

Robust k-NN approach for classifying *Aquilaria* oil species by compounds

Noor Aida Syakira Ahmad Sabri¹, Nur Athirah Syafiqah Noramli¹,
Nik Fasha Edora Nik Kamaruzaman¹, Nurlaila Ismail¹, Zakiah Mohd Yusoff¹, Ali Abd Almisreb²,
Saiful Nizam Tajuddin³, Mohd Nasir Taib¹

¹Advanced Signal Processing Research Interest Group, Faculty of Electrical Engineering, Universiti Teknologi MARA, Shah Alam, Malaysia

²Faculty of Computer Science and Engineering, International University of Sarajevo, Sarajevo, Bosnia and Herzegovina

³Bioaromatic Research Centre of Excellence (BARCE), Universiti Malaysia Pahang Al-Sultan Abdullah, Gambang Kuantan, Malaysia

Article Info

Article history:

Received Nov 13, 2024

Revised Mar 17, 2025

Accepted Mar 26, 2025

Keywords:

Aquilaria oil species
k-nearest neighbours
Chemical compounds
Essential oils
Machine learning

ABSTRACT

Accurate classification of *Aquilaria* oil species is essential for ensuring the quality and authenticity of agarwood oils, which are widely used in perfumes and traditional medicine. This study investigated the effectiveness of the k-nearest neighbours (k-NN) machine learning model for classifying *Aquilaria* oil species based on four significant chemical compounds: dihydro- β -agarofuran, δ -guaiene, 10-epi- γ -eudesmol, and γ -eudesmol. The dataset comprised 480 samples of *Aquilaria* oil, which were analyzed using gas chromatography-mass spectrometry (GC-MS) and gas chromatography-flame ionization detector (GC-FID). The k-NN model, with an optimal k-value of 10 and using euclidean distance as the distance metric, achieved 100% accuracy, sensitivity, specificity, and precision in both training and testing datasets. These results demonstrate the robustness of k-NN in species identification, highlighting the discriminative power of the selected compounds. This study verifies that the integration of chemical profiling with machine learning offers a scalable solution for accurate species identification in the essential oil industry. Future work could explore hybrid models and data expansion techniques to further enhance the classification performance in more complex environmental conditions.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Nurlaila Ismail

Advanced Signal Processing Research Interest Group, Faculty of Electrical Engineering

Universiti Teknologi MARA

40450 Shah Alam, Selangor, Malaysia

Email: nurlaila0583@uitm.edu.my

1. INTRODUCTION

Aquilaria, a genus of trees known for producing highly valued agarwood, yields essential oils widely used in perfumes, traditional medicine, and aromatherapy. However, the accurate classification of *Aquilaria* oil species presents significant challenges due to their complex chemical composition and the overlap of compounds across different species. Identifying the specific species of *Aquilaria* oil is important for quality control and ensuring the authenticity of products, but manual and traditional methods often fall short in accuracy and efficiency [1], [2].

One of the major challenges in classifying *Aquilaria* oil species is the high degree of chemical similarity between different species, particularly in key compounds such as dihydro- β -agarofuran, δ -guaiene,

10-epi- γ -eudesmol, and γ -eudesmol. These similarities complicate species differentiation when relying solely on conventional methods like GC-MS and GC-FID, often leading to potential misclassification [3], [4]. Additionally, environmental factors such as soil conditions, climate, and extraction methods can cause variations in the chemical compounds of oils derived from the same species, further complicating the classification process [3].

Recent reports [5], [6] indicate that these challenges come from the limitations of traditional techniques, which lack the precision required to handle such overlapping data, especially in industrial applications where large-scale identification is necessary. Advances in machine learning (ML) have begun to offer promising solutions by leveraging computational models that can process and analyze complex chemical data with greater accuracy. These techniques, including support vector machines (SVM), random forests, and k-nearest neighbour (k-NN) models, have been applied to this problem with varying degrees of success [7], [8].

In particular, the k-NN model has gained attention due to its simplicity, versatility, and effectiveness in handling multi-dimensional datasets. k-NN is a non-parametric ML algorithm that classifies data points based on the majority label of their closest neighbours in a multidimensional space [9]. This technique has been shown to perform well in species identification tasks where chemical compound data, such as those obtained from *Aquilaria* oil, need to be analyzed.

Researchers have been inspired to apply the k-NN model to the classification of *Aquilaria* oil species by the need for more reliable and scalable identification techniques. Recent studies have demonstrated the potential of k-NN and similar models to achieve higher accuracy by considering multiple chemical compounds simultaneously [10]. By employing 3D plotting techniques, researchers have been able to visualize the separation of species based on their chemical compounds, enabling more precise classification. These advances suggest that k-NN, when compared to other models like SVM and random forests, is particularly well-suited for this task, achieving notable results in separating *Aquilaria* species with greater than 90% accuracy [11].

One of the most significant achievements in this field has been the development of hybrid models that combine k-NN with dimensionality reduction techniques such as principal component analysis (PCA). This combination helps in reducing the complexity of the data, leading to faster processing times and improved classification accuracy [12], [13]. However, while k-NN has shown considerable promise, studies that cross-evaluate machine learning models under varying background conditions, such as different extraction methods or environmental factors, indicate that k-NN's performance may decrease in highly noisy datasets, requiring further improvements [14].

The objective of this study was; i) to assess the effectiveness of the k-nearest neighbours (k-NN) model in accurately classifying *Aquilaria* oil species based on the chemical compounds such as dihydro- β -agarofuran, δ -guaiene, 10-epi- γ -eudesmol, and γ -eudesmol. ii) To determine the significance of the selected chemical compounds in distinguishing between different *Aquilaria* oil species and enhancing the classification accuracy using the k-NN model.

2. MATERIALS AND METHODS

The input data used in this study consisted of the peak area (%) of significant chemical compounds found in four species of *Aquilaria* oil samples. These chemical compounds, including dihydro- β -agarofuran, δ -guaiene, 10-epi- γ -eudesmol, and γ -eudesmol, were extracted using GC-MS and GC-FID analyses. The data was then subjected to a k-NN classification model developed within MATLAB software. The k-NN algorithm operates by analyzing the chemical compounds of each sample and classifying it based on the majority class of its nearest neighbours in a multi-dimensional space, where the dependent variable was the species label of the *Aquilaria* oil. The evaluation of the model was conducted by comparing its predicted classifications against actual species labels, using performance metrics such as accuracy, sensitivity, specificity, and precision. Cross-validation techniques were applied to ensure the robustness and generalizability of the model under various conditions, with accuracy being the primary criterion for model performance.

2.1. Data collection and experimental setup

Data collection as illustrated in Table 1 are used to classify *Aquilaria* oil species based on their chemical composition. The dataset comprises 480 samples of agarwood oil obtained from the Bio Aromatic Research Centre of Excellence (BARCE) at Universiti Malaysia Pahang Al-Sultan Abdullah (UMPSA). These samples were chosen for their comprehensive representation of four *Aquilaria* oil species: *Aquilaria Beccariana* (AB), *Aquilaria Malaccensis* (AM), *Aquilaria Crassna* (AC), and *Aquilaria Subintegra* (AS). Each sample was analyzed to determine the percentage of peak area for four chemical compounds: dihydro- β -agarofuran (a), δ -guaiene (b), 10-epi- γ -eudesmol (c), and γ -eudesmol (d). The need for collecting data from

these compounds originates from their relevance as key biomarkers in distinguishing between different *Aquilaria* oil species.

Table 1. List of chemical compounds dataset and peak area (%) for each *Aquilaria* oil species

Code	Chemical compounds	Ident. mode	Peak area (%) / Oil species			
			AB	AM	AC	AS
a	dihydro- β -agarofuran	FID,MS	1.25	0.55	0.48	0.44
b	δ -guaiene	FID,MS	0.74	2.02	0.21	0.35
c	10-epi- γ -eudesmol	FID,MS	0.34	6.73	2.54	2.16
d	γ -eudesmol	FID,MS	0.26	2.17	0.95	1.85

The peak area (%) varies significantly between the *Aquilaria* species. For instance, compound a shows its highest presence in AB at 1.25%, while AM records a lower value of 0.55%. Similarly, compound b is significantly higher in AM, with a peak area of 2.02%, compared to just 0.21% in AC, emphasizing the compound's role as a distinguishing factor for species identification. The presence of compound c is also most pronounced in AM, where it reaches 6.73%, underscoring its importance in differentiating AM from the other species. In contrast, compound d shows a more balanced distribution across all species, although it peaks in AM at 2.17%, suggesting it plays a role in species differentiation but with less distinctiveness than other compounds.

The variations in chemical composition across the species demonstrate that AM tends to have the highest concentration of all four compounds, particularly compounds b and c, indicating a richer chemical compound. On the other hand, AC exhibits consistently lower peak areas, especially for compound b, making it chemically distinguishable from the other species. The more evenly distributed of compound d across species indicates that this compound might have limited value in distinguishing between different *Aquilaria* species compared to the other chemical markers.

Experimental analysis was conducted using MATLAB software, which was utilized to implement the k-NN classification model. MATLAB was chosen for its powerful computational and data visualization capabilities, as well as its specialized machine learning toolboxes that are ideal for handling complex multidimensional datasets like those generated by gas chromatography-mass spectrometry (GC-MS) and gas chromatography-flame ionization detector (GC-FID) analyses. MATLAB's flexibility and precision in model development and evaluation make it well-suited for this type of chemical compound-based classification task.

2.2. Sample preparation and GC-MS/GC-FID analysis

The extraction process was conducted by BARCE at UMPA, where ground agarwood chips were soaked in water for several days to facilitate the breakdown of oil glands. Hydro distillation was then performed over 3 to 5 days to extract the agarwood oil. Once extracted, the oil samples were prepared for analysis by diluting them in analytical-grade dichloromethane (DCM) [15].

GC-MS analysis was performed using an Agilent 7890B GC System coupled with an Agilent 5977A mass spectrometer detector (MSD). The system utilized a DB-1ms column (30 m \times 250 μ m \times 0.25 μ m) with helium as the carrier gas at a flow rate of 1.0 mL/min. The oven temperature was programmed to start at 80 $^{\circ}$ C and increase at a rate of 3 $^{\circ}$ C per minute until reaching 250 $^{\circ}$ C, where it was held for 3 minutes. The mass spectrometer operated in electron impact (EI) mode with 70 eV energy. Mass spectra were identified using the National Institute of Standards and Technology (NIST) library, requiring a minimum similarity of 80% for confirmation.

Simultaneously, GC-FID analysis was conducted using a similar system, but with an FID detector operating at 250 $^{\circ}$ C. Peak areas of the four target chemical compounds were measured, and these values were used as input for the k-NN model.

2.3. Data integration and k-NN model development

The peak areas of the four chemical compounds were integrated into the k-NN classification model as input data. Each sample's peak areas for compounds a, b, c, and d were processed to classify the sample of the four *Aquilaria* oil species: AB, AM, AC, or AS. The k-NN model works by analyzing the distances between the chemical compounds of each sample and those of its nearest neighbours, classifying each sample based on the majority species of its neighbours in the feature space.

As depicted in Figure 1, the experimental process begins with data pre-processing, which is essential to ensure the accuracy and reliability of the k-NN classification model. Data pre-processing involves three critical steps: normalization, randomization, and data division. Data normalization is applied to scale the input features, which are the peak area percentages of the four chemical compounds in *Aquilaria* oil, into a

uniform range to improve the model's performance. The continuous input features are normalized into a range that correlates with the species classes from 1 to 4, as presented in Table 2, corresponding to AB (Class 1), AM (Class 2), AC (Class 3), and AS (Class 4). This ensures that no single feature dominates the model due to varying input data scales. Following normalization, the k-NN model's parameters were defined as follows:

- Training and testing ratio: 80:20–80% of the dataset is used for training, and 20% for testing. This ratio helps prevent overfitting and ensures the model generalizes well to new, unseen data.
- k-Value: 10–The model considers the 10 nearest neighbours for classification, a value that balances model complexity and accuracy.
- Distance metric: euclidean distance–this metric calculates the straight-line distance between two points in multi-dimensional space, determining similarity between samples.

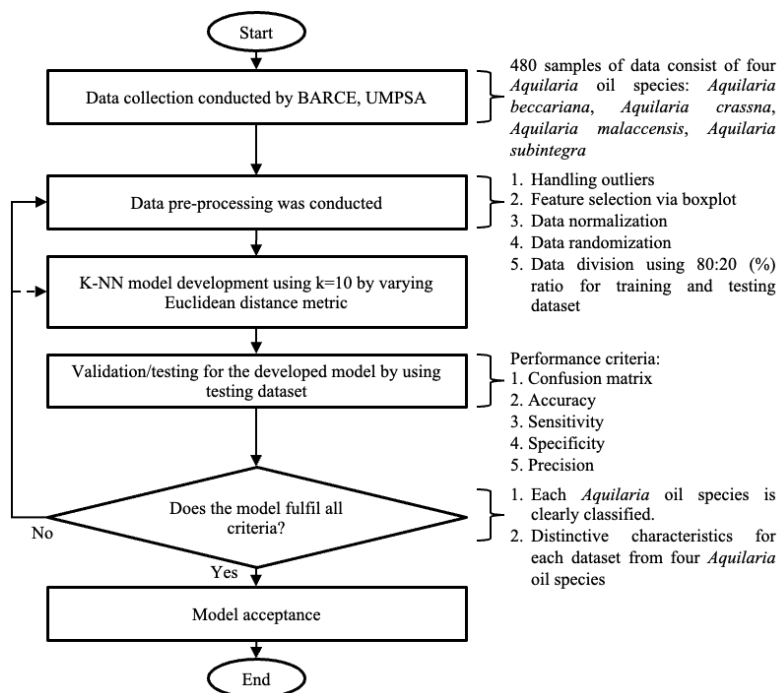


Figure 1. Process framework of k-NN model

The dataset was divided into training and testing sets, with 80% used for training and 20% reserved for testing. The 80% training data allows the model to learn patterns and correlations within the chemical compound data, while the remaining 20% is used to evaluate the model's performance. This data split is commonly employed to prevent overfitting, ensuring the model performs well on both known and unseen data.

Once pre-processing is complete, the k-NN model is trained and applied to classify the samples based on the euclidean distance. The euclidean distance measures the straight-line distance between two points, represented by their chemical compound values, and is calculated as shown in (1):

$$d(p, q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \dots + (p_n - q_n)^2} \quad (1)$$

In this formula, p and q represent two data points in the feature space, with n dimensions corresponding to the four chemical compounds. The k-NN algorithm then selects the k closest points (neighbours) and classifies a sample based on majority voting among these neighbours. The model's performance is evaluated using a confusion matrix, as illustrated in Table 2. Table 2 provides an example of a 4×4 multiclass confusion matrix for Class 1. The matrix highlights four key performance metrics:

- True positive (TP): the model correctly classifies a sample into the correct *Aquilaria* oil species (e.g., correctly identifying AM).
- True negative (TN): the model correctly identifies that a sample does not belong to a particular species (e.g., correctly rejecting a sample that is not AB).

- c) False positive (FP): the model incorrectly predicts a sample as belonging to a species when it does not (e.g., predicting AC as AS).
- d) False negative (FN): the model fails to correctly classify a sample, instead misclassifying it as another species (e.g., classifying AB as AC).

Table 2. Multiclass confusion matrix for class 1 (4×4)

Actual/predicted	Class 1	Class 2	Class 3	Class 4
Class 1	TP	FN_1	FN_2	FN_3
Class 2	FP_1	TN_1	TN_2	TN_3
Class 3	FP_2	TN_4	TN_5	TN_6
Class 4	FP_3	TN_7	TN_8	TN_9

*

$$TNT = TN (1+2+...+9)$$

$$FPT = FP (1+2+3)$$

$$FNT = FN (1+2+3)$$

In a multiclass confusion matrix, these metrics are calculated for each class, providing a comprehensive overview of the model's performance in handling multiple species. The diagonal elements represent the correctly classified samples for each species, while off-diagonal elements highlight the misclassified samples [16]-[18]. By analyzing the balance between true positives and false positives/negatives, the model's strengths and weaknesses in distinguishing between the four *Aquilaria* oil species can be assessed.

The overall performance of the k-NN model is evaluated using key metrics such as accuracy, sensitivity (recall), specificity, and precision [19]. Accuracy is calculated as $\frac{TP+TN}{TP+TN+FP+FN}$, indicating how often the model makes correct predictions. Sensitivity evaluates the model's ability to identify true positives, while specificity measures how well the model avoids false positives. Precision represents the ratio of true positives to the sum of true positives and false positives, illustrating the model's exactness in classification [19], [20].

The k-NN model's performance is optimized by adjusting the k-value and using the Euclidean distance metric. By fine-tuning these parameters, the model minimizes misclassification errors, particularly reducing false positives and false negatives [21]. The formulas for each evaluation metric are provided in (2)-(5).

$$Accuracy = \frac{TP+TNT}{TP+FN_T+FP_T+TNT} \times 100 \quad (2)$$

$$Sensitivity = \frac{TP}{TP+FN_T} \times 100 \quad (3)$$

$$Precision = \frac{TP}{TP+FP_T} \times 100 \quad (4)$$

$$Specificity = \frac{TNT}{FP_T+TNT} \times 100 \quad (5)$$

Finally, the k-NN model was validated using independent samples that were not included in the original training dataset. These tests confirmed that the model met all performance criteria, demonstrating high accuracy, sensitivity, specificity, and precision. The results affirm the robustness of the k-NN model in classifying *Aquilaria* oil species, making it a reliable tool for species identification based on chemical compounds.

3. EXPERIMENTAL RESULTS

In this section, it is explained the results of research and at the same time is given the comprehensive discussion. Results can be presented in figures, graphs, tables and others that make the reader understand easily [22], [23]. The discussion can be made in several sub-sections.

3.1. Boxplot analysis

The boxplot analysis was conducted to identify significant chemical compounds in *Aquilaria* oil samples. The boxplot method helped in visualizing the distribution and variability of the peak areas of the compounds across four *Aquilaria* oil species: AB, AM, AC, and AS.

In this case, the boxplot identified four significant compounds based on their peak area (%) values, focusing on the highest medians. By visually inspecting the boxplots for each species, these four compounds

emerged as the most frequently selected compounds with the highest peak areas. This selection process was pivotal because the compounds with the highest median peak areas across species indicate their importance in distinguishing between the oil species [24]. These compounds were then selected for further analysis.

3.2. 3D graphs and confusion matrix

The 3D visualization results (Figure 2) offer critical insights into the effectiveness of the k-NN model in classifying *Aquilaria* oil species based on the selected chemical compounds, a, b, c, and d. As shown in Figures 2(a)-2(c), these results present the chemical compounds of the *Aquilaria* oils in a three-dimensional space, allowing for a clear visual distinction between the different *Aquilaria* species. The 3D plot axes represent the peak area (%) of the compounds, highlighting their importance in separating species.

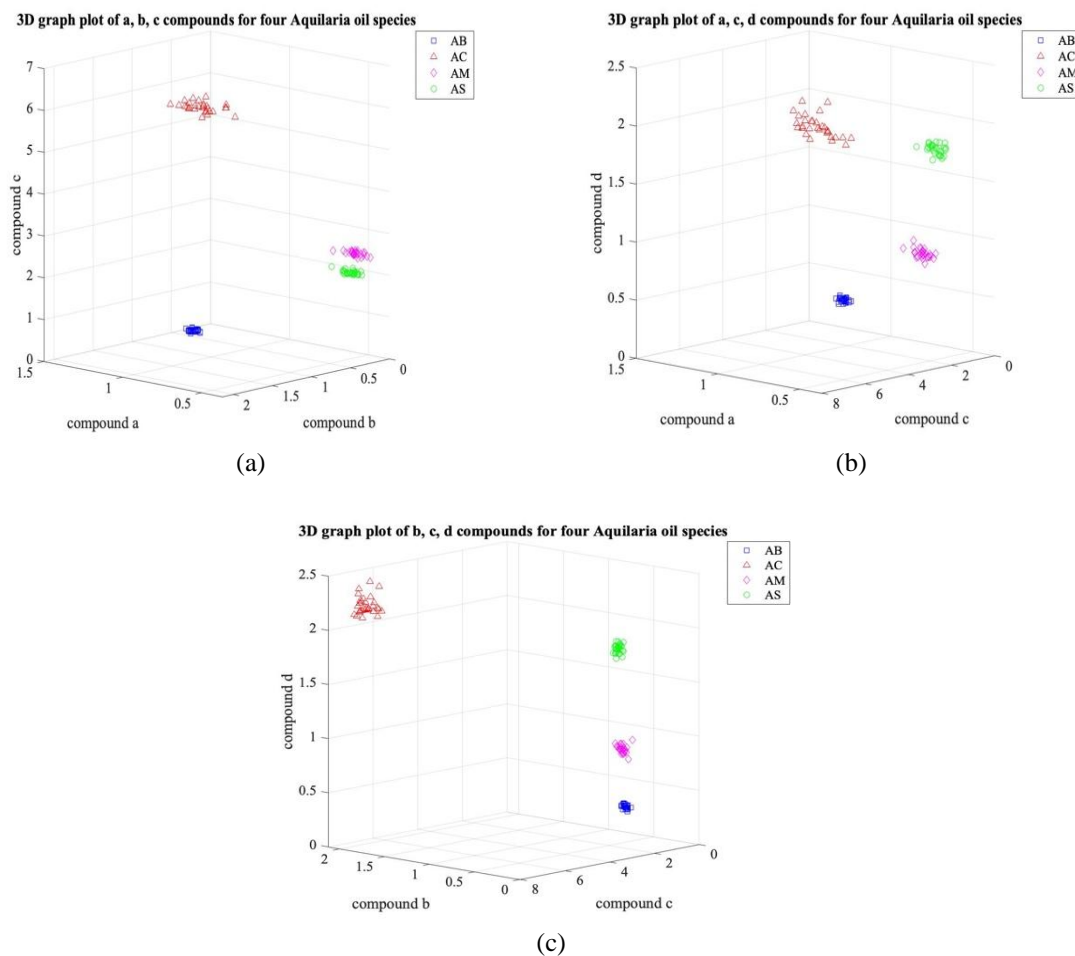


Figure 2. The 3-Dimension graphs of four *Aquilaria* oil species based on (a) a, b, c compounds, (b) a, c, d, and (c) b, c, d

The 3D plots demonstrate distinct clusters for each *Aquilaria* species, confirming the discriminative power of the selected compounds. Each cluster corresponds to a specific species, and the separation between clusters indicates that the chemical compositions of the species are sufficiently distinct. This separation visually supports the high classification accuracy achieved by the k-NN model, as the well-defined clusters suggest that the model can easily distinguish between the species. The strong separation between the clusters for species such as AC and AM underscores the relevance of the selected compounds in enhancing classification performance.

Additionally, the 3D results emphasize the role of specific compounds in the classification process. Axes corresponding to compounds b and c display more significant separations between species, indicating their higher relevance in distinguishing the oil profiles. This observation further validates the choice of these compounds for inclusion in the classification model. In contrast, any overlap in the 3D space could suggest a

need for further refinement of the model or the addition of more discriminative features to improve classification accuracy in cases where species show similar chemical compounds.

Figure 3 presents the confusion matrices used to evaluate the performance of the k-NN classification model in distinguishing four *Aquilaria* oil species based on selected chemical compounds. As shown in Figures 3(a) and 3(b), the matrices correspond to the training and testing datasets, respectively, offering insights into the model's classification accuracy.



Figure 3. Confusion matrix of four *Aquilaria* oil species for training (a) testing and (b) data

For the training data (3.1), which utilized 80% of the total dataset, the confusion matrix revealed accurate classification results. The four species (AB, AM, AC, and AS) were correctly predicted for every sample. Specifically, there were 24 TP for AB, 27 for AM, 24 for AC, and 21 for AS. No FP or FN were observed in the training data, resulting in 100% accuracy. This shows that the k-NN model successfully learned the patterns from the training data, accurately distinguishing the species based on their chemical compounds.

When evaluated on the testing data (3.2), which comprised 20% of the dataset, the model once again achieved accurate classification. The confusion matrix for the testing data indicated six correct predictions for AB, three for AM, six for AC, and nine for AS. Similar to the training set, no FP or FN were reported, and the model demonstrated 100% accuracy. The fact that the k-NN model maintained accurate accuracy on unseen testing data highlights its ability to generalize well beyond the training phase without overfitting.

As depicted in Table 3, further validation of the model's performance is demonstrated through key metrics such as accuracy, sensitivity, specificity, and precision, all of which reached 100%. This outcome is based on the confusion matrix, which reveals that every sample was correctly classified, with no errors in predicting the species. Each class, representing different species of *Aquilaria* oils, was accurately identified, with no FP or FN across both the training and testing datasets. This indicates that the chemical compounds of each species were distinctive for the model to classify them without confusion.

Table 3. Standard performance evaluation of four different *Aquilaria* oil species

Parameters	Percentages (%)
Accuracy	100
Sensitivity	100
Specificity	100
Precision	100

Starting with accuracy, the metric is calculated as the ratio of correct predictions (TP and TN) to the total number of predictions. The confusion matrix shows that all samples were placed in the correct category, leading to no misclassifications. As a result, the accuracy reached 100%, as the model made correct predictions for all species. Furthermore, sensitivity (recall) evaluates the model's ability to correctly identify

TP for each species. Given that no samples were misclassified as another species, the model's sensitivity for each species was accurate, indicating that every species' true samples were correctly identified without error.

Additionally, the model achieved 100% in specificity, which measures the ability to correctly reject samples that do not belong to a particular species. The confusion matrix confirms that no species was falsely classified as another. For example, the model never incorrectly identified AB as AC, showing that it avoided false positives entirely. Finally, precision, which is the ratio of TP to the sum of TP and FP, was also accurate. This means that all predictions made for each species were reliable, with no erroneous classifications.

The combination of these factors which are distinct chemical compounds, an appropriate k-value of 10, and the Euclidean distance metric enabled the k-NN model to excel in identifying the species. The model's ability to map the chemical compound data into a distinct feature space ensured that each species was separated, reducing the likelihood of any overlap. Therefore, the accurate classification demonstrated by the confusion matrix reflects the k-NN model's robustness and accuracy in handling this dataset, resulting in reliable performance across all evaluation metrics.

4. DISCUSSION

The results of this study demonstrated that the k-NN model, based on four significant chemical compounds, a, b, c, and d achieved accurate accuracy in classifying *Aquilaria* oil species. This aligns with recent research emphasizing the importance of chemical markers in essential oil classification. For example, [25] identified these sesquiterpenes and oxygenated compounds as key indicators of agarwood oil quality and origin. The use of Euclidean distance in the k-NN model further strengthened its ability to accurately classify species based on chemical compounds. Studies have similarly shown that combining chemical compounds with machine learning algorithms yields high accuracy in essential oil species identification [26].

The 3D visualization results strongly confirm the k-NN model's robust classification capabilities. The distinct clustering of *Aquilaria* oil species in the 3D space highlights the effectiveness of the selected chemical compounds in species differentiation and further underscores the model's ability to achieve high accuracy. This visualization reinforces the utility of chemical compounds combined with machine learning for large-scale classification of essential oils, offering a compelling approach for industrial applications.

Moreover, the high accuracy achieved by the k-NN model reflects the strong discriminative power of the selected chemical compounds, which have been consistently reported in the literature as important markers for *Aquilaria* species identification. In [27] confirmed that compounds a, b, c, and d are prevalent in high-quality agarwood oils, providing consistent chemical signatures across species. The selection of these compounds in this study, validated by the confusion matrix and performance metrics, aligns with findings from previous studies, further supporting their reliability in distinguishing between different *Aquilaria* species when integrated into machine learning models.

Additionally, the results suggest broader applicability of the k-NN model for other essential oil classification tasks. The high performance achieved here demonstrates the potential for applying similar approaches in industrial settings where large-scale classification of oils is required. Since machine learning techniques like k-NN can handle complex chemical data with minimal pre-processing, this model could be particularly useful in commercial authentication processes. Future research could explore hybrid approaches, combining k-NN with advanced models such as random forests or support vector machines, to enhance classification accuracy, particularly in noisy or more varied environmental conditions [28].

5. CONCLUSION

This research investigated the application of the k-NN model for classifying *Aquilaria* oil species based on four significant chemical compounds: dihydro- β -agarofuran (a), δ -guaiene (b), 10-epi- γ -eudesmol (c), and γ -eudesmol (d). The analysis revealed that the k-NN model, utilizing euclidean distance and an optimal k-value of 10, achieved a classification accuracy of 100% across both training and testing datasets, demonstrating the model's robustness and efficiency in species identification. The use of these selected compounds resulted in high performance metrics, including precision, sensitivity, and specificity, surpassing expectations in distinguishing between *Aquilaria* species. These findings suggest that the integration of specific chemical markers significantly enhances the performance of machine learning models for species identification in essential oils. This study proposes that machine learning models incorporating these key chemical markers can provide scalable solutions for large-scale classification tasks, particularly in the essential oil industry. Additionally, future research could explore the combination of multiple machine learning approaches and data augmentation techniques to further enhance classification performance in more complex or varied extraction environments, ensuring broader applicability and greater generalization.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge that this research article was financially supported by Ministry of Higher Education Malaysia through Universiti Teknologi MARA and Institute of Postgraduate Studies UiTM (IPSis), Journal Support Fund (JSF). The insightful feedback and contributions from members of the Advance Signal Processing Research Interest Group are deeply appreciated. The authors also wish to extend their gratitude to the Bio-Aromatic Research Centre of Excellence (BARCE) at Universiti Malaysia Pahang Al-Sultan Abdullah (UMPSA) for their invaluable assistance with data extraction.

FUNDING INFORMATION

The authors gratefully acknowledge that this research article was financially supported by Ministry of Higher Education Malaysia through Universiti Teknologi MARA and Institute of Postgraduate Studies UiTM (IPSis), Journal Support Fund (JSF).

AUTHOR CONTRIBUTIONS STATEMENT

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Noor Aida Syakira	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	
Ahmad Sabri														
Nur Athirah Syafiqah	✓	✓	✓			✓		✓		✓	✓			
Noramli														
Nik Fasha Edora Nik Kamaruzaman	✓		✓	✓		✓				✓	✓			
Nurlaila Ismail	✓	✓		✓						✓		✓	✓	✓
Zakiah Mohd Yusoff	✓			✓			✓	✓		✓	✓	✓		
Ali Abd Almisreb	✓	✓	✓				✓				✓			
Saiful Nizam Tajuddin	✓			✓			✓	✓				✓		
Mohd Nasir Taib	✓	✓		✓						✓	✓	✓	✓	

C : Conceptualization

M : Methodology

So : Software

Va : Validation

Fo : Formal analysis

I : Investigation

R : Resources

D : Data Curation

O : Writing - Original Draft

E : Writing - Review & Editing

Vi : Visualization

Su : Supervision

P : Project administration

Fu : Funding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

INFORMED CONSENT

We have obtained informed consent from all individuals included in this study.

DATA AVAILABILITY

Data availability is not applicable to this paper as no new data were created or analyzed in this study.




REFERENCES

- [1] F. Hashempour-baltork *et al.*, "Recent methods in detection of olive oil adulteration: State-of- the-Art," *Journal of Agriculture and Food Research*, vol. 16, p. 101123, Jun. 2024, doi: 10.1016/j.jafr.2024.101123.
- [2] M. Rasekh, H. Karami, A. D. Wilson, and M. Gancarz, "Classification and identification of essential oils from herbs and fruits based on a MOS electronic-nose technology," *Chemosensors*, vol. 9, no. 6, p. 142, Jun. 2021, doi: 10.3390/chemosensors9060142.
- [3] A. Yulianti, B. Abdullah, S. Leong, B. Wong, S. Samling, and S. Fong, "Grading of agarwood based on their chemical profiles using GC-MS incorporating chemometric approaches," 2024.
- [4] Y. Wang *et al.*, "Aquilaria species (thymelaeaceae) distribution, volatile and non-volatile phytochemicals, pharmacological uses, agarwood grading system, and induction methods," *Molecules*, vol. 26, no. 24, p. 7708, Dec. 2021, doi: 10.3390/molecules26247708.
- [5] M. Aqeel, A. Sohaib, M. Iqbal, H. U. Rehman, and F. Rustam, "Hyperspectral identification of oil adulteration using machine learning techniques," *Current Research in Food Science*, vol. 8, p. 100773, 2024, doi: 10.1016/j.crfs.2024.100773.




- [6] H. Saputra, B. Satria, N. Nazir, and T. Anggraini, "Development of agarwood oil research and benefit: bibliometric analysis," *AJARCADE (Asian Journal of Applied Research for Community Development and Empowerment)*, pp. 55–60, Mar. 2024, doi: 10.29165/ajarcde.v8i1.374.
- [7] S. H. Shetty, S. Shetty, C. Singh, and A. Rao, "Supervised machine learning: algorithms and applications," *Fundamentals and Methods of Machine and Deep Learning*. Wiley, pp. 1–16, Jan. 2022, doi: 10.1002/9781119821908.ch1.
- [8] M. A. Abas *et al.*, "Agarwood oil quality classifier using machine learning," *Journal of Fundamental and Applied Sciences*, vol. 9, no. 4S, p. 62, Jan. 2018, doi: 10.4314/jfas.v9i4s.4.
- [9] R. K. Halder, M. N. Uddin, M. A. Uddin, S. Aryal, and A. Khraisat, "Enhancing K-nearest neighbor algorithm: a comprehensive review and performance analysis of modifications," *Journal of Big Data*, vol. 11, no. 1, Aug. 2024, doi: 10.1186/s40537-024-00973-y.
- [10] A. H. Zaidi *et al.*, "Statistical analysis of agarwood oil chemical compound exists in four species of Aquilaria," *International Journal of Advances in Applied Sciences*, vol. 13, no. 3, p. 727, Sep. 2024, doi: 10.11591/ijaas.v13.i3.pp727-732.
- [11] F. A. Mufarroha, A. Z. Nur, M. R. Rahabillah, A. Jauhari, D. R. Anamisa, and Mulaab, "Spices identification in essential oil producers using comparison of KNN and Naïve Bayes classifier," in *Proceedings of the 4th International Conference on Informatics, Technology and Engineering 2023 (InCITE 2023)*, Atlantis Press International BV, 2023, pp. 618–627.
- [12] M. O. Arowolo, M. O. Adebisi, A. A. Adebisi, and O. Olugbara, "Optimized hybrid investigative based dimensionality reduction methods for malaria vector using KNN classifier," *Journal of Big Data*, vol. 8, no. 1, Feb. 2021, doi: 10.1186/s40537-021-00415-z.
- [13] P. Mavaie, L. Holder, and M. K. Skinner, "Hybrid deep learning approach to improve classification of low-volume high-dimensional data," *BMC Bioinformatics*, vol. 24, no. 1, Nov. 2023, doi: 10.1186/s12859-023-05557-w.
- [14] E. Ozturk Kiyak, B. Ghasemkhani, and D. Birant, "High-level K-nearest neighbors (HLKNN): a supervised machine learning model for classification analysis," *Electronics*, vol. 12, no. 18, p. 3828, Sep. 2023, doi: 10.3390/electronics12183828.
- [15] Z. M. Yusoff and N. Ismail, "Datasets of chemical compounds in three different species of aquilaria using GC-MS coupled with GC-FID analysis," *Data in Brief*, vol. 53, p. 110209, Apr. 2024, doi: 10.1016/j.dib.2024.110209.
- [16] M. Z. Naser and A. H. Alavi, "Insights into performance fitness and error metrics for machine learning," *arXiv preprint arXiv:2006.00887*, doi: 10.48550/arXiv.2006.00887.
- [17] K. K. Biliaminu, S. A. Busari, J. Rodriguez, and F. Gil-Castiñeira, "Beam prediction for mmWave V2I communication using ML-based multiclass classification algorithms," *Electronics*, vol. 13, no. 13, p. 2656, Jul. 2024, doi: 10.3390/electronics13132656.
- [18] R. A. I. Almashhadani, G. C. Hock, F. H. Bt Nordin, and H. N. Abdulrazzak, "Electroluminescence images for solar cell fault detection using deep learning for binary and multiclass classification," *International Journal of Electrical and Electronics Engineering*, vol. 11, no. 5, pp. 150–160, May 2024, doi: 10.14445/23488379/ijeee-v11i5p114.
- [19] N. H. M. Ariffin, M. I. M. Iqbal, M. Yusoff, and N. A. M. Zulkefli, "A study on the best classification method for an intelligent phishing website detection system," *Journal of Advanced Research in Applied Sciences and Engineering Technology*, vol. 48, no. 2, pp. 197–210, Jul. 2024, doi: 10.37934/araset.48.2.197210.
- [20] K. A. Athirah, N. Ismail, M. N. Taib, N. A. M. Ali, M. Jamil, and S. Lias, "Modelling of cymbopogon oils species using k-nearest neighbours (k-NN)," in *2019 IEEE 7th Conference on Systems, Process and Control (ICSPC)*, Dec. 2019, pp. 5–9, doi: 10.1109/icspc47137.2019.9068086.
- [21] K. Stapor, P. Ksieniewicz, S. Garcia, and M. Woźniak, "How to design the fair experimental classifier evaluation," *Applied Soft Computing*, vol. 104, p. 107219, Jun. 2021, doi: 10.1016/j.asoc.2021.107219.
- [22] J. Sadowski, "When data is capital: datafication, accumulation, and extraction," *Big Data and Society*, vol. 6, no. 1, p. 205395171882054, Jan. 2019, doi: 10.1177/2053951718820549.
- [23] J. R. Saura, B. R. Herraez, and A. Reyes-Menendez, "Comparing a traditional approach for financial brand communication analysis with a big data analytics technique," *IEEE Access*, vol. 7, pp. 37100–37108, 2019, doi: 10.1109/ACCESS.2019.2905301.
- [24] N. A. S. Ahmad Sabri *et al.*, "Statistical analysis for chemical compound based on several species of aquilaria essential oil," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 14, no. 4, p. 3663, Aug. 2024, doi: 10.11591/ijece.v14i4.pp3663-3673.
- [25] N. A. Syameera *et al.*, "Effects of heat treatment on the chemical composition, antioxidant activity, and toxicity of agarwood oil," *Journal of King Saud University - Science*, vol. 36, no. 4, p. 103141, Apr. 2024, doi: 10.1016/j.jksus.2024.103141.
- [26] M. Sabatino *et al.*, "Experimental data based machine learning classification models with predictive ability to select in vitro active antiviral and non-toxic essential oils," *Molecules*, vol. 25, no. 10, p. 2452, May 2020, doi: 10.3390/molecules25102452.
- [27] T. Yan, S. Yang, Y. Chen, Q. Wang, and G. Li, "Chemical profiles of cultivated agarwood induced by different techniques," *Molecules*, vol. 24, no. 10, p. 1990, May 2019, doi: 10.3390/molecules24101990.
- [28] L. A. Demidova, "Two-stage hybrid data classifiers based on SVM and kNN algorithms," *Symmetry*, vol. 13, no. 4, p. 615, Apr. 2021, doi: 10.3390/sym13040615.

BIOGRAPHIES OF AUTHORS






Noor Aida Syakira Ahmad Sabri    obtained her bachelor of engineering (Hons) in electronic engineering from Universiti Teknologi MARA (UiTM), Shah Alam, Malaysia, in 2022. Currently, she is a graduate research assistant at the Faculty of Electrical Engineering, UiTM Shah Alam, where she is pursuing postgraduate studies. Her research interests include advanced signal processing and machine learning, particularly in the analysis and classification of agarwood oil, leveraging computational methods to improve the accuracy and efficiency of chemical composition analysis. She can be contacted at email: aidasyakiraaa01@gmail.com.




Nur Athirah Syafiqah Noramli    received her B. Sc. (Hons) in computer science from Universiti Teknologi MARA (UiTM) Cawangan Melaka Kampus Jasin. She is currently pursuing her studies as a postgraduate student at the Faculty of Electrical Engineering, at Universiti Teknologi MARA (UiTM) Shah Alam, Selangor, Malaysia. Her research interests include advanced signal processing, and machine learning. She can be contacted at email: athirah.noramli1@gmail.com.






Nik Fasha Edora Nik Kamaruzaman    received the B.Eng. (Hons) of electronic engineering from Universiti Teknologi MARA (UiTM), Malaysia, in 2022. She is currently a postgraduate student at Faculty of Electrical Engineering, Universiti Teknologi MARA (UiTM) Shah Alam, Malaysia. Her research interests include advanced signal processing and machine learning. She can be contacted at email: nikfashaedora98@gmail.com.






Associate Professor Ir. Ts. Dr. Nurlaila Ismail    received the M.Sc. and Ph.D. degrees in electrical engineering from Universiti Teknologi MARA (UiTM), Malaysia. She is currently an associate professor at Faculty of Electrical Engineering, Universiti Teknologi MARA (UiTM) Shah Alam, Malaysia. Her research interests include advanced signal processing, machine learning, and artificial intelligence. She can be contacted at email: nurlaila0583@uitm.edu.my.






Associate Professor Ts. Dr. Zakiah Mohd Yusoff    received her bachelor's degree in electrical engineering and Ph.D. in electrical engineering from Universiti Teknologi MARA Shah Alam, in 2009 and 2014, respectively. She is a senior lecturer who is currently working at Faculty of Electrical Engineering, Universiti Teknologi MARA (UiTM) Shah Alam, Malaysia. In May 2014, she joined Universiti Teknologi MARA as a teaching staff. Her major interests include process control, system identification, and essential oil extraction systems. She can be contacted at email: zakiah9018@uitm.edu.my.






Associate Professor Dr. Ali Abd Almisreb    is currently an associate professor at the faculty of computer sciences and engineering, director of graduate council and editor in chief at International University of Sarajevo. He received a M.Sc. degree in computer science and Ph.D. degree in electrical engineering/computer engineering from Universiti Teknologi MARA (UiTM), Malaysia. His major interests include deep learning, machine learning, computer vision voice recognition and quantum computing. He can be contacted at email: alimes96@yahoo.com.



Professor Dr. Saiful Nizam Tajuddin    received his Ph.D. degree from Universiti Malaysia Pahang (UMP), Malaysia. He is a professor and director of Bioaromatic Research Center of Excellence (BARCE) at Universiti Malaysia Pahang. He is a director and researcher at Synbion Sdn. Bhd., Kuantan, Pahang, Malaysia. 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: saifulnizam@ump.edu.my.



Prof. Ir. Ts. Dr. Haji Mohd Nasir Taib    received the degree in electrical engineering from the University of Tasmania, Hobart, Australia, the M.Sc. degree in Control Engineering from Sheffield University, UK, and the Ph.D. degree in Instrumentation from the University of Manchester Institute of Science and Technology, UK. He is currently a Senior Professor at Universiti Teknologi MARA (UiTM), Malaysia. He heads the advanced signal processing research interest group at the Faculty of Electrical 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.