Grading Of Agarwood Oil Quality Based On Its Chemical Compounds Using Self Organizing Map (SOM)

Mohamad Hushnie Haron[1*], Mohd Nasir Taib[1], Nurlaila Ismail[1], Nor AzahMohdAli[2],
Saiful Nizam Tajuddin[3]

[1]Faculty of Electrical Engineering, UniversitiTeknologi MARA (UiTM), Selangor, Malaysia,
*hushnie_haron@yahoo.com.my
[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

Agarwood oil is derived from resin saturated heartwood of Agarwood tree using solvent extraction, hydro-distillation and etc. The oil consists ofcomplex mixtures of chromone derivatives, oxygenated sesquiterpenes and sesquiterpene hydrocarbonsand they are significant in determining the quality of Agarwood oil. In this paper, a study focusing on the determination of high and low grades of Agarwood oil quality using SOMwas carried out. Three chemical compounds namely β-agarofuran, α-agarofuran and 10-epi-φ-eudesmol were used as inputs for SOM. In addition, KNN was used to confirm the Agarwood oil grades. This technique wasthen validated using independent samples. The results show that SOM was able to determine the high and low gradesof Agarwood oil with average silhouette indexes between reasonable and excellent, and KNN was able to confirmthese grades with 100% accuracies. In conclusion, the Agarwood oil quality gradingtechnique based on SOM was successfulin determining the Agarwood oil quality into high and low.

Key words : agarwood oil quality, gas chromatography-mass spectrometry, k-nearest neighbor, self-organizing map

1. INTRODUCTION

Agarwood tree is a beneficial tree because each component is valuable; branches and tree trunks can be turned into flakes of uniform quality powder, chips or essential oils [1-5]. The essential oil or Agarwood oil is traded based on its quality which is direct proportional to the price [1, 6, 7]. The quality of Agarwood oil is characterized by consumer perception, high fixation properties, color and fragrance [5, 8]. A few terms were used to define the quality of Agarwood oil such as high and low [9, 10], and A+, A and B [11]. In the Agarwood oil trade, it is necessary to acquire information about the grade of Agarwood oil and therefore, there is a need to realize a comprehensive and systematic standard grading system that

will be recognized across Agarwood sourcing region [12] or worldwide [1]. This standard grading system will lead to stability in the prices of Agarwood oil [8] and a better bargaining power for sellers [12]. Technically, the grading of Agarwood oil is conducted by qualified human graders utilizing their sensory panels based on their physical appearances[13, 14]. However, this method has limitation for example fatigue [7], subjectivity, reproducibility, time consumption and labor costs [15].

To extract the Agarwood's oil, techniques such as supercritical fluid extraction, solvent extraction, hydro-distillation, and others have been used. Prior to extraction, a pre-treatment on Agarwood samples such as chemical treatment, soaking in water and sonication must be performed [15]. Some studies use Gas Chromatography-Mass Spectrometer (GC-MS) [3, 8, 9, 16] and Solid-Phase Microextraction (SPME) [9, 16] to further analyze the extracted oil. The resin content of Agarwood oil determines its quality [1, 4, 5, 17]. It inhibits aromatic terpenes in tree extractives [3, 5] and a highly complex mixture of volatile organic compounds and produces particular strong woody aroma. The combination of these fragrant compounds gives the plant a distinctive characteristic odor and is often used by humans in essential oil. [18]. This oil consists of a complex chemical constituent, namely hydrocarbons of sesquiterpene, oxygenated sesquiterpenes and derivatives of chromone. They are active compounds and considered significant [1, 3, 4, 8, 16, 19]. These compounds decide the quality of Agarwood oil. The names of these compounds are α-guaiene, β-agarofuran, α-agarofuran, 10-epi-eudesmol and valerianol. They are important compounds found in Agarwood oil of high quality. [1, 3, 4, 8, 9, 16].

There are some studies that use statistical analysis such as z-score [8]to determine the grades of Agarwood oil quality and machine learning algorithm such as ANN [8, 20], support vector classifier [21] and random forest [22] to confirm the grades. The z-score method is based on variations in the pattern of abundances of each compounds. Seven compounds namely β-agarofuran, α-agarofuran, 10-epi-ɤ-eudesmol, ɤ-eudesmol, longifolol, hexadecanol and eudesmolplay a major role in the grading of Agarwood oil quality[21].

β-agarofuran, α-agarofuran, 10-epi-ɤ-eudesmol and ɤ-eudesmolcontribute to the grading of high quality oil. Meanwhile,longifolol, hexadecanol and eudesmol contribute to the grading of low quality oil[8]. In the case of Agarwood oil samples, JBD and MA2were of high quality[8, 9] and CKE, HD and R5 were oflow quality[8]. Based on the Agarwood oil quality grading of z-score method, ANN and random forest have accurately classified the Agarwood oil as high and low with a smaller prediction error [8, 20, 22].

Normalization is a scaling technique and particularly useful for classifying algorithms such as neural networks, or distance measurements (nearest neighbor classification and clustering)[23-25]. In term of nearest neighbor classification, normalization can increase the value of accuracy compared to without normalization [26]. Min. max. normalization does a linear transformation on the initial data. Min-max normalization converts a value '*d*' of '*P*' to a new value '*d'*' in the range [*new_min(p), new_max(p)*]. Where *min(p)* and *max(p)*are the minimum and maximum value of attribute respectively. The min-max normalization is determined by (1) [23-25].

$$d' = \frac{d - min.(p)}{max.(p) - min.(p)} \\ * (new\ max.(p) - new\ min.(p)) \\ + new\ min.(p) \tag{1}$$

The self-organizing map (SOM) is an artificial neural network that uses a clustering algorithm to realize high-dimensional visualization [27]. It is also known as the Kohonen network that TeuvoKohonen introduced in 1981. The benefits of using SOM are [28]:

- Reduction of dimensions can be realized and the results of clustering can be easily interpreted. High dimensional input space can be converted to low dimensional output space thus preserving the initial topological relationship.
- It can be applied to cases that do not fully understand the characteristics of input data.
- The algorithm is clear and it's simple to calculate.

The neural network architecture of the SOM is shown in Figure 1. It comprises of the input and output layer (competition layer). The number of input layer neurons is calculated by the number of vectors in the input network, and the neurons position the output layer into a two-dimensional node matrix. Weights W completely link the neurons in the input layer to the neurons in the output layer [28]. Every neuron in the output layer can be viewed as a class or cluster representative of the inputs [27].
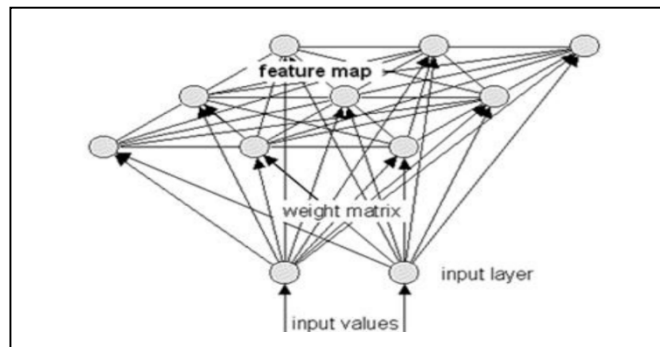

Figure 1: SOM Architecture [27]

A SOM is a competitive neural network that meets the principle of competitive learning rules. The winning neurons are classified according to the concept of minimal distance for all neurons in the output layer (competition layer), then the winning neurons and the winning neighborhood neurons are modified by the weight vector. Here are the measures for Kohonen's SOM [27, 28]:

i. The output layer weight vectors that are recorded as $\widehat{W}_j(j = 1, 2, 3, \dots, m)$ are given smaller and random numbers. m is the number of the output layer neurons.

ii. Normalize the input vector $X^p(j = 1, 2, 3, \dots, n)$ and the weight vector Wj in the competition layer as displayed in (2) and (3) where input data dimension isn. The normalized input vector is referred to as $\widehat{X}^p(j = 1, 2, 3, \dots, n)$, where n is input data dimension.

$$\hat{X} = \frac{X}{||X||} \tag{2}$$

$$\widehat{W}_j = \frac{W_j}{||W_j||} \tag{3}$$

iii. Calculate the distance between input $\hat{X}^p$ and all output neurons, and pick the winning neuron that has the smallest distance. Euclidean distance as displayed in (4) is selected for the distance formula.

$$d_j = ||\hat{X} - \widehat{W}_j|| = \sqrt{\sum_{j=1}^{m}(X - W_j)^2} \tag{4}$$

iv. The weight adjustment region of t time is determined by taking the winning node j* as the center. Setting the neighborhood distance d, the area within the d range of the distance from the winning node is considered the winning neighborhood.

v. The connection weights between the winning node and all other nodes in the winning neighborhood are modified based on gradient descent process. The modification is displayed by (5). Where $W_{ij}(t)$ denotes the weight of neuron i to neuron j at the time t; where $0<\alpha<1$ denotes the learning rate that influences the convergence and stability of the algorithm and decreases with time; where $N(j,t)$ represents the topological distance between the jth neuron and the winning neuron j* in the neighborhood at training time t.

$$W_{j*}(t + 1) = \widehat{W}_{j*}(t) + \Delta W_{j*}$$
$$= \begin{cases} \widehat{W}_{j*}(t) + \propto \left(\hat{X}^p - \widehat{W}_{j*}(t)\right); j = j^* \\ \widehat{W}_{j*}(t) + \propto N(t)\widehat{W}_{j*}(t); j \in N_j(t) \\ \widehat{W}_{j*}(t); j \notin N_j(t) \end{cases} \quad (5)$$

vi. The training process is completed when the learning rate $\alpha(t) \leq \alpha$ min or iterations $t \geq T$ were satisfied, otherwise return back to Step 2 and continue the iteration.

The results of SOM learning indicate the distance of neighboring neurons (SOM neighbor distances) [29, 30] and distribution of weight value (SOM weight planes) [29-32] and they are represented by color map. In SOM Neighbor Distances, the hexagons and red lines represent the neurons and connection of neurons respectively. Darker color represents greater distance and vice versa [29, 30]. In SOM Weight Planes, the association between the color and the weight of the output neuron is displayed in Figure 2. Lighter and darker colors represent larger and smaller weights, respectively [29-32]. If the connection patterns of two inputs are very similar (the shape and the color of neuron are the same for the two inputs), it can be concluded that the inputs were highly correlated [31, 32].
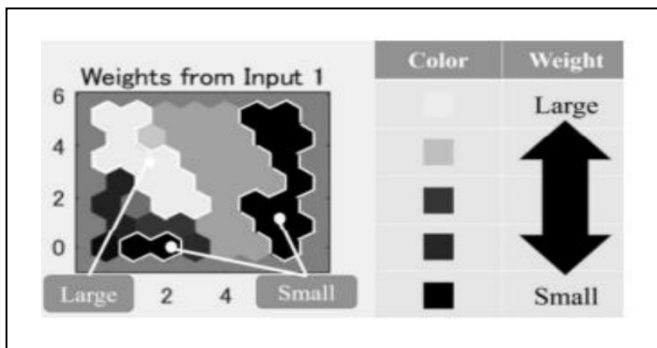

Figure 2: SOM Architecture [31]

Silhouette index (SI) is used to verify clusters by determining which objects are well positioned within their cluster and which objects are only somewhere between clusters [33]. Figure 3 displays two cluster: A and C.
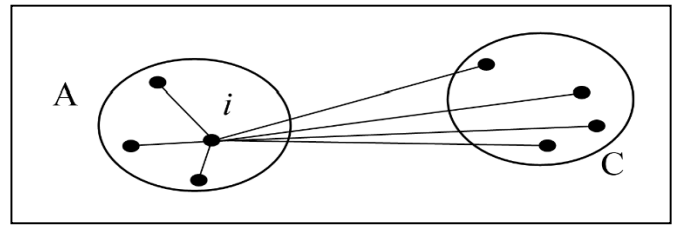

Figure 3: Computing Silhouette Index (SI) [33]

To determine the SI, take and mark an object in A as *i,* and then determine *a(i). a(i)* is the average dissimilarity *i* to all other objects of A. It shows the average length of all lines within A. Next, determine the *c(i,C). c(i,C)* is the average dissimilarity of *i* to all objects of C. It indicates the average length of all lines going from *i* (in cluster A) to C. After that, determine all *c(i,C)* values. Then, pick the smallest number and represent it by (6). The SI is summarized as in (7) [33-35].

$$b(i) = minimum \; c(i, C), C \neq A \quad (6)$$

$$SI(i) = \frac{b(i) - a(i)}{max.\,[a(i).\,b(i)]} \quad (7)$$

The SI range varies between −1 and 1. A value of SI close to 1 indicates that the object *i* is much closer to the other objects in the same cluster than the objects in the nearest cluster. A value of SI close to 0 means that the clustering of the focal object is questionable. A value of SI near −1 suggests a misclustering [34, 35]. The SI value can be translated as follows: "Excellent Split" if it is between 0.71 and 1.00. "Reasonable Split" if it is between of 0.51 and 0.70. "Weak Split" if it is between 0.26 and 0.5. "Bad Split" if it is less than 0.5 [33]. Averaging the SIs over a cluster provides an evaluation of the "goodness" of that cluster [34, 35].

The k-nearest neighbor (KNN) algorithm is based on the assumption that objects of the same class should be closer in distance. The KNN algorithm implementation process is as follows: first, apply an Euclidean distance equation to measure the distance in order to determine the number of nearest points of test data x against training data k. If $x = [x_1, x_2, \dots, x_k]$ and $y = [y_1, y_2, \dots, y_k]$ are two points in k dimensional space, the Euclidean distance between the two can be represented by (8):

$$d(x, y) = \sqrt{\sum_{i=1}^{k}(y_i - x_i)^2} \quad (8)$$

If test data x contains more representatives than a certain data category, it is assumed that x belongs to a certain category [36, 37]. Implementation of nearest neighbors techniques requires three variables: distance metric d, number of nearest neighbors k, and a scheme to weight individual neighbors when predictions are measured. The distance metric is used to evaluate similarity. The number of nearest neighbors is often

chosen from 1 to 10 and used when predictions are determined [37].

The results of classifier validation are often presented as confusion matrices. A confusion matrix for k class classification is a contingency table k x k whose cells [i,j] (i=1,2,...,k, j=1,2,...,k) present observations frequencies with actual class Ci and inferred class Cj. When there are only two classes, a binary confusion matrix is a special case: C (positive class) and not-C (negative class). It is always possible to represent a k x k confusion matrix as a set of k binary confusion matrices, one for each class, Ci [38]. In a binary confusion matrix, observations that were correctly classified into the positive class are considered True Positives (TP) and observations that were correctly classified into the negative class are called True Negatives (TN). Positive instances that were falsely classified as negative are considered False Negatives (FN) and negative instances that were falsely classified as positive are called False Positives (FP). Classification performance in terms of accuracy, precision, sensitivity and specificity can be measured using these classes. The formulas are displayed in (9), (10), (11), and (11) respectively [38, 39].

$$accuracy\,(acc.) = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (9)$$

$$precision\,(prec.) = \frac{(TP)}{(TP + FP)} \quad (10)$$

$$sensitivity\,(sen.) = \frac{(TP)}{(TP + FN)} \quad (11)$$

$$specificity\,(spe.) = \frac{(TN)}{(TN + FP)} \quad (12)$$

The study focuses on Agarwood oil quality grading using SOM. The objective is to determine the quality of Agarwood oil into high and low grade. The study consists of four sections: introduction, method, results and discussion, and conclusion. Introduction reviews the literature on Agarwood oil, SOM, KNN and confusion matrix. The method comprises of data acquisition and quality grading of Agarwood oil. The results obtained from method section are displayed and discussed in results and discussion section. In conclusion, the closure of this study will be presented.

2. METHOD

This section consists of two subsections: acquisition of data and grading of Agarwood oil quality. Under data acquisition, an analysis of Agarwood oil using GC-MS was conducted. Rescaling, determination and confirmation of grades and validation of Agarwood grading technique were carried out under subsection 2.2. Figure 4 shows the general methodology of this study.
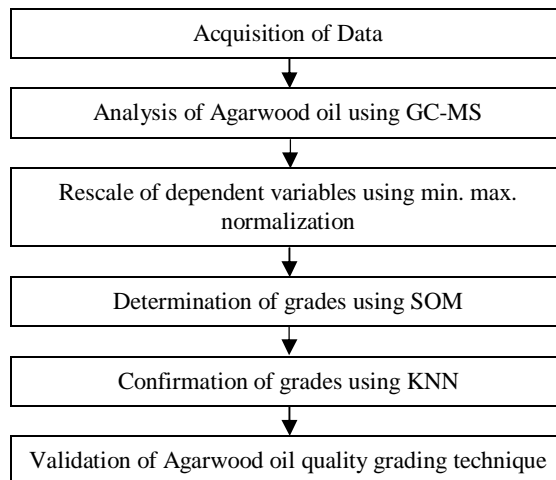

Figure 4: General methodology of the study

2.1 Data Acquisition

The samples of Agarwood oil used in this study are limited only to Aqualaria species. Thesampleswere obtained from Forest Research Institute Malaysia (FRIM) and Universiti Malaysia Pahang (UMP).A total of 660 samples with 22 primary samples wereusedfor this study. The primary samples were named as CKE, CM, EO2, EO3, EO4, HD, HG, JBD, KB, LA, LG, M, MA, MA1, MA2, MN, MNS, MPE, MS, R5, RG and T. Each sample contained103 chemical compounds. These chemical compounds were extracted and analyzed using GC-MS. The GC-MS apparatus settings were:

- The temperature of apparatus was initially set at 60º C for 10 min.
- The temperature ended at 230º C with an increment of 3º C per min.
- The flow rate of the helium gas carrier was 1ml per minute.
- The temperature of the ion source has been set at 280º C.

By matching them to the mass spectral library (HPCH2205.L; Wiley7Nist05a.L; NIST05a.L), the significant chemical compounds are identifiedwith the aid of a chemist.

2.2 Quality Grading of Agarwood Oil

This subsection consists of rescaling using min. max. normalization, determination of grades using SOM, confirmation of grades using KNN and validation of the Agarwood oil quality grading technique. All programs used in this study have been computed using MATLAB.

First, each dependent variables were rescaled between 0.01and 0.99 using min. max. normalization as in (1). Data with rescaled dependent variables were then clustered using SOM as shown in Figure 5. Out of 103 chemical compounds, only three were used as inputs for SOM. The three compounds were β-agarofuran, α-agarofuran and 10-epi-ɤ-eudesmol [40].

The output was the grade of Agarwood oil (high or low). The rules for SOM clustering were as follows:

- The ratio of training and testing: 80 to 20.
- 2 neurons: each neuron represents a cluster that is either high or low quality.
- Dimension: 1 by 2 grid
- Topology function: hextop (hexagonal pattern)
- Distance function: Euclidean distances
- coverSteps: 100
- initNeighbor: 1

These rules have been implemented into SOM algorithm. The data were first randomized per row. They were then divided into training and testing dataset with a ratio of 80 to 20. Next, each dataset was transposed. Before computing the SOM, the testing dataset was examined whether it contains all primary samples. After that, SOM properties such as dimension, coverSteps, initNeighbor, topology and distance function have been set. The datasets for training, testing and validation were then computed in order. Next, the silhouette values of each cluster were computed and inspected for negative values in training and testing. Only clusters with positive silhouette values were accepted. The program ended after the fulfilment of this condition.

SOM Algorithm

Input: data T, training data Tr, testing data Ts, validation data Tv, main samples M
Output: predicted clusterof training data Dtr, predicted cluster of testing data Dts, predicted cluster of validation data Dtv
1   **load** T
2   **while s**ilhouette values of Dtr$\leq$ 0 or Dts$\leq$ 0**do**
3   **While** Ts $\not\supset$ M**do**
4   randomize T
5   split T to Tr and Ts with 80 to 20 ratio
6   Tr' $\leftarrow$ Tr; Ts' $\leftarrow$ Ts
7   **end while**
8   set SOM properties: dimensions, coverSteps, initNeighbor, topology function, distance function
9   start training
10   start testing
11   **return** Dtr, Dts
12   calculate and plot silhouette values of Dtr, Dts
13   calculate average silhouette values of Dtr, Dts
14   **end while**
15   start validation
16   **return** Dtv

Figure 5: SOM Algorithm for clustering

Next, the data with the grades obtained by SOM were classified using KNN as shown in Figure 6. β-agarofuran, α-agarofuran, and 10-epi-eudesmol were used as inputs. The output was the grade of Agarwood oil (high or low). The rules for KNN classification were as follows:

- The ratio of training and testing: 80 to 20.
- Number of nearest neighbor: 5

- Distance: Euclidean distances
- Tie breaking algorithm: nearest (use the class with the nearest neighbor among tied groups)

These rules have been implemented into the KNN algorithm. The data were trained, tested and validated for five nearest neighbors. Confusion matrix was then computed for each kto obtain accuracy, precision, sensitivity and specificity.

KNN Algorithm

Input:SOMtraining data Tr, SOM testing data Ts, validation data Tv; SOM predicted cluster of training data Dtr, SOM predicted cluster of testing data Dts, SOM predicted cluster of validation data Dtv, number of nearest neighbor k
Output: predicted classof training data Ctr, predicted class of testing data Cts, predicted class of validation data Ctv
1   set KNN properties: k, distance metric, tie breaking algorithm
2   **for** each k **do**
3   training (Tr, Dtr)
4   testing (Ts, Dts)
5   validation(Tv, Dtv)
6   **return** Ctr, Cts, Ctv
7   **end for**
8   calculate accuracy, precision, specificity and sensitivity ofCtr. Cts, Ctv

Figure 6: KNN algorithm for confirmation of clusters

Finally, the SOM based Agarwood oil quality grading technique was validated using independent samples. Universiti Malaysia Pahang (UMP) provided two independent samples of Agarwood oil that were analyzed using GC-MS. They were labelled as UMP1 and UMP2 and their grades were high and low respectively. Data consisting of β-agarofuran, α-agarofuran and 10-epi-ɤ-eudesmol in both samples were rescaled between 0.01and 0.99 using min. max. normalization. These independent samples were computed as validation datasets as shown in Figure 5 and 6.

3. RESULTS AND DISCUSSION

This section consists of three sub-sections: determination of Agarwood oil grades using SOM, confirmation of grades using KNN and validation of Agarwood oil quality grading technique.

3.1 Determination of Agarwood Oil Grades using SOM

The results of weight distance between neurons, weight of compounds to neurons, silhouette values for training and testing, and assignment of neuron to Agarwood oil samples will be displayed and discussed in this subsection. Figure7 shows the weight distance between neurons. The neurons were portrayed by a blue hexagon. The neuron at the bottom was numbered as neuron 1 and the neuron at the top as neuron 2.. The color of the region between the neurons indicates the distance between neurons. Darker color was more distant and vice versa.. In Figure7, the color of the region was red. Therefore, the distance between the neurons is moderate.
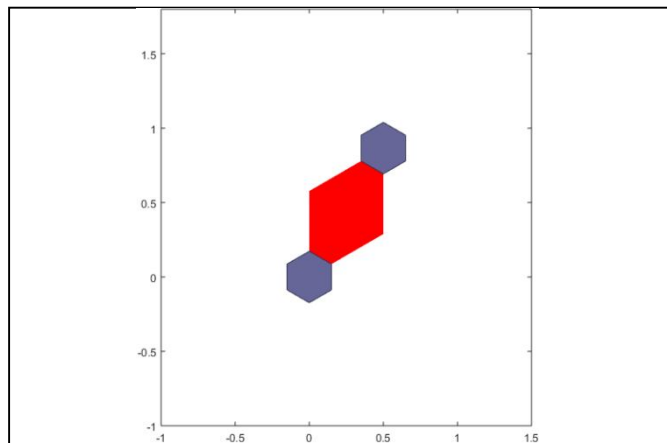
Figure 7: Weight distance between neurons

Figure8 shows the weight of compounds to neurons. The hexagons at the bottom and top represent the neurons labelled as neurons1 and 2. The colors of the neurons indicate the weight of compounds to the neurons. Lighter and darker colors represent the larger and smaller weight contribution respectively. As a result, the three compounds specifically β-agarofuran, α-agarofuran and 10-epi-ɤ-eudesmol contributed more to neuron 1 compared to neuron 2. The literatures have shown that these compounds are substantial compounds found in high-quality oil [1, 3, 4, 8, 9, 16]. Therefore, neuron 1 represents a high-grade cluster and neuron 2 a low-grade cluster. The connection pattern of these compounds was identical for all inputs. Neuron 1 was yellow, and neuron 2 was black. Thus, the compounds are highly correlated.
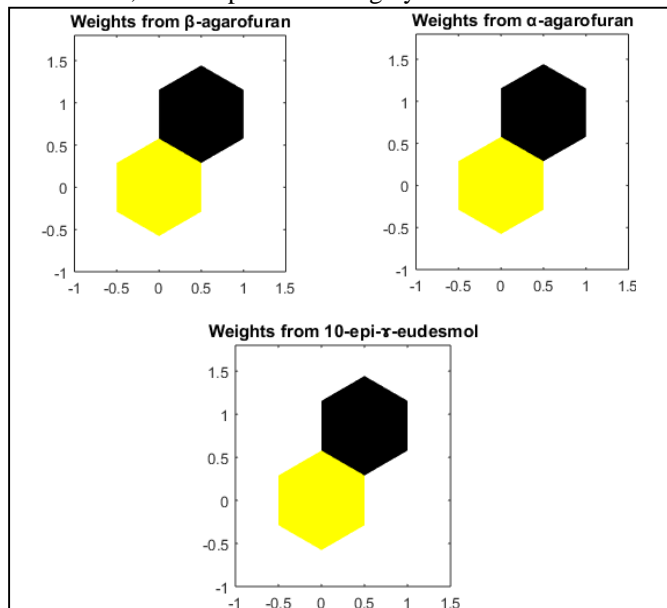

Figure 8: Weight of compounds to neurons

Figure 9 shows silhouette plots of each neuron for training (top) and testing (bottom). There were 528 training samples and 132 test samples. In training and testing, all samples had positive silhouette values. In training, the average silhouette values for neuron 1 (high grade cluster) and neuron 2 (low grade cluster) were 0.82 and 0.67 respectively. In testing, the average silhouette values for neuron 1 (high grade cluster) and neuron 2 (low grade cluster) were 0.79 and 0.58 respectively.

As a result, the high grade cluster (neuron 1) had better average silhouette values compared to the low grade cluster (neuron 2) in both training and testing. The plot indicates that the samples are well matched to their own cluster which is low and high grade and poorly matched to their neighboring cluster.
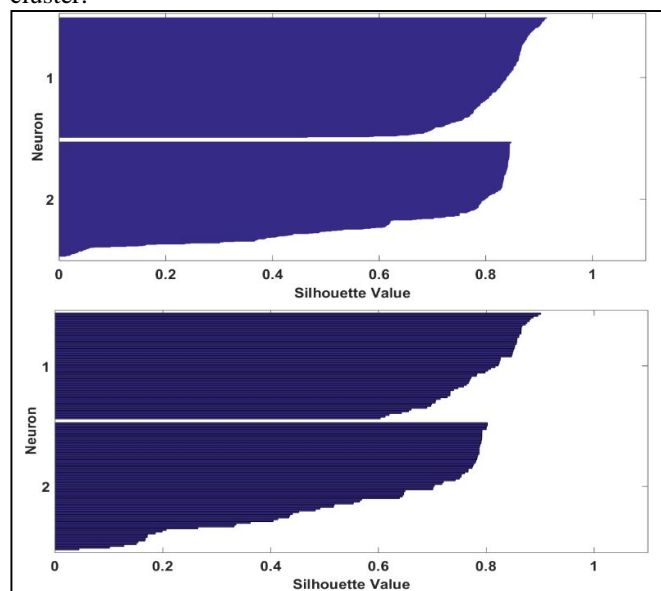

Figure 9: Silhouette plot for training and testing dataset

The assignment of primary samples to Agarwood oil grades for training and testing is shown in Table 1. For both training and testing, HG. JBD, KB, LA, MA, MA1, MA2, MNS, MPE, RG and T belonged to the high grade cluster (neuron 1). CKE, CM, EO2, EO3, EO4, HD, LG, M, MN, MS and R5 belonged to the low grade cluster (neuron 2). The numbers shown in the table are the number of samples.

Table 1: Assignment of major samples to neurons in training

Samples	Training		Testing	
	Neuron1 (high quality)	Neuron2 (low quality)	Neuron1 (high quality)	Neuron2 (low quality)
CKE		25		5
CM		23		7
EO2		21		9
EO3		21		9
EO4		23		7
HD		20		10
HG	27		3	
JBD	28		2	
KB	25		5	
LA	26		4	
LG		24		6
M		24		6
MA	26		4	
MA1	27		3	
MA2	22		8	
MN		27		3
MNS	23		7	
MPE	23		7	
MS		25		5
R5		25		5
RG	20		10	
T	23		7	

3.2 Confirmation of Agarwood Oil Grades using KNN

The result of the KNN confusion matrix for training and testing is shown in Table 2. There are a total of 528 training and 132 testing samples. In training, 270 and 258 samples were classified as high and low grades respectively. In testing, 60 and 72 samples were classified as high and low grades. There is no misclassification in both training and testing. As a result, KNN achieved100% accuracy, precision, sensitivity and specificity in training and testing for k=1 until k=5. It can be concluded that KNN has confirmed the grade of Agarwood oil determined by SOM.

Table 2: KNN confusion matrix for training and testing

	k	N=528	Predicted: HIGH	Predicted: LOW
Training	1	Actual: HIGH	270	0
		Actual: LOW	0	258
	2	Actual: HIGH	270	0
		Actual: LOW	0	258
	3	Actual: HIGH	270	0
		Actual: LOW	0	258
	4	Actual: HIGH	270	0
		Actual: LOW	0	258
	5	Actual: HIGH	270	0
		Actual: LOW	0	258
	k	N=132	Predicted: HIGH	Predicted: LOW
Testing	1	Actual: HIGH	60	0
		Actual: LOW	0	72
	2	Actual: HIGH	60	0
		Actual: LOW	0	72
	3	Actual: HIGH	60	0
		Actual: LOW	0	72
	4	Actual: HIGH	60	0
		Actual: LOW	0	72
	5	Actual: HIGH	60	0
		Actual: LOW	0	72

3.3 Validation of Agarwood Oil Quality Grading Technique

This subsection addresses the validation of Agarwood oil quality grading technique using two independent samples: UMP1 and UMP2. It is known from UMP that UMP1 has a better quality than UMP2.

The SOM assignment of independent samples to the grades of Agarwood oil is shown in Table 3. From the table, UMP1 belongs to the high grade cluster (neuron 1) and UMP2 to the low grade cluster (neuron 2). The grade determination shown here is the same as the grade determination from UMP. As a results, SOM has been able to determine the grades Agarwood oil using independent samples. The numbers shown in the table are the number of samples.

Table 3: SOM assignment of independent samples to neurons in the validation

Samples	Neuron1 (high quality)	Neuron2 (low quality)
UMP1	1	
UMP2		1

The confusion matrix for the KNN classification of independent samples is shown in Table 4. There was no misclassification from the confusion matrix in classifying each sample to the low and high grade from k=1 until k=5.As a result, KNN produced100% accuracy, precision, sensitivity and specificity for each k. It can be concluded that KNN has confirmed the grade of Agarwood oil determined by SOM

Table 4: KNN confusion matrix of independent samples

k	N=2	Predicted: HIGH	Predicted: LOW
1	Actual: HIGH	1	0
	Actual: LOW	0	1
2	Actual: HIGH	1	0
	Actual: LOW	0	1
3	Actual: HIGH	1	0
	Actual: LOW	0	1
4	Actual: HIGH	1	0
	Actual: LOW	0	1
5	Actual: HIGH	1	0
	Actual: LOW	0	1

4. CONCLUSION

Considering all the findings in results and discussion section, it can be concluded that, using only three chemical compounds, β-agarofuran, α-agarofuran, and 10-epi-π-eudesmol, the quality of Agarwood oil can be determined to be either high or low grade. This is consistent with the literatures[1, 3, 4, 8, 9, 16]. Furthermore, SOM is able to cluster the Agarwood oil into high and low grade with average silhouette values between 0.58 and 0.82, and KNN is capable of classifying the Agarwood oil into high and low grades with 100% accuracy, precision, sensitivity and specificity until five nearest neighbors. The SOM based grading technique was successfully validated using independent samples. To conclude, the quality grading of Agarwood oil into high and low grade based on β-agarofuran, α-agarofuran, and 10-epi-π-eudesmol using SOM has been proved to be effective and reliable. For future work, the quality of Agarwood oil can be extended into high, medium and low grades.

REFERENCES

1. 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**, *Jurnal Teknologi,* vol. 68, no. 1, 2014. https://doi.org/10.11113/jt.v68.2419

2. M. N. Azah, N. Ismail, J. Mailina, M. Taib, M. Rahiman, and Z. M. Hafizi, **Chemometric Study of Selected Agarwood Oils by Gas Chromatography—Mass Spectrometry**, *Journal of Tropical Forest Science,* pp. 382-388, 2014.

3. M. N. Azah, N. Ismail, J. Mailina, M. Taib, M. Rahiman, and Z. M. Hafizi, **Chemometric Study of Selected Agarwood Oils by Gas Chromatography-Mass Spectrometry**, *Journal of Tropical Forest Science,* pp. 382-388, 2014.

4. S. N. Ismail *et al.*, **Discriminative Analysis of Different Grades of Gaharu (Aquilaria malaccensis Lamk.) via 1H-NMR-Based Metabolomics Using PLS-DA and Random Forests Classification Models**, *Molecules,* vol. 22, no. 10, p. 1612, 2017. https://doi.org/10.3390/molecules22101612

5. M. H. Haron, M. N. Taib, N. Ismail, N. A. M. Ali, and S. N. Tajuddin, **Statistical Analysis of Agarwood Oil Compounds Based on GC-MS Data**, in *2018 9th IEEE Control and System Graduate Research Colloquium (ICSGRC)*, 2018, pp. 27-30: IEEE.

6. M. Nor Azah, Y. Chang, J. Mailina, S. Saidatul Husni, H. Nor Hasnida, and Y. Nik Yasmin, **Comparison of Chemical Profiles of Selected Gaharu Oils from Peninsular Malaysia**, *Malaysian Journal of Analytical Sciences,* vol. 12, no. 2, pp. 338-340, 2008.

7. N. Ismail, M. Ibrahim, S. Zareen, M. H. F. Rahiman, S. N. Tajuddin, and M. N. Taib, **Analysis of Agarwood Smoke Chemical Compounds using Solvent Trap, GC-FID and GC-MS**, *Pertanika Journal of Science and Technology,* vol. 25, pp. 1-8, 2017.

8. N. Ismail, N. A. M. Ali, M. Jamil, M. H. F. Rahiman, S. N. Tajuddin, and M. N. Taib, **Differentiating Agarwood Oil Quality Using Artificial Neural Network**, *Malaysian Journal of Analytical Sciences,,* vol. 17, no. 3, pp. 490-498, 2013.

9. N. Ismail, M. A. N. Azah, M. Jamil, M. H. F. Rahiman, S. N. Tajuddin, and M. N. Taib, **Analysis of High Quality Agarwood Oil Chemical Compounds by Means of SPME/GC-MS and Z-Score Technique**, *Malaysian Journal of Analytical Sciences,* vol. 17, no. 3, pp. 403-413, 2013.

10. M. Ishihara, T. Tsuneya, and K. Uneyama, **Components of the Agarwood Smoke on Heating**, *Journal of Essential Oil Research,* vol. 5, no. 4, pp. 419-423, 1993.

11. P. Chetpattananondh, **Overview of the Agarwood Oil Industry**, in *IFEAT international conference in Singapore*, 2012, pp. 131-8.

12. R. Mohamed and S. Y. Lee, **Keeping Up Appearances: Agarwood Grades and Quality**, in *Agarwood*: Springer, 2016, pp. 149-167. https://doi.org/10.1007/978-981-10-0833-7_10

13. M. S. Najib, M. N. Taib, N. A. M. Ali, M. N. M. Arip, and A. M. Jalil, **Classification of Agarwood Grades using ANN**, in *International Conference on Electrical, Control and Computer Engineering 2011 (InECCE)*, 2011, pp. 367-372: IEEE.

14. A. A. Rashid and Y. A. Zuhaidi, **Tapping the Wealth from Karas (Aquilaria Malaccensis) Tree**. Reka Cetak Sdn Bhd, 2011.

15. N. C. Radzi *et al.*, **Analysis of Sesquiterpenes in Agarwood Essential Oil from Hhydrodistillation Process**, *Malaysian Journal of Analytical Sciences,* vol. 22, no. 2, pp. 353-357, 2018.

16. N. Ismail, N. A. M. Ali, M. Jamil, M. H. F. Rahiman, M. N. Taib, and S. N. Tajuddin, **Major Volatile Chemical Compounds of Agarwood Oils from Malaysia Based on Z-Score Technique**, *Chemistry of Natural Compounds,* vol. 51, no. 4, pp. 776-779, 2015. https://doi.org/10.1007/s10600-015-1409-2

17. G. T. Pasaribu, T. K. Waluyo, and G. Pari, **Analysis of Chemical Compounds Distinguisher for Agarwood Qualities**, *Indonesian Journal of Forestry Research,* vol. 2, no. 1, pp. 1-7, 2016.

18. E. H. A. Latib, M. S. Najib, C. M. A. C. Mohd, and S. N. Tajuddin, **Analysis of Different Quality Agarwood Oil (Aquilaria Malaccensis) and Sensory Study**, *Journal of Telecommunication, Electronic and Computer Engineering (JTEC),* vol. 10, no. 1-3, pp. 57-61, 2018.

19. R. Mohamed, P. L. Jong, and A. K. Kamziah, **Fungal Inoculation Induces Agarwood in Young Aquilaria Malaccensis Trees in the Nursery**, *Journal of forestry research,* vol. 25, no. 1, pp. 201-204, 2014. https://doi.org/10.1007/s11676-013-0395-0

20. N. Zubir *et al.*, **Identification of Agarwood Oil Qualities using Radial Basis Function (RBF) Neural Network**, *Journal of Electrical & Electronic Systems Research* vol. 15, 2019.

21. M. A. H. Abas, N. Ismail, N. A. Ali, S. Tajuddin, and N. M. Tahir, **Agarwood Oil Quality Classification using Support Vector Classifier and Grid Search Cross Validation Hyperparameter Tuning**, *International Journal of Emerging Trends in Engineering Research,* vol. 8, no. 6, pp. 2551-2556, 2020. https://doi.org/10.30534/ijeter/2020/55862020

22. M. A. H. Abas, N. S. A. Zubair, N. Ismail, A. I. Mohd, S. N. T. Yassin, and M. N. Taib, **Classification of Agarwood Oil Quality Using Random Forest And Grid Search Crossvalidation**, *Journal of Electrical & Electronic Systems Research* vol. 15, 2019.

23. S. R. M-Dawam and K. R. Ku-Mahamud, **Reservoir Water Level Forecasting using Normalization and Multiple Regression**, *Indonesian Journal of Electrical Engineering and Computer Science,* vol. 14, no. 1, pp. 443-449, 2019.

24. O. Akbulut, **Feature Normalization Effect in Emotion Classification Based on EEG Signals**, *Sakarya Üniversitesi Fen Bilimleri Enstitüsü Dergisi,* vol. 24, no. 1, pp. 60-66, 2020. https://doi.org/10.16984/saufenbilder.617642

25. F. Zhang, I. Keivanloo, and Y. Zou, **Data Transformation in Cross-Project Defect Prediction**,

Empirical Software Engineering, vol. 22, no. 6, pp. 3186-3218, 2017.

26. D. A. Anggoro and N. D. Kurnia, **Comparison of Accuracy Level of Support Vector Machine (SVM) and K-Nearest Neighbors (KNN) Algorithms in Predicting Heart Disease**, *International Journal of Emerging Trends in Engineering Research,* vol. 8, no. 5, pp. 1689-1694, 2020.

27. H. D. Wijaya and S. Dwiasnati, **Clustering of Earthquake Data Using Kohonen Self Organizing Maps (SOMs) Algorithm**, *Scholars Bulletin,* pp. 383-388, 2019.

28. B. Xia *et al.*, **Using Self Organizing Maps to Achieve Lithium-Ion Battery Cells Multi-Parameter Sorting Based on Principle Components Analysis**, *Energies,* vol. 12, no. 15, pp. 2980-2996, 2019.

29. K. Saxena, S. Jaloree, R. Thakur, and S. Kamley, **Self-Organizing Map (SOM) based Modelling Technique for Student Academic Performance Prediction**, *Int J Future Revol Comput Sci Commun Eng,* vol. 3, no. 9, pp. 115-120, 2017.

30. H. Yao, H. Yang, A. Zhang, C. Fang, and Y. Guo, **WLAN Interference Self‐Optimization using SOM Neural Networks**, *Concurrency and Computation: Practice and Experience,* vol. 29, no. 3, p. e3913, 2017.

31. Y. Wakabayashi, K. Chanthasen, S. Yoshida, Y. Kawabata, T. Ogawa, and T. Sugibayashi, **Texture Evaluation of Transparent Acrylic Resin using Self-Organizing Maps**, in *Proceedings on the International Conference on Artificial Intelligence (ICAI)*, 2018, pp. 289-292: The Steering Committee of The World Congress in Computer Science, Computer ….

32. O. Murad and M. Malkawi, **An Optimization Method Using Clustering Technique for the Human Emotions Detection Artificial Neuro-Fuzzy Logic System**, *International Journal of Computers & Technology,* vol. 15, no. 9, pp. 7090-7096, 2016. https://doi.org/10.24297/ijct.v15i9.695

33. A. R. Mamat, F. S. Mohamed, M. A. Mohamed, N. M. Rawi, and M. I. Awang, **Silhouette Index for Determining Optimal k-Means Clustering on Images in Different Color Models**, *International Journal of Engineering and Technology,* vol. 7, no. 2.14, pp. 105-109, 2018.

34. A. Lengyel and Z. Botta‐Dukát, **Silhouette Width using Generalized Mean—A Flexible Method for Assessing Clustering Efficiency**, *Ecology and Evolution,* vol. 9, no. 23, pp. 13231-13243, 2019.

35. C. Subbalakshmi, G. R. Krishna, S. K. M. Rao, and P. V. Rao, **A Method to Find Optimum Number of Clusters based on Fuzzy Silhouette on Dynamic Data Set**, *Procedia Computer Science,* vol. 46, pp. 346-353, 2015.

36. C.-H. Chen, W.-T. Huang, T.-H. Tan, C.-C. Chang, and Y.-J. Chang, "**Using k-nearest neighbor classification to diagnose abnormal lung sounds**," *Sensors,* vol. 15, no. 6, pp. 13132-13158, 2015. https://doi.org/10.3390/s150613132

37. G. Chirici *et al.*, **A Meta-Analysis and Review of the Literature on the k-Nearest Neighbors Technique for Forestry Applications that Use Remotely Sensed Data**, *Remote Sensing of Environment,* vol. 176, pp. 282-294, 2016.

38. S. Ruuska, W. Hämäläinen, S. Kajava, M. Mughal, P. Matilainen, and J. Mononen, **Evaluation of the Confusion Matrix Method in the Validation of an Automated System for Measuring Feeding Behaviour of Cattle**,*Behavioural processes,* vol. 148, pp. 56-62, 2018. https://doi.org/10.1016/j.beproc.2018.01.004

39. S. Aljawarneh, M. Aldwairi, and M. B. Yassein, **Anomaly-Based Intrusion Detection System through Feature Selection Analysis and Building Hybrid Efficient Model**, *Journal of Computational Science,* vol. 25, pp. 152-160, 2018. https://doi.org/10.1016/j.jocs.2017.03.006

40. M. H. Haron, M. N. Taib, N. Ismail, N. A. Mohd Nor, and S. N. Tajuddin, **Determination of Substantial Chemical Compounds of Agarwood Oil for Quality Grading**, *Journal of Electrical & Electronic Systems Research,* to be published.