ISSN 2347 - 3983
Volume 8. No. 7, July 2020
International Journal of Emerging Trends in Engineering Research
Available Online at http://www.warse.org/IJETER/static/pdf/file/ijeter0872020.pdf
https://doi.org/10.30534/ijeter/2020/03872020

Linear Regression of Gaharu Oil Significant Compounds for Oil Quality Differentiation

Noratikah Zawani Mahabob[1], Zakiah Mohd Yusoff[2*], Nurlaila Ismail[1], Mohd Nasir Taib[1]
[1]Faculty of Electrical Engineering, Universiti Teknologi MARA (UiTM), Selangor, Malaysia,
[2]Faculty of Electrical Engineering, Universiti Teknologi MARA (UiTM), Cawangan Johor, Kampus Pasir Gudang, Johor, Malaysia
*zakiah9018@uitm.edu.my

ABSTRACT

This study presented the Linear Regression model that is trained in Feed Forward Neural Network (FFNN). The model is trained using Matlab version R2017a. The sample of dataset of Gaharu oil that used in this research study was obtained from Forest Research Institute Malaysia (FRIM), Selangor Malaysia, and BioAromatic Research Centre of Excellence (BARCE), Universiti Malaysia Pahang (UMP), Malaysia. As Feed-Forward Neural Network consists of input layer, hidden layer and output layer, this related to situation that is used in the research. In this experiment, the seven significant compounds of gaharu oil that consists of 96 data samples from high and low quality are representing the input layer. The hidden layer is varying from 1 to 10 hidden neurons to evaluate each of the performances. The output layer is presenting the quality of gaharu oil within the range of 0 and 1 which is low and high, respectively. The experiment involved the data pre-processing consists of data normalization, randomization and division. The parameter concerned is on the value of correlation coefficient, R and the mean squared error (MSE) for each of the hidden neurons. The training algorithm is involving Levenberg Marquadt as a default algorithm in FFNN besides has a good stability and fast in convergence. Based on the results of the study, hidden neurons number 2 outperforms others due to successfully show the best performance towards regression value and MSE.

Key words: Gaharu oil, linear regression, Levenberg Marquadt, correlation coefficient and MSE.

1. INTRODUCTION

Gaharu also known as “*oudh*”[1]“*woods of Gods”*[2] and “*eaglewood*”[1]is a word that describes dark resinous heartwood that forms in various genus plants family Thymelaeaceae (which found in Aquilaria and Gyrinops trees). The resinous wood is formed due to fungal infection. Gaharu itself has a pale and light colored but when the wood is at the state of matured, it will converts into dark colored; both dark brown or black and also produce aromatic resin. Nowadays, the variable usage of gaharu oil have been made the oil become highly demand in the market especially in the countries from the Middle East, China, Japan also including Malaysia. The special usage of gaharu oil has been applied in wide areas such as incense, ingredient for perfume (especially the dark color), and traditional medical preparation (treating such as asthma and jaundice) [3]. In addition, in the Middle East has been used gaharu oil and gaharu smoke as their symbol of wealth and during the wedding ceremony [4].

The gaharu oil is grading accordingly to its quality which is high and low. The high and low quality of the gaharu oil are differentiated by its color, odor, and price. For high quality oil, the odor is long lasting and it has a dark color, while the low quality oil is vice versa. Normally, high quality oil is more costly compared to low quality oil. The cost for high quality gaharu oil is normally from USD126 to USD633 per tola (12ml) [4]. In other countries, they had their own way of grading the gaharu oil. For example, in Malaysia, high quality oil known as “kelambak” while low quality classified as “gaharu”. In Japan, it classified the gaharu oil as kanankoh and jinkoh referring to highest and low quality respectively [5]. Traditionally, human sensory panel (human nose) is one of the methods used to analyze the odor of gaharu oil, but this method had weakness on repeatability and subjectivity. This is because human nose cannot control continuous production and a bulk numbers of samples since the human nose can fatigues rapidly [6], [7][8].

The most commonly method that used in statistical area which investigates between two variables is known as linear regression (LR) [9]–[11]. In regression area, linear regression is one of the simplest regression models that are used for predicting some results other than cox regression and logistic regression [12], [13]. The variables in single linear regression involve the relationship between single dependent variable and single independent variable which denoted as y and x, respectively in a linear graph [9], [10], [12]–[14].Based on Figure 1 and Figure 2, the linear regression can be an increase ratio or decrease ratio depends on the research data in research study mentioned in [9]. Theoretically, the dependent variables are the results by the estimation from independent variables [13]. The mathematical relationship of linear regression is based on linear equation which is y=mx+c. The m referring to the slope of the graph and c is the interceptof the fit line. Linear regression calculates the variables such as correlation coefficient (R), coefficient of determination (R^2) and error. For determining the best fit line, the correlation coefficient, R value has to be closeness to -1 or 1 similar to the value of coefficient of determination, R^2(the higher the value R^2, the more accurate the model) [9], [13].

Linear regression method has been used in research study [12] for predicting the students' admission into master program through their profiles. Instead of using linear regression, support vector regression, decision tress and random forest also methods used in this study. The methods are comparable each other and conclude that linear regression outperforms other due to highest result of R and lowest MSE. Next, linear regression has been proposed in predicting the correlation between test score for different courses and classroom attendance [15]. The result from scatter diagram in the study shows a positive correlation that students' attendance gives impacts for the test score. LR also successfully applied in research study [11]to predict online shopping based on sales value.

Next, Levenberg Marquadt (LM) algorithm or known as dumped least square method [16], [17]was developed by Kenneth Levenberg and Donald Marquadt. LM defines as an algorithm that provides a numerical solution to the problem of minimizing a non-linear function [16], [18]. In addition, LM is designed without having to compute the hessian matrix by approaching the second order training speed [19]–[21]. LM integrates between two algorithms which is Steepest Descent method and Gauss Newton algorithms which both algorithms have good speed and stability, respectively. LM is faster than Steepest Descent method but slower than Gauss Newton method. This algorithms proving to have advantages in convergence by terms of speed, stability while LM algorithms have limited training for only small and medium sized problems [16].

This paper presents on using regression model to be fed into neural network to predict which regression line of hidden neurons performs the best for gaharu oil quality discrimination. In addition, the dataset is trained using Levenberg Marquadt algorithm due to the most widely used in optimization algorithm in many research studies[22] and has been recommended for training in neural network by paper [16].

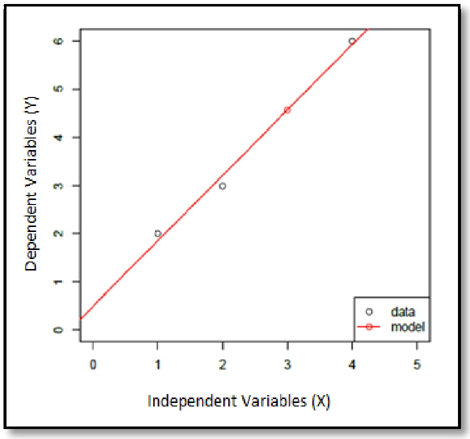


Figure 1: Increase Ratio Linear regression graph

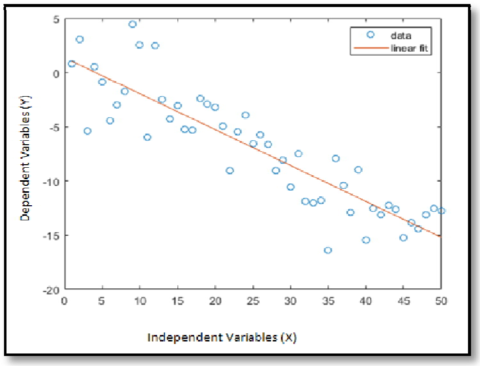


Figure 2: Decrease Ratio Linear regression graph

2. METHODOLOGY

The sample of dataset of gaharu oil that used in this research study was obtained from Forest Research Institute Malaysia (FRIM), Selangor Malaysia, and BioAromatic Research Centre of Excellence (BARCE), Universiti Malaysia Pahang (UMP), Malaysia consists of the total of 96 gaharu oil samples from low and high qualities. The output classified as a low or high quality (0 and 1 respectively), while the input will be the abundances (%) of compounds. The experiment is conducted using Matlab version R2017a to train and evaluate the performance of network model of machine learning.

Firstly, the experiment start with the data acquisition where the dataset contains of 96 samples of gaharu oil and out of that, 18 samples are from low quality while another 78 samples are from high quality classes. It is consists of 7 selected significant compounds referring to C1 until C7. They are coded as β-agarofuran, α-agarofuran, 10-epi-ϒ-eudesmol, ϒ-Eudesmol, Longifolol, Hexadecanol and Eudesmol.

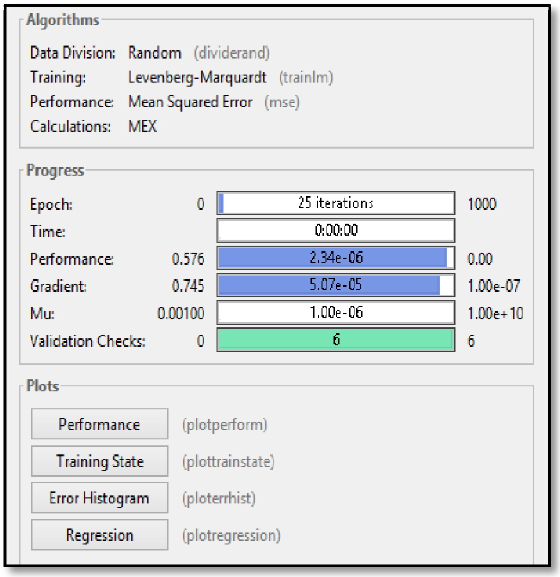
The next step, the data will undergo pre-processing stage. In this stage, where the data is normalized and randomized before dividing them into three dataset for training, validation and testing with the ratio of 70%, 15% and 15%. Data normalization is using min-max scaling technique as the feature scaling technique. The purpose of min-max scaling on gaharu oil dataset is to scale the continuous variable for input features to range of 0 to 1.

The process continued by training a network to create the network object. Feedforward neural network consists of a series of layers. The first layer has a connection from the network input. Each subsequent layer has a connection from the previous layer. The final layer produces the output of the network by purelin transfer function. The 'feedforwardnet' function is used to create a multilayer feed-forward network with regression in this experiment. The seven significant compounds of gaharu oil act as input layer, output layer is represents the quality of gaharu oil at the end of the experiment, while the number of neurons in hidden layer is varied from 1 to 10. 10 hidden neurons are selected for performance evaluation. A researcher observed that the number of hidden neurons can affected the model. If there is few number of hidden neuron will cause an inaccurate model while higher number of hidden neuron will cause a weak generalization ability [23].

Figure 3 shows the neural network training tool, which the Levenberg-Marquadt (LM) is used as a training algorithm in this experiment as recommended and mention in paper [16]due to the stability and fast convergence and it is a default training function in feed-forward network with the performance of mean squared error (MSE). During training, the matlab function calculates the value or correlation coefficient, R for each of dataset (training, validation and testing). The value of correlation coefficient, R is tabulated in Table 1, which the value referred to the relationship between output and target values [13]. Therefore, the model is concluding to have a good fitting when the value of R closes to 1. The 'plotregression' in Figure 3 also creates three regression plots for training, validation and testing subsets of outputs and targets results.

Table 1: Relationship on linear specification of R value

Correlation coefficient, R	Relationship
R < 0	Inverse, negative relationship
R = 0	Non-linear relationship
R > 0	Positive relationship
R = ±1	Perfect linear and stronger relationship



Algorithms

Data Division: Random (dividerand)
Training: Levenberg-Marquardt (trainlm)
Performance: Mean Squared Error (mse)
Calculations: MEX

Progress

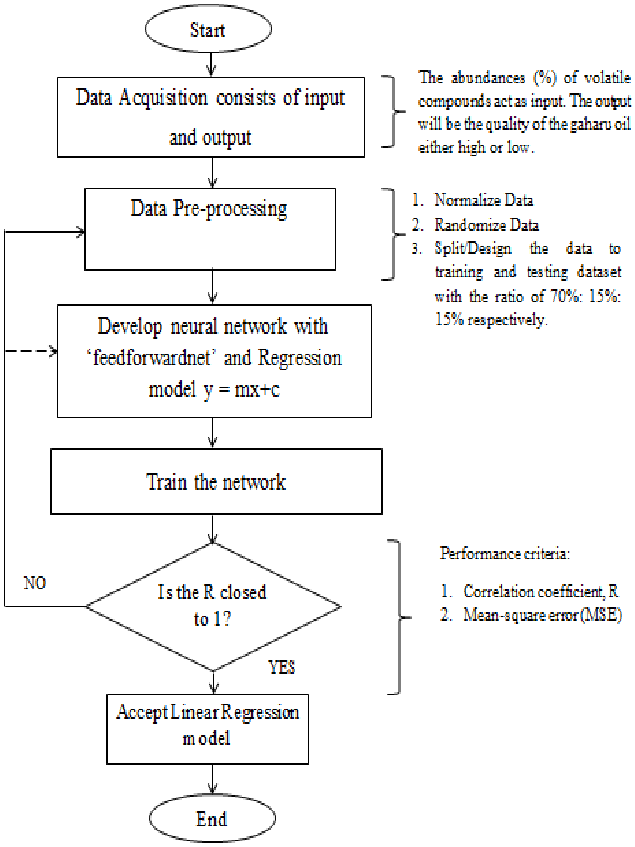
Epoch:	0	25 iterations	1000
Time:		0:00:00	
Performance:	0.576	2.34e-06	0.00
Gradient:	0.745	5.07e-05	1.00e-07
Mu:	0.00100	1.00e-06	1.00e+10
Validation Checks:	0	6	6

Plots

Performance (plotperform)
Training State (plottrainstate)
Error Histogram (ploterrhist)
Regression (plotregression)

Figure 3: Matlab Neural Network Training tool

In order to choose which neurons in hidden layer performs the best fit line of regression, other than the value of correlation of coefficient, the performance of mean squared error (MSE) is taken as the important criteria. Each time of the training is simulate, MSE and value or R in training dataset will be compared among 10 neurons and select the best that outperformed. The progress of the experiment is clearly explained through flowchart in Figure 4.



Start
Data Acquisition consists of input and output
The abundances (%) of volatile compounds act as input. The output will be the quality of the gaharu oil either high or low.
Data Pre-processing
1. Normalize Data
2. Randomize Data
3. Split/Design the data to training and testing dataset with the ratio of 70%: 15%: 15% respectively.
Develop neural network with 'feedforwardnet' and Regression model y = mx+c
Train the network
Is the R closed to 1?
NO
YES
Performance criteria:
1. Correlation coefficient, R
2. Mean-square error (MSE)
Accept Linear Regression model
End

Figure 4: Flowchart of details experiment

3. RESULTS AND DISCUSSION

This section shows the result obtained by training the regression model using feed forward neural network. It consists of value of correlation of coefficient, R, equation of linear regression and mean-squared error (MSE). Figure 5 shows the architecture of Feed forward Neural Network with input layer, hidden layer and output layer. The input layer has seven neurons presenting the seven selected compounds contributed to high and low quality of gaharu oil, the hidden layer is varied from 1 to 10 neurons while the output layer is representing the quality of gaharu oil (high=1 or low=0).

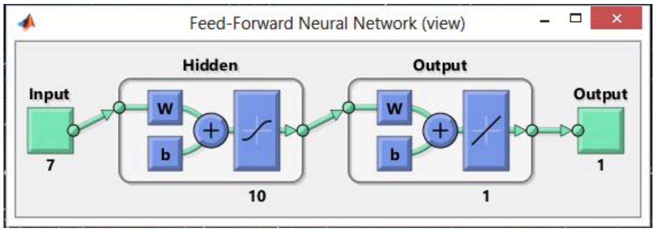


Figure 5: Architechure of Neural Network Training

3.1 Results for Regression

Result in Figure 6 shows the graph relationship between correlation coefficient, R versus number of hidden neurons. Between the ranges 1 to 10 hidden neurons, the maximum value of R achieved 1 at hidden neurons 1, 2, 3 and 9. The lowest value of R is obtained by hidden neurons number 6 which is 0.7611. Based on theory, the R value should have value +1 or -1 for the best fit line of regression. Therefore, according to the results obtained in the graph, four hidden neurons are chosen as they achieved the perfect fit line of regression according to the achievement of maximum R value which is neurons number 1, 2, 3 and 9.

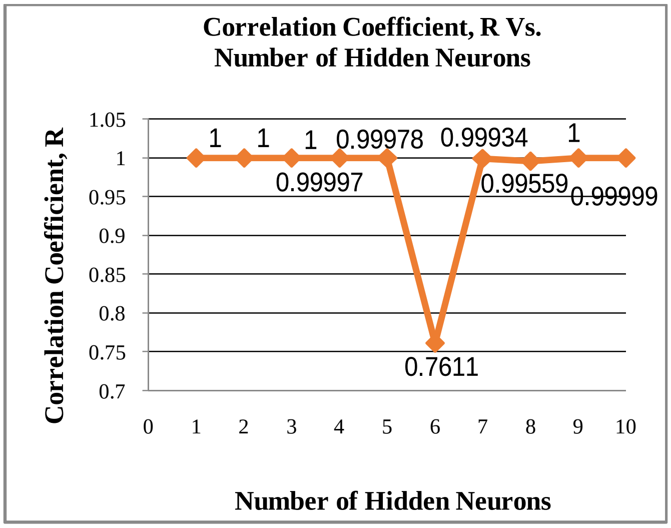


Figure 6: Graph of Correlation coefficient, R versus Number of Hidden Neurons

3.2 Results for Mean Square Error (MSE)

The result of mean squared error according to 10 hidden neurons is tabulated in the Table 2. It is found that, the highest MSE found at hidden neurons number 6 which is 4.88x10^{-02}, while the lowest value MSE is 7.69x10^{-15} which is at hidden neurons number 2.

Table 2: MSE vs. Number of Hidden Neurons

Number of hidden neurons	Mean squared error (MSE)
1	2.02x10^{-13}
2	*7.69x10^{-15}
3	6.85x10^{-14}
4	4.37x10^{-06}
5	1.21x10^{-09}
6	4.88x10^{-02}
7	5.15x10^{-05}
8	8.22x10^{-04}
9	1.81x10^{-07}
10	2.34x10^{-06}

**the lowest MSE*

3.3 Summarize Results

Figure 7, 8, 9 and 10 shows scatter diagram of the regression R for the training target of the chosen hidden neurons number 1, 2, 3 and 9, respectively. All of the training target have perfect value of R=1. The regression equation of training target are Output~=1(target)+2.4e^{-07}, Output~=1(target)+5.8e^{-08}, Output~=1(target)+8.2e^{-08} and Output~=1(target)+0.00051 for hidden neurons 1, 2, 3 and 9 respectively.

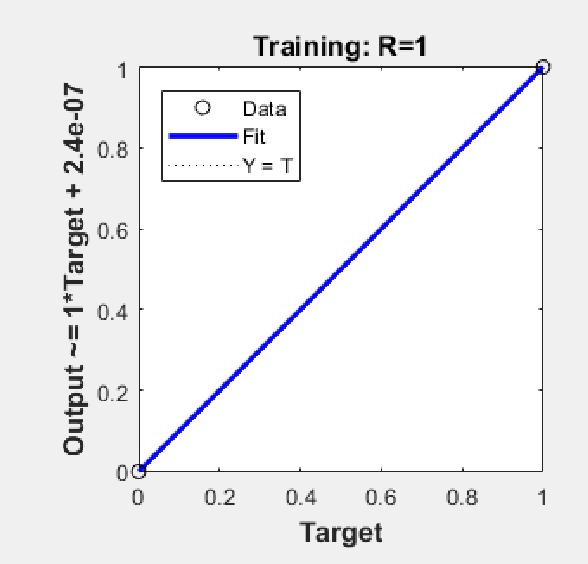


Figure 7: Training target of Hidden neurons 1

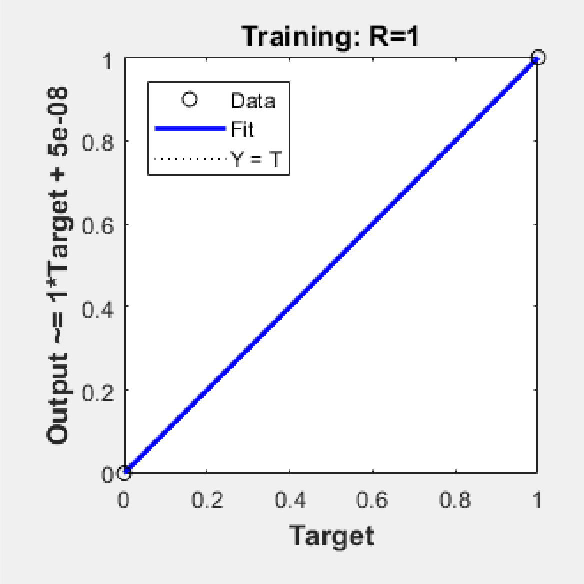


Figure 8: Training target of Hidden neurons 2

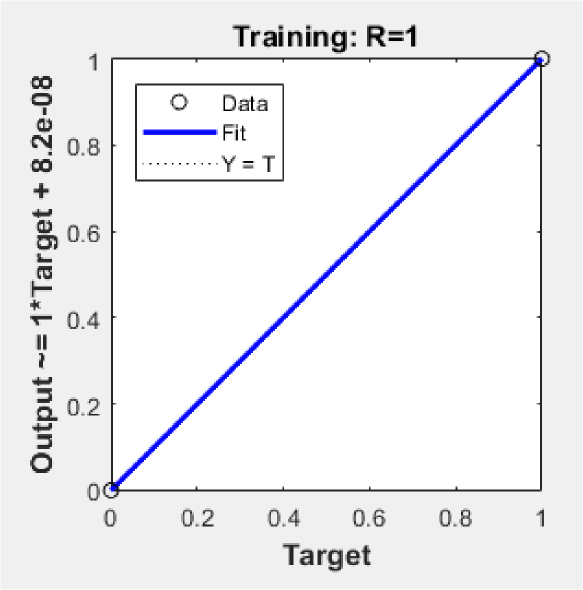


Figure 9 : Training target of Hidden neurons 3

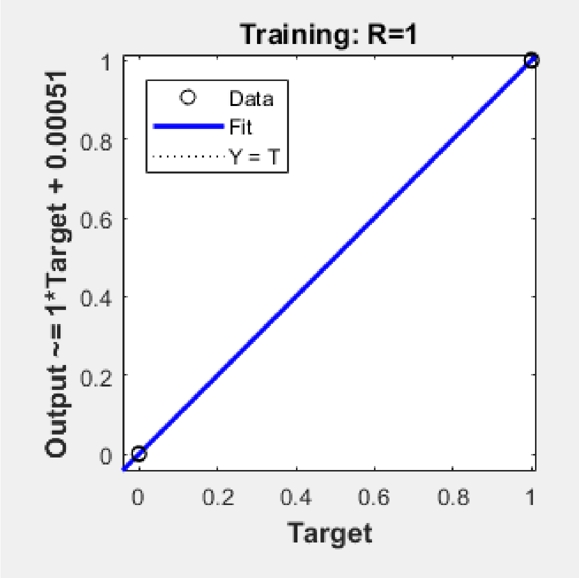


Figure 10: Training target of Hidden neurons 9

Table 3 tabulated the results of 4 chosen hidden neurons consists of neurons number 1, 2, 3 and 9 based on the performance on regression, R and MSE. The performances are comparable, where the hidden neurons number 2 outperforms others. It achieved R=1 which belong to an accurate results of linear relationship between the outputs and targets and obtained the lowest MSE among four hidden neurons (1, 2, 3 and 9 neurons) also the lowest among all hidden neurons.

Table 3 : Four selected Hidden Neurons vs. MSE

Hidden neurons	Value of R	MSE
1	1	2.02x10^{-13}
*2	1	7.69x10^{-15}
3	1	6.85x10^{-14}
9	1	1.81x10^{-07}

**The best hidden neurons*

Table 4 summarizes the final design parameter for the feed forward neural network architecture and training parameters. The number of input layer is set 7 compounds (C1 until C7). The quality of gaharu oil with high and low is represented at 1 output layer of neuron. The best neuron was found in hidden neurons 2 with the R value is 1 is the best regression with the regression equation is Output~=1(target)+5.8e^{-08} where the (target) is either high or low quality. Besides, it is proved by the lowest MSE by hidden neurons number 2 which is 7.69x10^{-15}.

Table 4: Final Design Parameter

Parameter	Value
Number of input layer	7
Number of hidden nodes	2
Number of output layer	1
Value of R	R=1
Linear Regression Equation	Output~=1(target)+5.8e^{-08}
Mean Square Error	7.69x10^{-15}

4. CONCLUSION

This study has successfully applied regression model to predict the gaharu oil quality discrimination. Seven selected compounds are selected as input data which contain high and low quality of gaharu oil and the output is either low quality as 0 or high quality as 1. The value of correlation coefficient, R according to scatter diagram and the value of MSE are concerned and comparable within the range of 1 to 10 hidden neurons. Levenberg Marquadt (LM) is used as training algorithm due to the good performance in many papers. Results of the study strongly showed that, gaharu oil obtained a best fit linear regression line with value of R exactly 1 at hidden neurons number 2. Besides, the finding is proven by the results of mean square error in hidden neurons number 2 which is the lowest compared to other neurons.

ACKNOWLEDGEMENT

This research is funded by Institute of Research Management and Innovation (IRMI), Universiti Teknologi MARA (UiTM) Kampus Pasir Gudang ,Universiti Teknologi MARA (UiTM) Selangor,and under Grant No: 600-IRMI/FRGS 5/3 (224/2019).The authors would like to thank Dr Nor Azah Mohd Ali from Forest Research Institute Malaysia (FRIM) and Assoc Prof Dr Saiful Nizam Tajuddin from BARCE (UMP Gambang, Pahang) for their valuable knowledge contribution in this article.

REFERENCES

[1] S. Akter, M. T. Islam, M. Zulkefeli, and S. I. Khan, “Agarwood Production - A Multidisciplinary Field to be Explored in Bangladesh,” *Int. J. Pharm. Life Sci.*, vol. 2, no. 1, pp. 22–32, 2013. https://doi.org/10.3329/ijpls.v2i1.15132

[2] R. Kalra and N. Kaushik, “A review of chemistry, quality and analysis of infected agarwood tree (Aquilaria sp.),” *Phytochem. Rev.*, vol. 16, no. 5, pp. 1045–1079, 2017.

[3] M. Syafiq, A. Ishak, Y. Yusof, M. H. Ahmad, and N. E. Alias, “Agarwood Grading Estimation Using Artificial Neural Network Technique and Carving Automation,” *J. Electr. Eng.*, vol. 16, no. 3, pp. 36–41, 2017.

[4] N. Ismail, N. A. M. Ali, M. Jamil, M. H. F. Rahiman, S. N. Tajuddin, and M. N. Taib, “A review study of agarwood oil and its quality analysis,” *J. Teknol. (Sciences Eng.*, vol. 68, no. 1, pp. 37–42, 2014. https://doi.org/10.11113/jt.v68.2419

[5] R. Naef, “The volatile and semi-volatile constituents of agarwood , the infected heartwood of Aquilaria species : A review .,” no. January, pp. 73–89, 2011.

[6] W. Hidayat, A. Y. M. Shakaff, M. N. Ahmad, and A. H. Adom, “Classification of Agarwood oil using an electronic nose,” *Sensors*, vol. 10, no. 5, pp. 4675–4685, 2010.

[7] P. E. Keller, “Mimicking biology: Applications of cognitive systems to electronic noses,” *IEEE Int. Symp. Intell. Control - Proc.*, pp. 447–451, 1999.

[8] M. Aqib *et al.*, “Agarwood Oil Quality Classification using Support Vector Classifier and Grid Search Cross Validation Hyperparameter Tuning,” *Int. J. Emerg. Trends Eng. Res.*, vol. 8, no. 6, pp. 6–11, 2020. https://doi.org/10.30534/ijeter/2020/55862020

[9] K. Lee *et al.*, “Comparison and Analysis of Linear Regression & Artificial Neural Network,” vol. 12, no. 20, pp. 9820–9825, 2017.

[10] O. S. Maliki, A. O. Agbo, A. O. Maliki, L. M. Ibeh, and C. O. Agwu, “Comparison of Regression Model and Artificial Neural Network Model for the prediction of Electrical Power generated in Nigeria,” vol. 2, no. 5, pp. 329–339, 2011.

[11] T. Gopalakrishnan, R. Choudhary, and S. Prasad, “Prediction of Sales Value in online shopping using Linear Regression,” *2018 4th Int. Conf. Comput. Commun. Autom.*, pp. 1–6, 2018. https://doi.org/10.1109/CCAA.2018.8777620

[12] M. S. Acharya, “A Comparison of Regression Models for Prediction of Graduate Admissions,” pp. 0–4, 2019.

[13] A. Schneider, G. Hommel, and M. Blettner, “Linear Regression Analysis,” vol. 107, no. 44, pp. 776–782, 2010.

[14] W. Tang and S. Ma, “Application of Regression and Artificial Neural Network in Ground Temperature Processing,” *2019 Int. Conf. Meteorol. Obs.*, pp. 1–4, 2019.

[15] F. L. Teaching and V. Blerkom, “Research on Correlation Analysis between Test Score and Classroom Attendance Based on Linear Regression Model,” pp. 545–548, 2010.

[16] H. Yu and B. M. Wilamowski, “Levenberg-Marquadt Training.”

[17] N. S. A. Zubir *et al.*, “Analysis of algorithms variation in Multilayer Perceptron Neural Network for agarwood oil qualities classification,” *2017 IEEE 8th Control Syst. Grad. Res. Colloquium, ICSGRC 2017 - Proc.*, no. August, pp. 122–126, 2017. https://doi.org/10.1109/ICSGRC.2017.8070580

[18] N. Mohamad, F. Zaini, A. Johari, I. Yassin, and A. Zabidi, “Comparison between Levenberg-Marquardt and Scaled Conjugate Gradient training algorithms for Breast Cancer Diagnosis using MLP,” *2010 6th Int. Colloq. Signal Process. its Appl.*, pp. 1–7, 2010. https://doi.org/10.1109/CSPA.2010.5545325

[19] Ö. Ki and E. Uncuo, “Comparison of three back-propagation training algorithms for two case studies,” vol. 12, no. October, pp. 434–442, 2005.

[20] P. Sehgal, “Comparative Study of GD , LM and SCG Method of Neural Network for Thyroid Disease Diagnosis,” vol. 1, pp. 34–39, 2015.

[21] N. Coskun and T. Yildirim, “The effects of training algorithms in MLP network on image classification,” *Proc. Int. Jt. Conf. Neural Networks*, vol. 2, no. see 17, pp. 1223–1226, 2003.

[22] S. Pahwa, “Comparative Study of Support Vector Machine with Artificial Neural Network Using Integer Datasets,” *Int. J. Adv. Res. Comput. Sci. Softw. Eng.*, vol. 6, no. 11, pp. 200–205, 2016.

[23] D. A. Petrosov, R. A. Vashchenko, A. A. Stepovoi, and N. V. Petrosova, “Application of artificial neural networks in genetic algorithm control problems,” *Int. J. Emerg. Trends Eng. Res.*, vol. 8, no. 1, pp. 177–181, 2020. https://doi.org/10.30534/ijeter/2020/24812020