The k-nearest neighbor modelling by varying Mahalanobis and correlation in distance metric for agarwood oil quality classification

Noor Syafina Mahamad Jainalabidin[1], Aqib Fawwaz Mohd Amidon[1], Nurlaila Ismail[1], Zakiah Mohd Yusoff[1], Saiful Nizam Tajuddin[2], Mohd Nasir Taib[3]
[1]School of Electrical Engineering, College of Engineering, Universiti Teknologi MARA, Shah Alam, Malaysia
[2]Bioaromatic Research Centre, Universiti Malaysia Pahang, Pahang, Malaysia
[3]Malaysia Institute of Transport (MITRANS), Universiti Teknologi MARA, Shah Alam, Malaysia

Article Info

Article history:

Received Mar 22, 2022
Revised Apr 13, 2022
Accepted May 19, 2022

Keywords:

Agarwood oil
Classification
k-nearest neighbor
Mahalanobis

ABSTRACT

Agarwood oil is well known for its unique scent and has many usages; as an incense, as ingredient in perfume, is burnt during religious ceremonies and is used in traditional medical preparation. Therefore, agarwood oil has high demand and is traded at different price based on its quality. Basically, the oil quality is classified by using physical properties (odor and color) and this technique has several problems: not consistent in term of accuracy. Thus, this study presented a new technique to classify the quality of agarwood oil based on chemical properties. The work focused on the k-nearest neighbor (k-NN) modelling by varying Mahalanobis and correlation in distance metric for agarwood oil quality classification. It involved of 96 samples of agarwood oil, data pre-processing (data randomization, data normalization, and data division to testing and training datasets) and the development of k-NN model. The training dataset is used to train the k-NN model, and the testing dataset is used to test the developed model. During the model development, Mahalanobis and correlation are varied in k-NN distance metric. The k-NN values are ranging from 1 to 10. Several performance criteria including resubstitution error (closs), cross-validation error (kloss) and accuracy were applied to measure the performance of the built k-NN model. All the analytical work was performed via MATLAB software version R2020a. The result showed that the accuracy of Mahalanobis distance metric has a better performance compared to correlation from k = 1 to k = 5 with the value of 100.00%. This finding is important as it proved the capabilities of k-NN modelling in classifying the agarwood oil quality. Not limited to that, it also contributed to the agarwood oil research area as well as its industry.

Corresponding Author:

Nurlaila Ismail
School of Electrical Engineering, College of Engineering, Universiti Teknologi MARA
40450 Shah Alam, Selangor, Malaysia
Email: nurlaila0583@uitm.edu.my

1. INTRODUCTION

Agarwood, also called '*Gaharu'*, is a resin-impregnated, fragrant, and highly valuable heartwood found in Aquilaria plants [1], [2]. Agarwood oil refers to the oil extracted from agarwood trees. Agarwood oil has been used in a wide variety of usages including as an incense, in perfumery ingredients, been burned during religious ceremonies and been used in traditional medicinal preparations [3]. The agarwood (known as '*kalambak*') is used in traditional Malay medicinal preparation to treat various illnesses such as fatigue, stomach or chest pain, oedema, as a tonic for men and women and postpartum medicine [4].

Researchers found that there are different qualities of agarwood oil [5]. This quality reflects the price of the oil traded in the market. High quality is sold at high price and low quality is sold at low price [5]. Currently, people use their experiences in identify the quality of agarwood oil based on its physical properties such as color and odor. Dark color and long-lasting scent are considered as high quality and been sold at a premium price. A high-quality oil usually cost between USD126 and USD633 per tola (12 ml) [6]. At the same time, wood prices are USD19 per kg for low quality and USD100,000 per kg for superior quality [6], [7].

The difference of agarwood oil quality is due to different chemical compounds exist in the oil. Researchers from Japan discovered several chemical compounds appear in high quality of agarwood oil such as β-agarofuran, α-agarofuran, γ-eudesmol and 10-epi-γ-eudesmol [8]. On the other hand, longifolol and hexadecanol usually are synonym to the low quality of agarwood oil [9].

There is a recommendation for the agarwood oil to be graded based on its chemical properties due to the limitations presence when the oil is grading based on its physical properties [10]. The limitations are human eye and nose cannot deal with a bulk sample of oil in one time. Then, it resulted to a non-consistent grading. Not limited to that the result of grading based on human sensory panel also varies due to different expertise and experiences [11]. Subjectivity, low reproducibility and time consuming are all disadvantages of current grading method using human sensory panel [12]. Thus, it is important to have an intelligent technique to classify the quality of agarwood oil.

Several methods have proven their successes in classifying or discriminating of different groups essential oil [13]. The methods are artificial neural network (ANN), adaptive neuro-fuzzy inference system (ANFIS), support vector machine (SVM) and k-nearest neighbor (k-NN) [13]–[17]. Among all, k-NN was the most preferred due to its efficiency in classifying and selecting samples under study to their respective groups [14]–[17].

As part of ongoing research in classifying the agarwood oil quality using its chemical properties, the objective of this study is to develop the k-NN modelling by varying Mahalanobis and correlation in distance metric for agarwood oil quality classification [18]. In addition, there is lacking research related to distance metric especially for Mahalanobis and correlation during the k-NN model development. Thus, this study is very significant and contributed to the agarwood oil grading research area.

The k-NN is the most widely used classification method, and it is one of the top ten most common and important algorithms [14]–[17]. It is known for being simple and straightforward. This algorithm is a classification algorithm that is often used. Hence, algorithm's goal is to identify new objects using their attributes and training information [19]. So, analysis based on modelling by using Mahalanobis and correlation of k-NN distance metric to ensure that agarwood oil result based on the quality specifications that measured.

This paper was organized in the following manners: section 1 is introdustion, section 2 is literature review and theoretical work, section 3 explains the methodology of this research and section 4 is the results and discussion and finally section 5 is the conclusion of the study.

2. LITERATURE REVIEW AND THEORETICAL WORK

2.1. The k-NN algorithm

Figure 1 shows the algorithm known as k-NN has the ability to identify, classify and categorize nonparametric components, such as hospitality signal, transformer winding, network, essential oils for daily use and electroencephalography (EEG) signals [20]–[23]. Non-parametric means that there are no parameters or that there are a fixed number of parameters regardless the size of the dataset. However, the size of the training dataset would be determining the parameters.

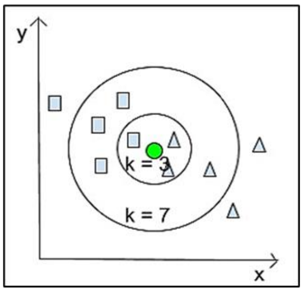

Figure 1. The concept of k-NN technique [21]

Next, k-NN undoubtedly has been recognized among the top ten methods in data mining [24]. Both the training and classification stages are provided in the k-NN basic [25]. The feature vectors and class labels of training samples are stored in a multidimensional feature space as a starting phase while in the following phase, all the data in training set was labelled then being used to sort into unlabeled data in the testing set. Hence, the distance between data enumerates and the nearest distance is selected to represent the number of k while the n-numeric attributes are used to describing training tuples and n-dimensional space is used to store all training examples [21]. Usually, the ratio between training data and testing is equal to 80:20% [26]–[28].

The main advantage of using k-NN is that it does not require assumptions about the underlying data distribution. In addition, k-NN also known as the laziest learning approach since it involves storing all training datasets and delaying until the test data is required to be created without being forced to create a learning model. The proper classification method of the agarwood oil quality can be proposed using machine learning model k-NN and widely implemented algorithms as shown in (1) [23].

$$d(x, y) = \sqrt{\sum_i^n (x_i - y_i)^2} \tag{1}$$

Where,
x,y = two points of object for n-space
x_i, y_i = vector that starting from the origin of the space (initial point)
n = n-space

2.2. Distance metric

2.2.1. Mahalanobis metric

The Mahalanobis distance metric calculates the straight-line distance between two points as it is the most common distance metric choice in NN algorithms [29]. However, the distance measure shows the relationship between different characteristics of a given object compared to the Chebyshev distance metric. The Mahalanobis distance can be used when dealing with problems such as clustering analysis, hypothesis testing, goodness of fit tests, classification techniques, outlier detection, and density estimation methods is of great importance. Equation (2) shows the Mahalanobis distance formula, assuming that the covariance matrix of x and y is represented as S [21].

$$d(x, y) = D^2 = \sqrt{(x_i - y_i)^T S^{-1} (x_i - y_i)} \tag{2}$$

Where,
D^2 = square of Mahalanobis distance
x_i = vector of the observation (row in a dataset)
y_i = vector of mean values of independent variables (mean of each column)
S^{-1} = the inverse covariance matrix of independent variables

The components of the vector 'x' and vector 'y' are represented by 'xi' and 'yi'. To overcome the problems of correlation and scale, Mahalanobis distance metric is calculated using the group of mean and variance of each variable. However, it cannot be calculated if the variables are highly correlated [21].

2.2.2. Correlation distance metric

The correlation distance metric measures both non-linear and the linear association or intensity between two random variables or degree of association. A symbolic attribute values would be considered as a random variable if given a set of 'n' for instances. Furthermore, correlation coefficients can be used between binary variables (classical statistics) which as a computation of the connection between two symbolic values [13]. As it goes beyond Pearson' correlation because it can detect more than linear associations while working in multi-dimensionally. The formula of correlation distance metric defined by (3).

The components of the vector 'x' and vector 'y' are represented by 'xi' and 'yi'. To overcome the problems of correlation and scale, Mahalanobis distance metric is calculated using the group of mean and variance of each variable. However, it cannot be calculated if the variables are highly correlated [17].

$$d_{st} = \frac{\sum (x_s - x'_s)(y_t - y'_t)}{\sqrt{\sum (x_s - x'_s)^2 (y_t - y'_t)^2}} \tag{3}$$

Where,
d_st = correlation coefficient of linear relationship between x and y

x_s = values of the x-variable in a sample
x'_s = mean of the values of the x-variable
y_t = values of the y-variable in a sample
y'_t = mean of the values of the y-variable

It ranges from 0 to 1, where 0 implies independence between x & y and 1 implies that the linear subspace of x & y is equal. However, the correlation distance is not the correlation between the distance itself, but it is a correlation between the scalar products which the "double centered" matrices are composed of.

2.3. Performance criteria

2.3.1. Resubtitution error (closs)

"Closs" which are the training error of the system also known as resubstitution error (apparent error) that describe on how the training set have been modified. It also evaluates the error rate based on outcome (output) versus the actual value from the same training dataset [30]. In other words, it will be derived by applying a model to the training dataset from which it was learned.

2.3.2. Cross-validation error (kloss)

Cross-validation error (kloss) is a great method for evaluating a model's prediction performance. However, a model may reduce the mean squared error on the training set data, but a cross-validation might be more optimistic in its predictive error by observing on split partitions [31], [32]. The most common kind of cross-validation is K-Fold, in which observations are split into K-partitions; the model is trained on K–1 partition, and the test error is predicted on the left partition K [32]. For example, the process will be repeated for k = 1, 2, 3, 4, …, K and the results is average. This kind of approach has lower bias where it is computationally affordable, but the estimation of each fold is highly correlated [32].

2.3.3. Accuracy (%)

The performance measure used in this research is the confusion matrix. Classification accuracy itself can be deceptive if the model dataset has a different number of observations in each class or have more than two classes in the dataset. The matrix row in the confusion matrix will donates the literal class while matrix column displays the expected class [33]. There are four elements that involve in the confusion matrix which are the number of correctly classified to the class examples, the number of correctly identified to the not class examples, the example of incorrectly identified to the class examples, and the example of incorrectly identified to the not class examples. The accuracy is the overall effectiveness of a classifier. As for the formula, it can be defined in (4) [28].

$$Accuracy = \frac{Cc+Cn}{Cc+In+Ic+Cn} \tag{4}$$

Which expressed as:
Cc is the number of correctly classified to the class examples
Cn is the number of correctly identified to the not class examples
Ic is the example of incorrectly identified to the class examples
In is the example of incorrectly identified to the not class examples

3. RESEARCH METHOD

The agarwood oil data collection used in this study is obtained from two institutions; the Forest Research Institute Malaysia (FRIM), Malaysia and BioAromatic Research Centre of Excellence (BARCE), University Malaysia Pahang (UMP), [34]. The work involved of 96 samples of agarwood oil, data pre-processing (data randomization, data normalization, and data division (testing and training datasets) and k-NN model development. The training dataset is used to train the k-NN model, and the testing dataset is used to test the develop model. During the model development, the distance metric is varied by using Mahalanobis and correlation. The k-NN with learning algorithms are used as classifier to classify the quality of the agarwood oil. The k-NN values are ranging from 1 to 10. Several performance criteria including resubstitution error (closs), cross-validation error (kloss) and accuracy (%) are applied in measuring the performance of the built k-NN model.

3.1. The k-NN modelling

Figure 2 shows the flowchart of the k-NN modelling. Firstly, it is the data collection consists of input and output data which are chemical compounds and oil quality of agarwood oil. Next, it is continued with the data pre-processing. At this stage, the data is normalized, randomized and divided into two groups: training and testing (with the ratio of 80% and 20%, respectively). Then, the process is followed by the k-NN model development using training dataset. During the development, distance metrics of k-NN are varied; Mahalanobis and correlation. For each learning algorithm, the value of 'k' is ranging from 1 to 10 as recommended by [35]. After that, the developed k-NN model is tested by using the tested dataset [36].

Next, in order to accept the developed k-NN model, several performance criteria are used. They are accuracy (%), closs, and kloss. If the model passed all these criteria, it can be accepted, and it is meaning that the model able to classify the agarwood oil quality to high and low. If it not passed (usually above 80% of accuracy [34]), the development of k-NN model have to return to the earlier stage either from data pre-processing or data training. All this analytical work is performed via MATLAB software version R2020a.

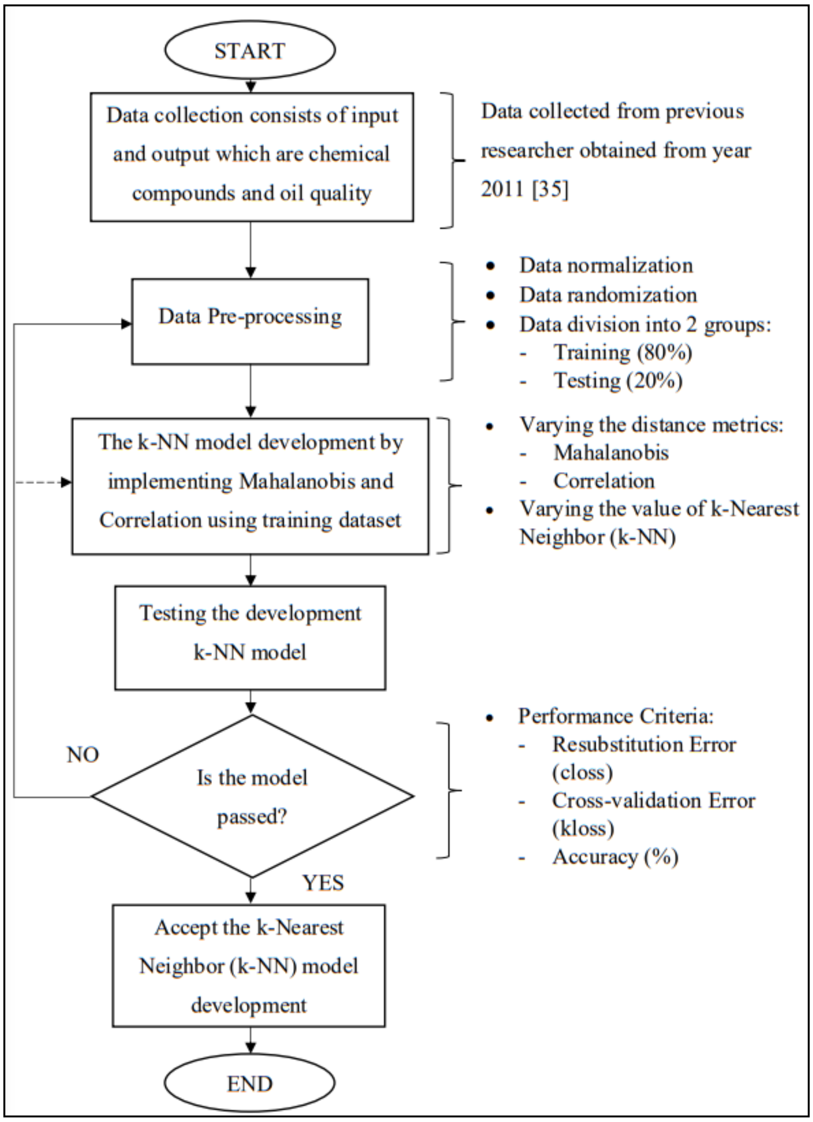
START
Data collection consists of input and output which are chemical compounds and oil quality
Data collected from previous researcher obtained from year 2011 [35]
Data Pre-processing
• Data normalization
• Data randomization
• Data division into 2 groups:
- Training (80%)
- Testing (20%)
The k-NN model development by implementing Mahalanobis and Correlation using training dataset
• Varying the distance metrics:
- Mahalanobis
- Correlation
• Varying the value of k-Nearest Neighbor (k-NN)
Testing the development k-NN model
NO
Is the model passed?
• Performance Criteria:
- Resubstitution Error (closs)
- Cross-validation Error (kloss)
- Accuracy (%)
YES
Accept the k-Nearest Neighbor (k-NN) model development
END

Figure 2. Flowchart of the classification model designed

3.2. The hierarchy for k-NN algorithm

The k-NN is a definition of an object that is graded in multidimensional space by the class of its k-NN in the test set. Two steps consist of the k-NN algorithm. The first was the preparation level and the second was the classification test [27]. The ratio of training and testing datasets usually equivalent to 80% and 20% [28]. The hierarchy for k-NN algorithm was described as follows:

Step 1: The k-value model development was identified. Where ‘k’ is the value number of rhe nearest or closest neighbor in multidimensional space of interest.
Step 2: Find value ‘k’ for training dataset. Here, the distance metric of Cosine and Mahalobis plays a rule and the implementation nearest rules in k-NN.
Step 3: The distance for all objects in the training dataset are sorted.
Step 4: The nearest neighbor based on the k-th minimun distance was determined.
Step 5: All the classes of the training dataset for the sorted values belong to k-model development were collected gathered.
Step 6: The majority of the nearest neighbor as prediction values was used.

3.3. Performance criteria for k-NN modelling

The developed k-NN model is tested its performances by using closs, kloss, and accuracy (%). The closs and kloss is calculated by using MATLAB script ‘`resubLoss`’ and ‘`crossval`’, respectively. Then, the accuracy (%) is calculated using in (4) as in section 2 literature review. These performance measures are done for Mahalanobis and correlation during k-NN modelling, and their results are discussed.

4. RESULTS AND DISCUSSION

This section presented the results of k-NN modelling for agarwood oil quality by varying its distance metric which are Mahalanobis and correlation. Figure 3 shows the accuracy of Mahalanobis and correlation distance metric. In this figure, from k = 2 to k = 10 the accuracy of Mahalanobis is higher than correlation. At k = 1 the accuracy of Mahalanobis is same as correlation. It means that different k-value of neighbor produce different accuracy of develop of k-NN model in classifying the agarwood oil quality. Then, the detail of the accuracy value for both Mahalanobis and correlation are tabulated in Table 1.

From Table 1, the accuracies of Mahalanobis and correlation distance metric in the developed k-NN models were gathered. For Mahalanobis; at k = 1 until k = 5 the accuracy is 100.00%, resulted as the maximum percentage. Then, the accuracy for k = 10 is 93.75%, resulted as the minimum percentage. The rest, at k = 6 until k = 9 the accuracy is 98.96%. Next, the accuracy for correlation distance metric is obtained. It can be seen that at k = 1, the accuracy is 100.00%, which is the maximum percentage. Then, at k = 2, k = 3, and k= 4 the accuracy is 97.92%. Following, at k = 5 the accuracy is 98.96% then at k = 6 the accuracy is 96.88% whereas at k = 7 is 95.83%. Meanwhile k = 8 and k = 9 the accuracy is similar which is 94.79%. Lastly, at k = 10 shows an accuracy of 92.71%.

Table 1. Accuracy of k-NN by using Mahalanobis and correlation from k = 1 to k =10

Parameter	Accuracy (%)	
(k-value)	Mahalanobis	Correlation
1*	100.00	100.00
2	100.00	97.92
3	100.00	97.92
4	100.00	97.92
5	100.00	98.96
6	98.96	96.88
7	98.96	95.83
8	98.96	94.79
9	98.96	94.79
10	93.75	92.71

Notes: *is the highest value for the accuracy of Mahalanobis and correlation

The accuracies of Mahalanobis and correlation as discussed in Table 1 was plotted for graphical observation and it is shown in Figure 3. By quick visual inspection, it can be observed that for k = 1, the accuracy is similar (100.00%) and after that from k = 2 to k = 10 the accuracy of Mahalanobis is better than correlation. The closs and kloss are used in determining the performance of Mahalanobis and correlation distance metric in the developed k-NN models. Table 2 shows the closs and kloss for Mahalanobis distance metric. In general, the resubstitution error from k = 1 until k = 5 is 0.0000 while from k = 6 until k = 9, has similar error which is 0.0104. Besides, the resubstitution error for k = 10 is 0.0625.

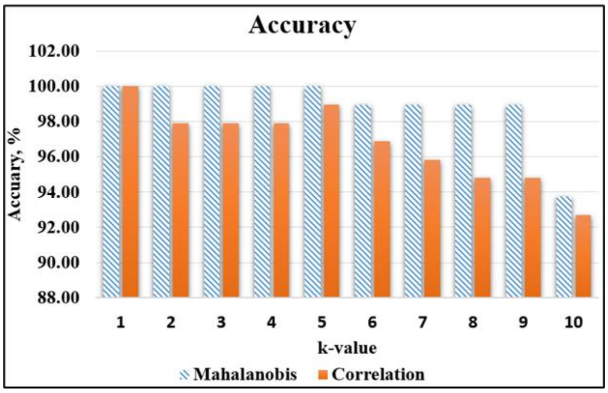

Figure 3. Accuracy of Mahalanobis and correlation distance metric

Table 2. Closs and kloss for Mahalanobis distance metric from k = 1 to k = 10

Parameter (k-value)	Mahalanobis Resubstitution error (closs)	Cross-validation error (kloss)
1	0.0000*	0.0000#
2	0.0000	0.0000
3	0.0000	0.0000
4	0.0000	0.0000
5	0.0000	0.0104
6	0.0104	0.0104
7	0.0104	0.0313
8	0.0104	0.0521
9	0.0104	0.0625
10	0.0625	0.0938

Notes: * is the lowest value for the closs, # is the lowest value for kloss

Furthermore, the cross-validation error in the table follows shows that the result from k = 1 until k = 4 was maintained in 0.0000 as well as from k = 5 and k = 6 with the kloss of 0.0104. For the next kloss, the result showed that at k = 7 is 0.0313, at k = 8 is 0.0521, at k = 9 is 0.0625, and at k = 10 is 0.0938. Hence, it is shown that the maximum value between the closs and kloss were 0.0625 and 0.0938 at k = 10. While the minimum value for both errors were same which is 0.0000. All these errors can be observed graphically in Figure 4.

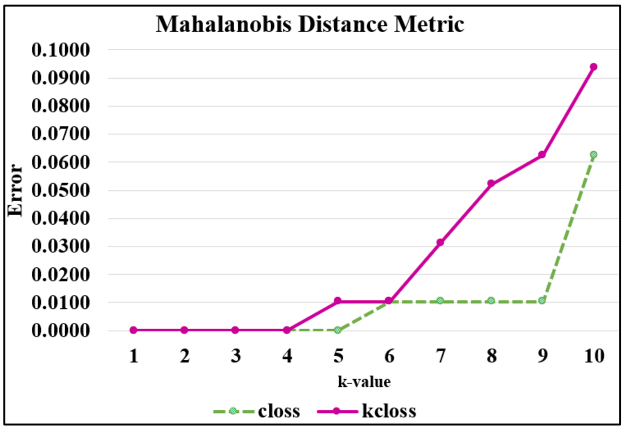

Figure 4. Closs and kloss for Mahalanobis distance metric

Table 3 shows the closs and kloss for correlation distance metric. The closs at k = 1 shows an error of 0.0000 while at k = 2, k = 3, and k = 4 shows closs of 0.0208. Next, the closs at k = 5, k = 6, and k = 7 shows an error of 0.0104, 0.0313, and 0.0417 respectively whereas k = 8 and k = 9 have the same resubstitution error which is 0.0521. Following, the resubstitution error at k = 10 is 0.0729. The maximum error for resubstitution error is 0.0729 at k = 10 and the minimum closs is 0.0000 at k = 1. Moreover, the klossr at k = 1 and k = 8 have error of 0.0521. Then, the kloss at k = 2 and k = 7 is 0.0417 and 0.0729 respectively. Furthermore, at k = 3 and k = 4 has the same kloss of 0.0208 while k = 5 and k = 6 value model both has kloss of 0.0625. In addition, k = 9 and k = 10 shows cross-validation error of 0.0833. Hence, the maximum error for kloss is 0.0833 at k = 9 and k = 10 on the other hand the minimum kloss is 0.0208 at k = 3 and k = 4. All these errors can be observed graphically in Figure 5.

Table 3. Closs and kloss for correlation distance metric from k = 1 to k = 10

Parameter	Correlation	
(k-value)	Resubstitution error (closs)	Cross-validation error (kloss)
1	0.0000*	0.0521
2	0.0208	0.0417
3	0.0208	0.0208#
4	0.0208	0.0208
5	0.0104	0.0625
6	0.0313	0.0625
7	0.0417	0.0729
8	0.0521	0.0521
9	0.0521	0.0833
10	0.0729	0.0833

Notes: * is the lowest value for the substitution error (closs), # is the lowest value for cross-validation error (kloss)

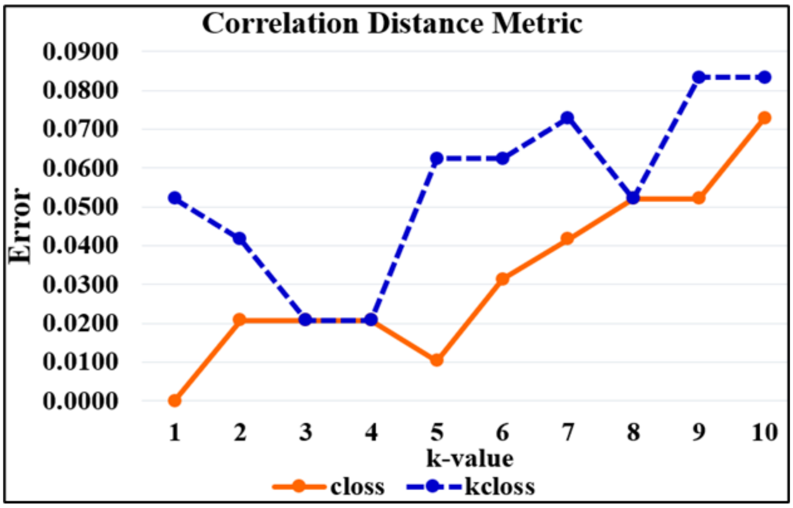

Figure 5. Closs and kloss for correlation distance metric

In summary, Mahalanobis is better than correlation in k-NN model to classify the agarwood oil quality. This is due to the concept of correlation which is a statistical measurement which its distance is computed from one minus the sample linear correlation between observations. Hence, the distance based on correlation is a measure of statistical dependence between two vectors. Next, the performance of Mahalanobis due to its concept as one of the distance metrics which is an effective multivariate distance metric to measure the distance between a point and a distribution. It is an extremely useful metric having, excellent applications in multivariate anomaly detection, classification on highly imbalanced datasets and one-class classification. If the variables in the dataset are strongly correlated, then, the covariance will be high. Dividing by a large covariance will effectively reduce the distance. Moreover, the samples of agarwood oil quality are not correlated, then the covariance is not high, and the distance is not reduced much. Hence, it supports the Mahalanobis to perform better than correlation. Thus, this study is very significant and contributed to the agarwood oil grading research area. Hence, the finding is in line agreement with the study as reported by Verdier [22].

5. CONCLUSION

This study has successfully presented the k-NN modelling by varying Mahalanobis and correlation in distance metric for agarwood oil quality classification as well as accomplished the objectives. The Mahalanobis and correlation distance metric were implemented during the k-NN modelling as well as the k-value from 1 to 10 is varied. After that, the developed k-NN models were tested using three performances criteria which are closs, kloss, and accuracy (%). The finding showed that for k = 1, the accuracy is similar (100.00%) and after that from k = 2 to k = 10 the accuracy of Mahalanobis is better than correlation. It accompanied by the errors of resubstitution and cross-validation, which errors for both are very small, i.e. close to 0.0000. It proved that the k-NN modelling developed in this study is capable in classifying the agarwood oil quality. This technique is significant as it used the chemical properties of the oil in its classification. Furthermore, it contributed and benefited to the agarwood oil industry especially to oil grading system. For future work it is recommended to build a portable device in order to classify the agarwood oil quality on site. The device should be installed with the k-NN programming that has being successfully done in this study.

ACKNOWLEDGEMENTS

The authors acknowledge all the staff involved and the School of Electrical Engineering, College of Engineering, Universiti Teknologi MARA, Shah Alam, for supporting this research.

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, May 2013, doi: 10.3329/ijpls.v2i1.15132.

[2] N. A. M.A *et al.*, “Comparison of chemical profile of selected gaharu oils from Peninsular Malaysia,” *Malaysian J. Anal. Sci.*, vol. 12, no. 2, pp. 338–340, 2008, [Online]. Available: https://www.researchgate.net/publication/264000994_Comparison_of_chemical_profile_of_selected_gaharu_oils_from_Peninsular_Malaysia.

[3] A. Barden, N. A. Anak, T. Mulliken, and M. Song, “Heart of the matter: agarwood use and trade and CITES implementation for Aquilaria malaccensis,” Cambridge, UK, 2000. [Online]. Available: https://portals.iucn.org/library/efiles/documents/Traf-072.pdf.

[4] 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 J. Anal. Sci.*, vol. 17, no. 3, pp. 403–413, 2013, [Online]. Available: http://www.ukm.my/mjas/v17_n3/Nurlaila.pdf.

[5] H. Chhipa, K. Chowdhary, and N. Kaushik, “Artificial production of agarwood oil in Aquilaria sp. by fungi: a review,” *Phytochem. Rev.*, vol. 16, no. 5, pp. 835–860, Oct. 2017, doi: 10.1007/s11101-017-9492-6.

[6] N. Ismail, N. A. Mohd 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.*, vol. 68, no. 1, Apr. 2014, doi: 10.11113/jt.v68.2419.

[7] A. P. Davydov and T. P. Zlydneva, “On the reduction of free photons speed in modeling of their propagation in space by the wave function in coordinate representation,” in *2016 13th International Scientific-Technical Conference on Actual Problems of Electronics Instrument Engineering (APEIE)*, Oct. 2016, pp. 233–240, doi: 10.1109/APEIE.2016.7806458.

[8] M. Ishihara, T. Tsuneya, and K. Uneyama, “Components of the Agarwood Smoke on Heating,” *J. Essent. Oil Res.*, vol. 5, no. 4, pp. 419–423, Jul. 1993, doi: 10.1080/10412905.1993.9698252.

[9] M. Ishihara, T. Tsuneya, and K. Uneyama, “Fragrant sesquiterpenes from agarwood,” *Phytochemistry*, vol. 33, no. 5, pp. 1147–1155, Jul. 1993, doi: 10.1016/0031-9422(93)85039-T.

[10] S. N. Tajuddin and M. M. Yusoff, “Chemical Composition of Volatile Oils of Aquilaria malaccensis (Thymelaeaceae) from Malaysia,” *Nat. Prod. Commun.*, vol. 5, no. 12, pp. 1965–1968, 2010, doi: https://doi.org/10.1177/1934578X1000501229.

[11] S. N. Tajuddin, N. S. Muhamad, M. A. Yarmo, and M. M. Yusoff, “Characterization of the Chemical Constituents of Agarwood Oils from Malaysia by Comprehensive Two-Dimensional Gas Chromatography–Time-of-Flight Mass Spectrometry,” *Mendeleev Commun.*, vol. 23, no. 1, pp. 51–52, Jan. 2013, doi: 10.1016/j.mencom.2013.01.019.

[12] L. Auria and R. A. Moro, “Support Vector Machines (SVM) as a Technique for Solvency Analysis,” *SSRN Electron. J.*, vol. 1, no. 1, 2008, doi: 10.2139/ssrn.1424949.

[13] S. Elkatatny and M. Mahmoud, “Development of a New Correlation for Bubble Point Pressure in Oil Reservoirs Using Artificial Intelligent Technique,” *Arab. J. Sci. Eng.*, vol. 43, no. 5, pp. 2491–2500, May 2018, doi: 10.1007/s13369-017-2589-9.

[14] N. da Silva e Silva *et al.*, “Artificial intelligence application for classification and selection of fish gelatin packaging film produced with incorporation of palm oil and plant essential oils,” *Food Packag. Shelf Life*, vol. 27, p. 100611, Mar. 2021, doi: 10.1016/j.fpsl.2020.100611.

[15] Y. Kusuma Arbawa and C. Dewi, “Soil Nutrient Content Classification for Essential Oil Plants using kNN,” in *Proceedings of the 2nd International Conference of Essential Oils*, 2019, pp. 96–100, doi: 10.5220/0009957400960100.

[16] C. Dewi and Y. K. Arbawa, “Performance Evaluation of Distance Function in KNN and WKNN for Classification of Soil Organic Matter,” in *2019 International Conference on Sustainable Information Engineering and Technology (SIET)*, Sep. 2019, pp. 196–199, doi: 10.1109/SIET48054.2019.8986030.

[17] M. E. M. Samad, N. Ismail, M. H. F. Rahiman, M. N. Taib, N. A. M. Ali, and S. N. Tajuddin, “Analysis of distance metric variations in KNN for agarwood oil compounds differentiation,” in *2017 IEEE Conference on Systems, Process and Control (ICSPC)*, Dec. 2017, pp. 151–156, doi: 10.1109/SPC.2017.8313038.

[18] N. Ismail, “ANN modelling of agarwood oil significant chemical compounds for quality discrimination,” in *IPSis Biannual Publication*, Shah Alam: UiTM Institutional Repositories, 2015.

[19] Okfalisa, I. Gazalba, Mustakim, and N. G. I. Reza, “Comparative analysis of k-nearest neighbor and modified k-nearest neighbor algorithm for data classification,” in *2017 2nd International conferences on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, Nov. 2017, pp. 294–298, doi: 10.1109/ICITISEE.2017.8285514.

[20] G. Parthasarathy and B. N. Chatterji, “A class of new KNN methods for low sample problems,” *IEEE Trans. Syst. Man. Cybern.*, vol. 20, no. 3, pp. 715–718, 1990, doi: 10.1109/21.57285.

[21] H. Zhang, “The impact of distance, feature weighting and selection for KNN in credit default prediction,” in *Master Degree Project in Informatics with a specialisation in Data Science Spring term 2020*, Dissertation, 2020.

[22] G. Verdier and A. Ferreira, “Adaptive Mahalanobis Distance and k-Nearest Neighbor Rule for Fault Detection in Semiconductor Manufacturing,” *IEEE Trans. Semicond. Manuf.*, vol. 24, no. 1, pp. 59–68, Feb. 2011, doi: 10.1109/TSM.2010.2065531.

[23] K. Chomboon, P. Chujai, P. Teerarassammee, K. Kerdprasop, and N. Kerdprasop, “An Empirical Study of Distance Metrics for k-Nearest Neighbor Algorithm,” in *The Proceedings of the 2nd International Conference on Industrial Application Engineering 2015*, 2015, pp. 280–285, doi: 10.12792/iciae2015.051.

[24] X. Wu *et al.*, “Top 10 algorithms in data mining,” *Knowl. Inf. Syst.*, vol. 14, no. 1, pp. 1–37, Jan. 2008, doi: 10.1007/s10115-007-0114-2.

[25] L. E. Stenroos and K. J. Siebert, “Application of Pattern-Recognition Techniques to the Essential Oil of Hops,” *J. Am. Soc. Brew. Chem.*, vol. 42, no. 2, pp. 54–61, Apr. 1984, doi: 10.1094/ASBCJ-42-0054.

[26] A. Kataria and M. D. Singh, “A Review of Data Classification Using K-Nearest Neighbour Algorithm,” *Int. J. Emerg. Technol. Adv. Eng.*, vol. 3, no. 6, 2013.

[27] S. Taneja, C. Gupta, K. Goyal, and D. Gureja, “An Enhanced K-Nearest Neighbor Algorithm Using Information Gain and Clustering,” in *2014 Fourth International Conference on Advanced Computing & Communication Technologies*, Feb. 2014, pp. 325–329, doi: 10.1109/ACCT.2014.22.

[28] P. Mulak and N. Talhar, “Analysis of Distance Measures Using K-Nearest Neighbor Algorithm on KDD Dataset,” *Int. J. Sci. Res.*, vol. 4, no. 7, 2015, [Online]. Available: https://www.ijsr.net/archive/v4i7/SUB156942.pdf.

[29] V. I. K. Pathirana and R. K. M., “Mahalanobis based k-nearest neighbor forecasting versus time series forecasting methods,” 2015.

[30] M. D. Buhmann *et al.*, “Resubstitution Estimate,” in *Encyclopedia of Machine Learning*, Boston, MA: Springer US, 2011, pp. 863–863.

[31] L. Rueda, D. Mery, and J. Kittler, “Progress in Pattern Recognition, Image Analysis and Applications,” 2007.

[32] D. Berrar, “Cross-Validation,” in *Encyclopedia of Bioinformatics and Computational Biology*, Elsevier, 2019, pp. 542–545.

[33] J. R. Parker, “Rank and response combination from confusion matrix data,” *Inf. Fusion*, vol. 2, no. 2, pp. 113–120, Jun. 2001, doi: 10.1016/S1566-2535(01)00030-6.

[34] M. N. Taib and N. Ismail, “System Identification Makes Sense of Complex Measurements,” in *2021 11th IEEE International Conference on Control System, Computing and Engineering (ICCSCE)*, Aug. 2021, pp. 240–245, doi: 10.1109/ICCSCE52189.2021.9530875.

[35] S. Zhang, X. Li, M. Zong, X. Zhu, and R. Wang, “Efficient kNN Classification With Different Numbers of Nearest Neighbors,” *IEEE Trans. Neural Networks Learn. Syst.*, vol. 29, no. 5, pp. 1774–1785, May 2018, doi: 10.1109/TNNLS.2017.2673241.

[36] T. M. Martin *et al.*, “Does Rational Selection of Training and Test Sets Improve the Outcome of QSAR Modeling?,” *J. Chem. Inf. Model.*, vol. 52, no. 10, pp. 2570–2578, Oct. 2012, doi: 10.1021/ci300338w.

BIOGRAPHIES OF AUTHORS

Noor Syafina Mahamad Jainalabidin is graduated from School of Electrical Engineering, College of Engineering, Universiti Teknologi MARA (UiTM), Shah Alam in 2021, Selangor. She obtained Diploma in Electrical Engineering from the similar university and now working as an engineer in the related field. She can be contacted at email: nurlaila0583@uitm.edu.my.

Aqib Fawwaz Mohd Amidon was born in Malaysia, on September 1996. He received his B. Eng. (Hons) of Electronic Engineering from Universiti Teknologi MARA (UiTM). He is currently a Software Engineer at Greatech Technology Berhad and at the same time as full time postgraduate students at School of Electrical Engineering, College of Engineering, Universiti Teknology MARA, UiTM Shah Alam, Malaysia. He can be contacted at email: aqibfawwaz.academic@gmail.com.

Nurlaila Ismail is a senior lecturer at School of Electrical Engineering, College of Engineering, Universiti Teknologi MARA (UiTM), Shah Alam, Selangor. She obtained her BSc, MSc and PhD in Electrical Engineering from UiTM Shah Alam. She joined UiTM during her postdoctoral service in 2016. She is a professional engineer in the discipline of teaching recognized by the Board of Engineer Malaysia (BEM), an active member in several organizations, including IEEE Malaysia, especially Control System Society, Malaysian Society for Computed Tomography and Imaging Technology (MyCT) and Institute of Engineer Malaysia (IEM) and Malaysia Board of Technologists (MBOT) as well as ASEAN Engineering Register (AER). She has published more than 50 technical papers, locally and at international level. Her research interests are in advanced signal processing and artificial intelligence. She can be contacted at email: nurlaila0583@uitm.edu.my.

Zakiah Mohd Yusoff is a senior lecturer who is currently working at UITM Pasir Gudang. She received the B. ENG in Electrical Engineering and PhD In Electrical Engineering from UITM Shah Alam, in 2009 and 2014, respectively. In Mei 2014, she joined UITM Pasir Gudang as a teaching staff. Her major interests include process control, system identification, and essential oil extraction system. She can be contacted at email: zakiah9018@uitm.edu.my.

Saiful Nizam Tajuddin is a senior lecturer at Faculty of Industrial Science & Technology (FIST). He has worked at Universiti Malaysia Pahang since 2005 and was one the members who first founded FIST Faculty in 2007. Later in 2013, he established the BioAromatic Research Center (BIOAROMATIK) and is currently appointed as Director of COE. His main research area is on natural products, focusing on aromatic plants, agarwood (locally known as Gaharu), spanning the upstream to downstream development. His research projects include development of methods for extraction of essential oil, and isolation and purification of complex mixtures of terpene compounds that are predominantly present in the essential oil. Email: saifulnizam@ump.edu.my.

Mohd Nasir Taib received his PhD from UMIST, UK. He is a Senior Professor at Universiti Teknologi MARA (UiTM). He heads the Advanced Signal Processing Research Group at the School of Electrical Engineering, College of Engineering, UiTM. He has been a very active researcher and over the years had author and/or co-author many papers published in refereed journals and conferences. He can be contacted at email: dr.nasir@uitm.edu.my.