impact on selecting the answer, thus if the new data does not have same distribution, the model will have problem to generalize.

Since decision trees will likely overfit on some part of data, the idea behind random forest is to build multiple decision trees that works well and overfit in different ways, thus having the ability to reduce the amount of overfitting by averaging the results. During training phase, the data is being sampled repeatedly with replacement where the process is known as bagging or also called bootstrap aggregating. The decision trees are being constructed using randomly subset of features. When making predictions and selecting the answer on unseen data, each decision tree is evaluated independently of each other, then the majority vote will be chose as the answer. Some of the advantages of Random Forest classifier are it is known to be very robust towards missing values, outliers and imbalanced dataset. Furthermore, the data does not need to be normalized before they are used to train the Random Forest classifier [7]. Random Forest computation can also be parallelized through multiple CPU cores easily as the decision trees in the model will be train independently. Random Forest model also have a disadvantage which is it requires a longer time to train as the number of trees increased.

Grid search is one of hyperparameter tuning technique used to find optimal parameters for machine learning model[6], [8]. It uses brute-force technique where it will try all possible combinations of the parameter values of interest. Cross-validation is a data splitting method of evaluating generalization performance of model. In cross-validation, the data is being split in multiple folds and a model will be trained for each fold. This makes cross-validation technique better than using single split technique (hold-out test set) as the dataset will be able to fully utilised and used effectively. Combination of grid search and cross-validation technique to find the best parameter of a model will be able to get a more stable and accurate prediction of parameters but at the cost of having a larger time taken to train all the models.

II. LITERATURE REVEW

There have been few studies that covers on agarwood oil usage and quality in the past few years. In [9], the author has carried out analysis of agarwood oil using the GC-MS data. The analysis was based on 64 chemical compounds from 7 samples of agarwood oil. The agarwood oil data used was complex mixture of chemical compounds and different from each sample. The author states that the distribution studied from the agarwood oil data shows that it is not normal and there are 5 main components identified from the 64 chemical compounds based on PCA application analysis.

In [10], the author used artificial neural network (ANN) to classify the agarwood oil quality. The significant of the compound is identified using Z-Score technique and the amount of agarwood oil data has been increase synthetically to increase the accuracy of ANN classification. The data of agarwood oil used has been increased synthetically due to limitation of ANN model that requires a high number of data to perform the analysis. The result shows an increase in accuracy from 75% without the synthetic generated data to 100% accuracy with the synthetic generated data.

Machine learning classifier model has also been used widely in agriculture domain. Hossam *et al.* [11], successfully developed a classification system that uses Random Forest algorithm to classify the fruit based on images. The fruit images of apples, strawberry and oranges were used and analysed. The images have been resized and undergone feature extraction before it was used in the classifier model. Based on the results obtained, the author states that Random Forest algorithm can provide a much better accuracy compared to Support Vector Machines (SVM) and *k*-Nearest Neighbour (*k*-NN).

In [12], the author uses Random Forest classifier to classify the fruit diseases on apple. The study was to classify between three mutual diseases of apple fruits which are apple scab, apple rot and apple blotch. The colour and texture feature of the images are fused together for better accuracy of Random Forest. Image fusion technique has been carried out in the study to join related information from more than one images into one image. The infected apple images are then segmented using k-means clustering technique on the diseased part.

The application of Random Forest algorithm has also been proven to be better than other classification algorithm such as SVM when classifying imbalanced dataset. Ziming Wu *et al.* [13] apply Random Forest algorithm to business data with imbalanced distribution and missing user features. The study combines the use of heuristic bootstrap sampling method and ensemble learning algorithm in a large-scale insurance business data. Random Forest also gives an advantage of the ability to run in parallel to maximize and fully utilize the parallel computing capability. The result shows the Random Forest algorithm is more suitable with the imbalanced insurance product recommendation and potential customer analysis data than other classifier algorithm.

III. EXPERIMENTAL SETUP

The whole process of the experiment which includes the building, training and evaluation of Random Forest algorithm has been done using anaconda software. The agarwood oil dataset that was used in the experiment is 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 compounds of agarwood oil extracted from GC-MS analysis. The output feature of the data samples is the quality of agarwood oil which is high quality and low quality.

Fig. 1 shows the flowchart of the overall experimental procedures used

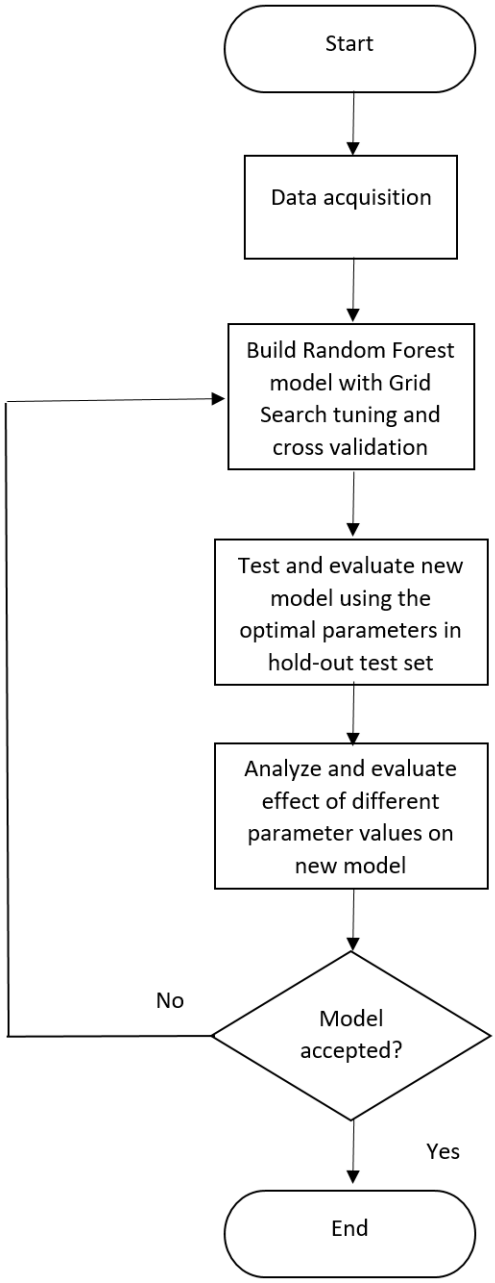


Fig.1 Flowchart of overall experimental procedures

The first step is data acquisition. The dataset used in the experiment consist of 132 agarwood oil sample with 7 input features and 1 target feature (output). 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 which are 'high quality' and 'low quality'. There are 78 samples of 'high quality' agarwood oil and 54 samples of 'low quality' agarwood oil. The dataset is slightly imbalanced with 59.1% of the data is high quality and remaining 40.9% of data is low quality.

Next, the data acquired will be used to train the Random Forest classifier with grid search cross-validation technique. The data used does not need to undergo any pre-processing method like in other classifier because Random Forest classifier is known to be robust with missing values, outlier and imbalanced dataset.

The parameters value grid of Random Forest that will be used are the criterion computation, maximum depth of tree and maximum number of features used. The other parameter on building the Random Forest classifier will be using the default value. The parameter values used to find for criterion measure is either Gini impurity measure or entropy impurity measure; for maximum depth of tree is between 2, 3 or 4 levels; and for maximum features used for each tree is between 2, 3 or 4 features. Overall there are 18 possible combinations of parameter values to be searched. The grid search technique will be used alongside cross-validation to have a better estimate of the generalization performance of classifier model [6]. The number of folds used for cross-validation process is 10.

In Decision Tree, the method used to split from its root node to leaf node is either Gini impurity measure [14]

$$G_i = 1 - \sum_{k=1}^{n} {P_{i,k}}^2 \quad (1)$$

or Entropy impurity measure

$$H_i = -\sum_{k=1}^{n} p_{i,k} log(p_{i,k}) \quad (2)$$

between Gini impurity measure and Entropy impurity measure, Gini impurity is slightly faster to compute as it does not have to compute the log equation. However, Entropy impurity produce more balanced trees, while Gini impurity tends to isolate most frequent class in its own branch.

After the optimal parameters for Random Forest has been identified, a new Random Forest model will be train using hold-out test set with ratio of 75% training set and 25% testing set [15]. This is done to evaluate and analyze the model performance by using the optimal parameters found earlier in grid search method in a new model.

Next, the model built will be tested, evaluated and analyzed in the model evaluation process. For evaluating the model performance, the performance measures used are classification accuracy, confusion-matrix based performance measure and precision, recall and F_1 measure score. Using precision, recall and F_1 measure score would give a better and more precise measurement information on the model built than classification accuracy.

The classification accuracy formula is given as [8]

$$classification\ accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (3)$$

where TP, TN, FP and FN are from confusion matrix result for True Positive, True Negative, False Positive and False Negative. The positive and negative in confusion matrix only describes binary output classes of 0 and 1. In this experiment positive is used to represent the low quality class, while negative used to represent high quality class. The formula for precision and recall are given as

$$precision = \frac{TP}{(TP + FP)} \quad (4)$$

and

$$recall = \frac{TP}{(TP + FN)} \quad (5)$$

Precision would measure for how many number of samples that is predicted positive are actually positive. Recall would measure how many number of samples that are actually positive has successfully been predicted as positive. In general, precision is the metric used when aiming to limit the false positive predictions while recall is the metric used when the goal is to identify all positive samples. There will always be a trade-off between optimizing recall measure and optimizing precision measure. Therefore, F_1 measure score always used together with precision and recall to obtain a more detailed information. It is the harmonic mean of precision and recall and given as [8]

$$F_1\ score = 2 \times \frac{(precision \times recall)}{(precision + recall)} \quad (6)$$

After the evaluation phase of using the parameter values found using grid search, a new Random Forest classifier model is build using the variation values for maximum depth and maximum features while other parameters are kept at constant and default value. This is done to analyse and evaluate the effect when using different values of parameters to build the Random Forest model based on the results obtained.

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 there is any error or problems with the score of performance measures, the model will be rebuilt with the data and the data and model will be checked for the errors or problems.

After the optimal parameters for Random Forest has been identified, a new Random Forest model will be train using hold-out test set with ratio of 75% training set and 25% testing set [15]. This is done to evaluate and analyze the model performance by using the optimal parameters found earlier in grid search method in a new model.

IV. RESULTS AND DISCUSSION

Table 1 tabulates the results of grid search method with cross-validation on the agarwood oil dataset for Random Forest classification after running through 18 possible combinations of parameter values. Based on the result it was found that the best criterion measure on tree split are using

combination of Gini impurity measure with 3 maximum depth of levels in the tree and 3 maximum number of features for bagging technique.

Table-1. Result of grid search method.

Parameter grid	Optimal value
criterion	Gini
maximum depth	3
maximum features	3

The experiment proceeds by using the parameter values found in Table 1 to build new Random Forest model for testing and evaluation of the model. Fig. 2 shows the feature importance of each input features used in the classifier. Note that the number starts with 0 to 6 instead of 1 to 7 because of the structure and format of the program used. Table 2 tabulates the result values of the feature importance shown in Fig. 2. The figure shows the rates of how important each feature is used for decision making that Random Forest makes. The values are between 0 to 1, where 0 means the feature does not used at all for the classification process and 1 means the feature perfectly predict the target. It can be seen that the γ-eudesmol has the largest importance value for the decision-making process in the classifier with value of 0.40220819, followed by β-agarofuran with 0.29472443, then 10-epi-γ-eudesmol with 0.2569628. The remaining 4 chemical compounds play less important role in Random Forest classification with having less than 0.05 value of importance. On the fourth place of importance value ranking is α-agarofuran with 0.02885749, followed by hexadecanol with 0.00906812, then eudesmol with 0.00606994 and lastly longifolol with only 0.00210903.

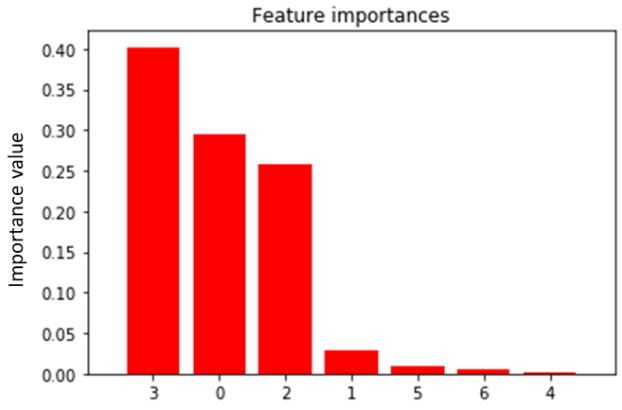


Fig.2 Feature importance of 7 input features used.

Table-2. Feature importance value of 7 input features used.

Chemical compound	Importance value
β-agarofuran	0.29472443
α-agarofuran	0.02885749
10-epi-γ-eudesmol	0.2569628
γ-eudesmol	0.40220819
longifolol	0.00210903
hexadecanol	0.00906812
eudesmol	0.00606994

Fig. 3 shows the confusion matrix of the model. It can be seen that the model has successfully predicted the 13 samples of low quality and 20 samples of high quality inside the testing set data.

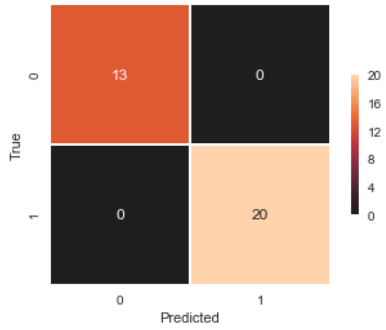


Fig.3 Confusion matrix of model.

Table 3 tabulates the performance measure score and accuracy score achieved by the model. Since the model are able to classify the quality of agarwood oil perfectly as shown in Fig. 3 above, the performance measure of precision, recall and F_1 measure has all achieved score of 1.0.

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

Fig. 4 shows that the accuracy of Random Forest model for both training and testing set is 100% when using 3 maximum depth level for the decision trees. When using maximum depth of 2, the training accuracy is at 96% while the testing accuracy is at 94%. At maximum depth of 4, the training accuracy achieve 100% but the testing accuracy drops to 97%.

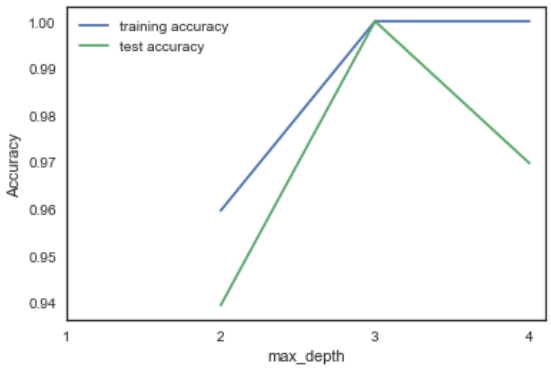


Fig.4 Classification accuracy of model using different values for maximum depth of trees.

Fig. 5 shows performance measure scores when using different values for maximum depth. It can be seen that the recall score is perfect for all 3 values. This indicates that the positive class (low quality) is successfully predicted accurately when using all 3 different values. However, the precision score when using value of 2 and 4 does not achieve the perfect value of 1.0 which is at 0.87 and 0.93 respectively. This indicates that there are misclassification of negative class (high quality) being positive (low quality) when using values of 2 and 4. The F_1 measure score when using value of 2 is 0.93, then it increased to 1.0 when using value of 3, then it dropped slightly to 0.9 when using value of 4.

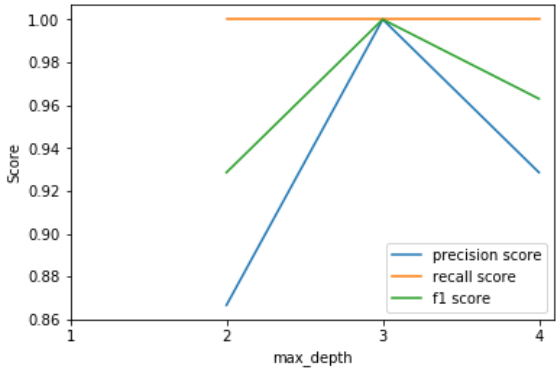


Fig.5 Performance measure score of model using different values for maximum depth of trees.

Fig. 6 shows that the accuracy of Random Forest model for training and testing set when using values of 2 is 97% and 94% respectively. Both accuracy increased to 100% when using value of 3 and 4 for the maximum features used.

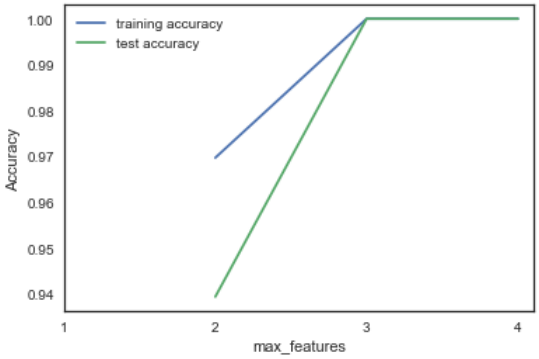


Figure-6 Classification accuracy of model using different values for maximum features used.

Fig. 7 shows the recall score is perfect when using all 3 values for maximum features used. However, for values 2, the precision score is at 0.87, which indicates a false positive prediction. The F_1 measure score when using value of 2 is 0.93, then increased to 1.0 for value 3 and 4.

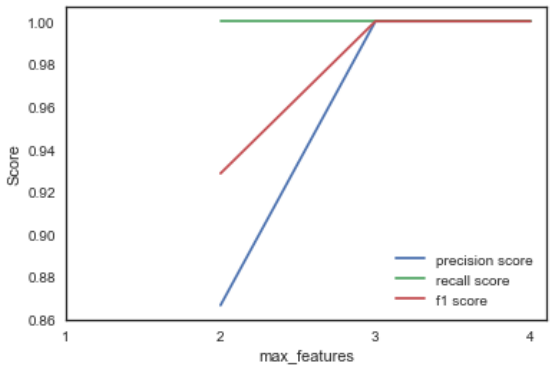


Fig.7 Performance measure score of model using different values for maximum features used.

Based on the accuracy and performance measure score in Fig. 6 and Fig. 7 and the grid search optimum value result in Table 1, it can be concluded that the grid search technique meets with the combination of using value 3 maximum features that have the highest accuracy before value 4. Thus the grid search method choose the value 3 as the optimum value for maximum features used.

V. CONCLUSION

The agarwood oil quality classifier using Random Forest model has been successfully built. Based on the results of optimum parameter value using grid search and the scores for classification accuracy and performance measure of the model, Random Forest model built using combination of maximum depth of 3 with maximum feature of 3 and Gini impurity measure achieve the highest classification score.

VI. ACKNOWLEDGEMENT

The dataset of agarwood oil used in this paper is from Forest Research Institute Malaysia (FRIM) and Universiti Malaysia Pahang, Malaysia. The authors' appreciation for all staff involved for their help on data collection from FRIM, UMP and Faculty of Electrical Engineering, UiTM Shah Alam.

VII. REFERENCES

[1] N. B. A. B. SIDIK, "Comparison Of Gaharu (Aquilaria Malaccensis) Essential Oil Composition Between Each Country," University Malaysia Pahang, 2008.

[2] N. A. M. Ali *et al.*, "Comparison Of Chemical Profiles Of Selected Gaharu Oils From Peninsular Malaysia," *Malaysian J. Anal. Sci.*, vol. 12, no. 2, pp. 338–340, 2008.

[3] R. Naef, "The volatile and semi-volatile constituents of agarwood , the infected heartwood of Aquilaria species : A review.," *Flavour Fragr. J.*, vol. 26, no. September 2010, pp. 73–89, 2011.

[4] N. Ismail, M. Hezri, and F. Rahiman, "The Grading of Agarwood Oil Quality using k- Nearest Neighbor (k-NN)," *IEEE Conf. Syst. Process Control*, pp. 13–15, 2013.

[5] 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.

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

[7] H. Brink, J. W. Richards, and M. Fetherolf, *Real-world machine learning*, 1st ed. Manning Publications, 2016.

[8] 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.

[9] N. A. M. Ali, N. Ismail, and M. N. Taib, "Analysis of Agarwood Oil (Aquilaria Malaccensis) Based on GC-MS Data," in *2012 IEEE 8th International Colloquium on Signal Processing and its Applications*, 2012, pp. 470–473.

[10] 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.

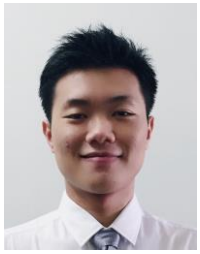
[11] H. M. Zawbaa, M. Hazman, M. Abbass, and A. E. Hassanien, "Automatic fruit classification using random forest algorithm," *2014 14th Int. Conf. Hybrid Intell. Syst. HIS 2014*, pp. 164–168, 2014.

[12] B. J. Samajpati and S. D. Degadwala, "Hybrid Approach for Apple Fruit Diseases Detection and Classification Using Random Forest Classifier," no. 2013, pp. 1015–1019, 2016.

[13] Z. Wu, W. Lin, Z. Zhang, A. Wen, and L. Lin, "An Ensemble Random Forest Algorithm for Insurance Big Data Analysis," *22017 IEEE Int. Conf. Comput. Sci. Eng. IEEE Int. Conf. Embed. Ubiquitous Comput.*, pp. 531–536, 2017.

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

[15] Pedregosa, "Scikit-learn: Machine Learning in Python," *JMLR*, vol. 12, pp. 2825–2830, 2011.



Mohamad Aqib Haqmi bin Abas was born in Malaysia, on December 1993. He received a degree in Bachelor of Engineering (Hons.) from Universiti Teknologi MARA (UiTM) in 2016.He is currently enrolled as a Master of Science student at Universiti Teknologi MARA (UiTM).

Dr Nurlaila Ismail was born in Malaysia, Dec 10 1982. She received her PhD in Electrical Engineering from Universiti Teknologi MARA, Malaysia in 2015, MSc in Electrical Engineering from Universiti Teknologi MARA, Malaysia in 2011 and Bachelor in Electrical & Electronics Engineering (Hons.) from Universiti Teknologi MARA, Malaysia in 2005.She currently lectures at Universiti Teknologi MARA, Malaysia. Her research interests are artificial intelligence and system identification.

Dr Ahmad Ihsan Mohd Yassin was born in Malaysia.He received his PhD in Electrical Engineering from Universiti Teknologi MARA (UiTM) in 2014 and MSc in Electrical Engineering from Universiti Teknologi MARA (UiTM) in 2008. His research interest are in System Identification, Artificial Intelligence (Fuzzy Logic, Neural Networks and Deep Learning), and Stochastic Optimization Methods.

Prof Mohd Nasir Taib was born in Malaysia. He received his degree in Electrical Engineering from the University of Tasmania, Australia, MSc in Control Engineering from Sheffield University, UK, and PhD in Instrumentation from UMIST, UK. Currently, he is a Professor and Dean of The Faculty of Electrical Engineering, Universiti Teknologi MARA (UiTM. Currently, his research interests are system and instrumentatation.