Decision Tree on Branch Node Observations for Significant Compounds of Agarwood Oil Different Grades

Shilton Lotthar Inau anak Samawi[1], Nurlaila Ismail[1*], Zakiah Mohd Yusoff[2], Mohd Hezri Fazalul Rahiman[1], Saiful Nizam Tajuddin[3], Mohd Nasib Taib[1]

[1]School of Electrical Engineering, College of Engineering, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia
[2]School of Electrical Engineering, College of Engineering, Universiti Teknologi MARA, Cawangan Johor, Kampus Pasir Gudang, Malaysia
[3]BioAromatic Research Centre of Excellence, Universiti Malaysia Pahang, Gambang, Malaysia

Corresponding author* email: nurlaila0583@uitm.edu.my
Available online 20 June 2023

ABSTRACT

For ages, agarwood oil has been utilized in a range of applications, including scented goods such as incense and fragrances, as well as medicinal therapies. There is currently no accepted method for precisely grading agarwood oil based on its chemical constituents, despite the fact that demand for it is constantly rising. Consequently, the goal of this work is to create a revolutionary method for classifying agarwood oil according to its chemical components. In this study, a machine learning technique based on the Decision Tree (DT) classification algorithm was created to address this issue. The DT model was based on branch node observations, with corresponding values of 11 for MinParent. The result showed that the accuracy of the DT model for branch node is 94.74%. The performance of the model was evaluated using confusion matrix, accuracy, resubstitution error and cross validation error in MATLAB software version R2021a. This study revealed the DT as an effective method for classifying agarwood oil grades based on chemical compounds, which has the potential to greatly benefit the agarwood oil industry and research field.

Keywords: Decision Tree, Machine Learning, Agarwood Oil, Oil Grades

1. Introduction

Aquilaria tree, or commonly known as agarwood tree is one of the most expensive raw material available in the world and its price can reach up to USD100,000 per kg depending on the quality of the woods [1]. The tree must undergo the inoculation process in order to produce agarwood. This procedure is used to introduce a particular kind of bacterial and fungal infection into the agarwood tree's healthy, odorless heartwood. The tree will consequently respond by creating a dark, damp, scented resin as a protection strategy. The wood chips that infused with the resin called Oud will be distilled into an essential oil. This oil can also be called as agarwood oil [2, 3]. Agarwood oil has been used for a variety of purposes, including the production of incense, perfumes, religious rituals, and conventional medications. Agarwood oil is highly prescribed on the international market, particularly in China, Japan, and the Middle Eastern nations, and Malaysia is no exception due to its wide range of uses [4].

The quality of agarwood oil can be determined by looking at the physical properties like color and odor. The aromatic scent from the resinous heartwood from Aquilaria consists of numerous families such as the *Leguminosae*, *Thymelaeaceae* and *Euphorbiaceae* [5]. The physical qualities of agarwood oil, such as color, durable odor, powerful fixative, and consumer perception, are currently used to grade the product's quality [6]. This method is impractical due to the human nose's inability to gather multiple samples at once. Particularly when dealing with persistent strong scent, which causes nasal fatigue. Furthermore, the human sensory system has limitations in terms of reproducibility, time consumption, subjectivity, and high labor costs [7]. Therefore, this study set out to implement the usefulness of an

intelligent classification using machine learning with the capacity to provide services to end users is needed to grade the agarwood oil quality [7]. The aim of this paper is to apply the Boxplot Analysis technique in analyses the Agarwood oil chemical profiles based on its abundance from four different qualities. There are seven chemical compounds for 660 oil samples of data from high and low qualities.

According to published research, it is possible to classify agarwood oil based on its chemical properties to improve grading precision and uniformity [6]. There have been suggestions made to allow the grade of agarwood oil to be correctly and consistently categorized based on its chemical properties in order to address this problem. But there is still no accepted system for grading the caliber of agarwood oil. Researchers from Japan found that several chemical compounds, including β-agarofuran, α-agarofuran, γ-eudesmol, eudesmol and 10-epi-eudesmol, were present in high-grade oil [8]. On the other hand, low-grade agarwood oil frequently contains the chemicals longifolol and hexadecanol [9]. The agarwood oil quality reflects the market price of the oil. Both high-grade and low-grade goods are sold according to their quality which higher grade is more expensive than the lower one [10]. Dark colors and strong scents are typically associated with high grade and are sold at a higher price. The price of a high-grade oil per tola (12ml) ranged from USD 126 to USD 633 [4]. Furthermore, agarwood wood can cost USD19 per kg for low grade and USD100,000 per kg for superior grade [4].

With the finding that agarwood oil may be categorized based on its chemical properties to increase grading accuracy and consistency, it has been suggested that the grade be determined precisely and consistently based on those qualities. [7]. Nonetheless, there is still no accepted system for grading agarwood oil. Agarwood oil must be divided into high and low grades in order for them to be easily distinguishable from one another [8]. Therefore, this research is conducted to filled up the research gap by introducing decision tree as one of the machine learning techniques to classify agarwood oil different grades. The finding obtained in this study is vital as it contributed to the standard of agarwood oil quality grading, its industry as well as research area.

2. Theoretical Work

Decision tree is one of the machine learning classifications techniques that can be used in grading the essential oil [11, 12]. Analysis based on decision tree modelling with different splitting processes is required to ensure that agarwood oil quality grading will produce a better result in terms of efficiency. Currently, decision tree techniques are popular in medical research and had been applied to diagnose a medical condition from the pattern of symptoms [12].

A decision tree is a hierarchical, flowchart-like structure in which internal nodes correspond to attribute tests, branches denote test outcomes, and leaf nodes (or terminal nodes) signify class labels. When presented with a tuple X, the tree evaluates the attribute values against its structure. A path is followed from the root to a leaf node, which provides the class prediction for the tuple. Transforming decision trees into classification rules is straightforward. Decision tree learning employs this model to link observations about an item with conclusions regarding its target value. The basic structure of decision tree is shown in Figure 1 [13]. This method is a common approach in statistics, data mining, and machine learning for predictive modeling. When the target variable takes a finite set of values, these tree models are referred to as classification trees, where leaves denote class labels and branches represent combinations of features leading to those labels.

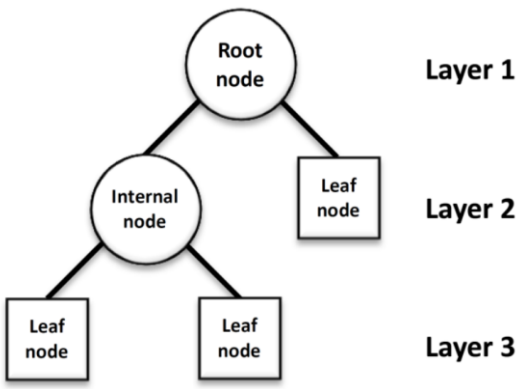

Figure 1 Basic structure of decision tree [13]

Decision tree performance in terms of precision, speed, and accuracy are superb compared to others method [14]. However up to the latest references in 2022, there are still no specific research on grading the agarwood oil quality using the decision tree techniques [15, 16]. In light of it, DT was chosen for this research. The rating of agarwood oil was determined using the decision tree technique.

3. Methodology

3.1 Data collection

The agarwood oil samples for this study are obtained from previous researcher [5]. It is a collaboration with the Forest Research Institute Malaysia (FRIM), Malaysia and Bio Aromatic Research Centre of Excellence (BARCE), University Malaysia Pahang (UMP), Malaysia. The previous researchers prepared the samples and evaluated the chemicals [5,17,18]. There are 660 samples of agarwood oil from high and low qualities. Gas chromatography-mass spectrometry is used to identify the chemical components in agarwood oil (GC-MS). Out of hundreds of chemical compounds exits, only seven chemicals are selected for this study and are fed to the decision tree modelling as inputs. The compounds are β-agarofuran, α-agarofuran, 10-epi-γ-eudesmol, γ-eudesmol, longifolol, hexadecanol, and eudesmol. The quality of the agarwood oil either 'high' or 'low' quality is set as output. After that, MATLAB software version R2021a is chosen to perform the analytical work decision tree model development. The "fitrtree" function with corresponding values of 11 for 'Minparent' is implemented during the development stage.

At the DT modelling stage, the process is executed by using the training dataset. The corresponding value chosen for the branch node is 11. The performance of DT modelling with different nodes as well as its different corresponding values were evaluated and tested.

Four performance criteria are used to check the performance of DT modelling such as resubstitution error (Closs), cross-validation error (Kloss), accuracy (%) and confusion matrix by using the testing dataset. If the model passes all the criteria, the model is accepted but if not, the model needs to return to the pre-processing stage for model performance enhancement.

3.2 Decision Tree (DT) model development

Figure 2 shows the whole process of the DT model development. Firstly, the data collection consists of input and output which are high and low grade of agarwood oil [5]. Following that it is data pre-processing, where the data had been divided into two groups: training and testing dataset (with the ratio of 80% of training and 20% of testing). Then, the development of DT model is performed by using training dataset with the input is the seven chemical compounds and the output is the quality. The branch node is varying and the Minparent value is 11. After that, the developed model is validated using the testing dataset. Four performance measures are used to check either the developed DT model can be accepted or not. They are resubstitution error (closs), cross-validation error (kloss), accuracy (%) and confusion matrix. The DT model is accepted once it passed all these performances with the accuracy (%) is above 80.00 %. If not, the model has to undergone pre-processing stage and re-train.

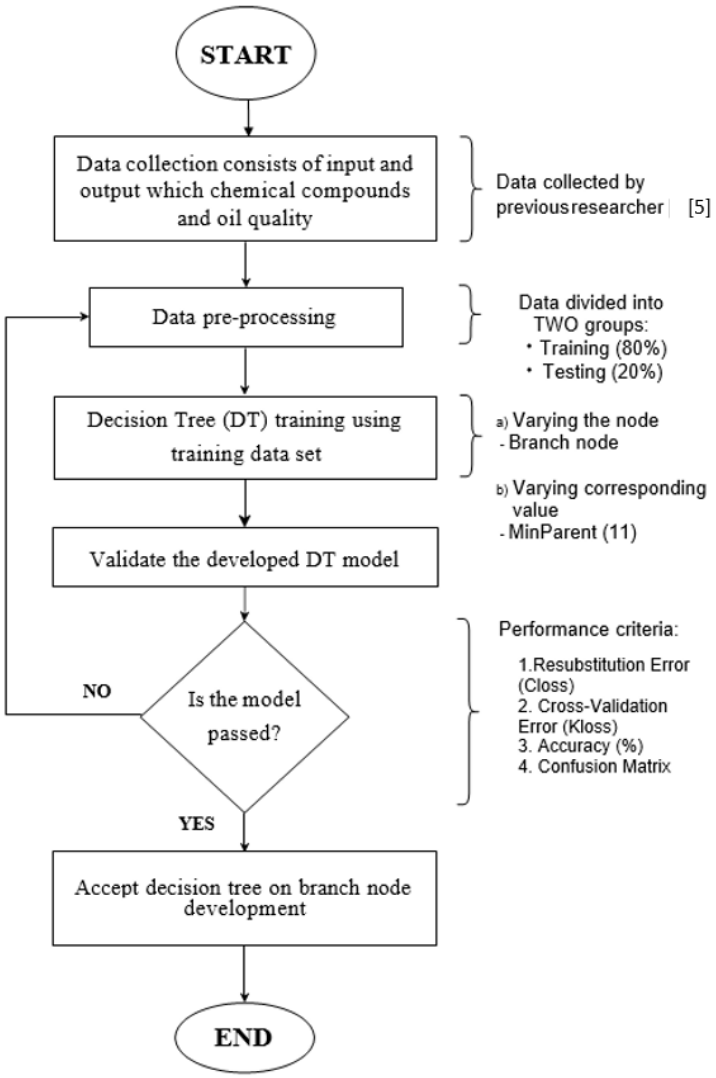

Figure 2 Flowchart of DT Model Development

To further subdivide a node into sub-nodes, decision trees can employ a variety of algorithms. Technically, the decision tree splits the nodes using all of the available variables, but ultimately it selects the split that produces the most homogeneous sub-nodes. In this situation, choosing the right algorithm is extremely important.

Iterative Dichotomiser 3 (ID3), C4.5, C5.0 (updated version of C4.5), Chi-square Automatic Interaction Detector (CHAID), Multivariate Adaptive Regression Splines (MARS), and Classification and Regression Trees (CART) are among the decision tree algorithms available [14].

In this study, CART algorithm had been selected as the decision tree algorithm. The CART algorithm was chosen because it can solve both classification and regression problems. Furthermore, unlike the ID3 and C4.5 algorithms, it divides datasets using the Gini index metric, as opposed to information gain or entropy and gain ratio [14].

With CART, the splitting process adopts a greedy strategy with the goal of lowering the cost function. The purity of the leaf nodes for classification problems is calculated using the Gini index as a cost function. The cost function for regression used by the algorithm to determine the best prediction is the Gini index [14].

The Gini index is used to determine the cost of feature splits in a dataset. The measure defines the likelihood that a randomly selected data point would be misclassified by a particular node. Gini index is given by the formula [14]:

$$Gini = 1 - \sum_{i=1}^{n} (pi)^2$$

4. Results and Discussion

This section presented the agarwood dataset together with the results of DT modelling based on the MinParent for agarwood oil quality grading. This DT modelling is capable of classified the quality of agarwood oil in the form of a DT branch node. The classification results show that the quality of agarwood oil was determined by analyzing the data of the agarwood oil compound.

4.1 Decision Tree Analysis

Figure 3 until Figure 9 illustrate the decision tree of all the seven compounds for the branch node. Four out of the seven compounds had the same decision tree pattern which is the 10-epi-γ-eudesmol, longifolol, hexadecanol and eudesmol. Longifolol, and hexadecanol ended with three root nodes, β-agarofuran, α-agarofuran, and eudesmol contained two root nodes meanwhile, 10-epi-γ-eudesmol, and γ-eudesmol ended with only one root node.

The value of '1' indicates the low grade of agarwood oil and the value of '2' indicate the high grade of agarwood oil. The results of the analysis can be seen in Table 1. The results of a DT based on the compounds composition in the agarwood oil and each of them have a different pruning level. In branch node, pruning level in compounds β-agarofuran and α-agarafuran have a similar pruning level which are 6. As for 10-epi-γ-eudesmol have a pruning of 2, γ-eudosmol have 3 pruning level, and eudesmol have 4 pruning level. On the other hand, longifolol has a pruning level of 7. Lastly, hexadecanol have 9 pruning level.

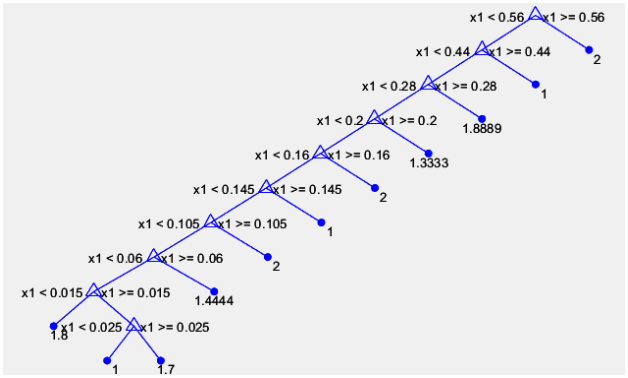

Figure 3 DT of β-agarofuran

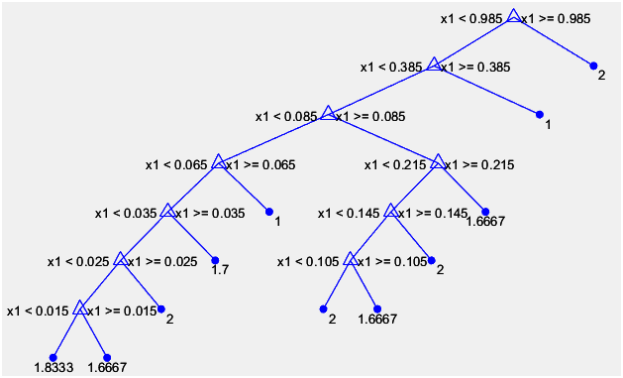

Figure 4 DT of α-agarofuran

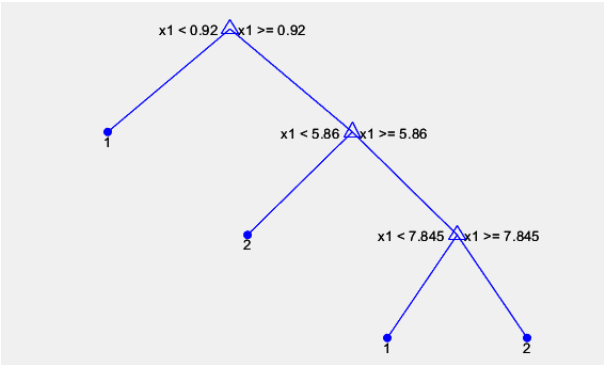

Figure 5 DT of 10-epi-eudesmol

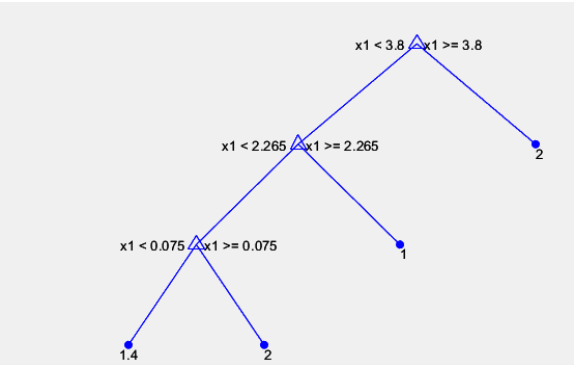

Figure 6 DT of γ-eudesmol

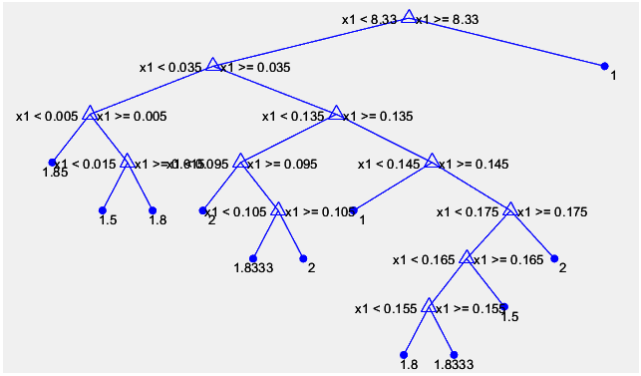

Figure 7 DT of longifolol

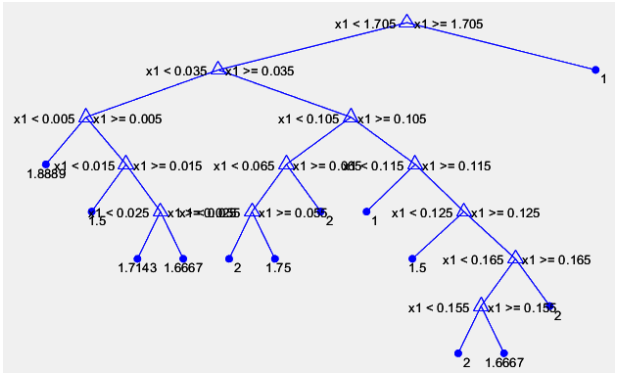

Figure 8 DT of hexadecanol

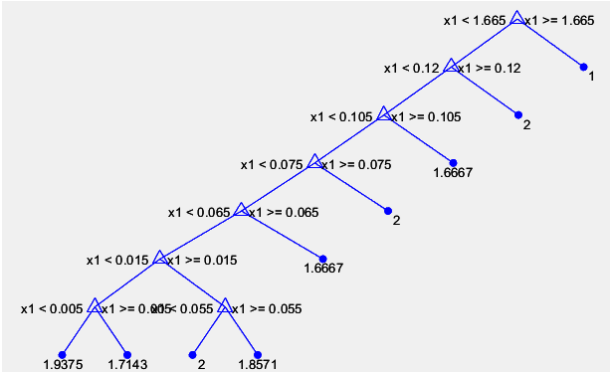

Figure 9 DT of eudesmol

4.2 Accuracy of Branch Node

For the branch node, several of the compounds have a similar pruning level as the leaf node which are 10-epi-γ-eudesmol, longifolol, hexadecanol and eudesmol. As for the rest, β-agarafuran, α-agarafuran and γ-eudosmol have a pruning level of 8, 7 and 4 respectively. Figure 10 plotted the accuracy values of branch node for each compound. The corresponding value for branch node is 11.

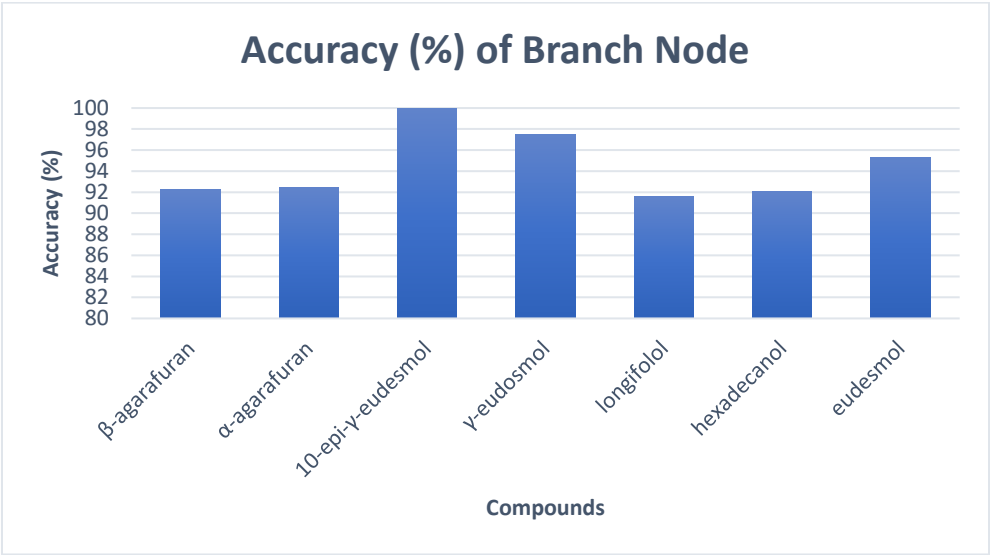

Figure 10 Accuracy of Branch Node

The detailed value of accuracy was tabulated in Table 2. Longifolol have the lowest accuracy which is 91.5451% and hexadecanol is the second lowest with a value of 92.0594%. As for the 10-epi-γ-eudesmol, it also has the highest

accuracy which is 100.0000%. The rest which are γ-eudosmol, eudesmol, β-agarafuran and α-agarafuran has the accuracy of 97.5000%, 95.235%, 92.2106% and 92.4306% respectively.

Table 2 Accuracy of DT for Branch Node

No	Compounds	Accuracy (%)
1	β-agarafuran	92.2106
2	α-agarafuran	92.4306
3	10-epi-γ-eudesmol	100.0000
4	γ-eudosmol	97.5000
5	longifolol	91.5451
6	hexadecanol	92.0594
7	eudesmol	95.2535

4.3 Resubstitution Error and Cross-Validation Error of Branch Node

The error was obtained from the DT modelling by using branch node with 11 as the corresponding value for each compound. From the result, it can be seen that the pattern where Kloss error is higher than the Closs except at 10-epi-γ-eudesmol compound where both of the errors are 0. Figure 11 plotted the errors in branch node.

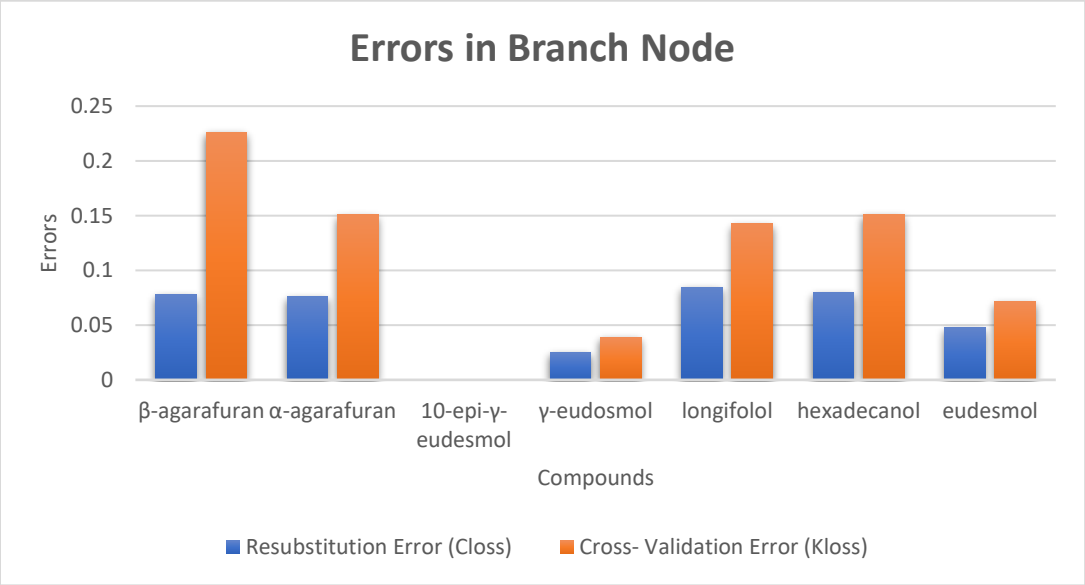

Figure 11 Errors in Branch Node

Table 3 shows the detailed results of the Closs and Kloss errors for the branch node. The results show that the Kloss for β-agarafuran had the highest value, which is 0.2263, while the Closs for longifolol had the highest value at 0.0845. The Closs and Kloss errors for the compound 10-epi-γ-eudesmol had the same lowest error value of 0. In addition, the errors varied, with γ-eudesmol having the second lowest value for both errors at 0.0250 for Closs and 0.0390 for Kloss. The remaining compounds had declining error values, with Closs values of 0.0794 for hexadecanol, 0.0779 for β-agarafuran, 0.0757 for α-agarafuran, and 0.0475 for eudesmol. Similarly, in declining order, the Kloss values for hexadecanol, α- agarafuran, longifolol, and eudesmol were 0.1510, 0.1507, 0.1427, and 0.0712, respectively.

Table 3 Resubstitution Error and Cross-Validation Error for Branch Node

No	Compounds	Resubstitution Error (Closs)	Cross-Validation Error (Kloss)
1	β-agarafuran	0.0779	0.2263
2	α-agarafuran	0.0757	0.1507
3	10-epi-γ-eudesmol	0.0000	0.0000
4	γ-eudosmol	0.0250	0.0390
5	longifolol	0.0845	0.1427
6	hexadecanol	0.0794	0.1510
7	eudesmol	0.0475	0.0712

4.4 Confusion Matrix

Table 4 shows the detailed results of the confusion matrix for the branch node. It revealed that the branch node has confusion matrix of 94.74%. The results show that the model was tested on a total of 19 datasets, of which it correctly predicted 18 datasets as true positives (TP) or true negatives (TN). Only 1 dataset was incorrectly predicted as a false negative (FN), and there were no false positives (FP) in the confusion matrix.

Table 4 Confusion Matrix of Branch Node

Data Class	Predicted as Positive	Predicted as Negative
Positive	2 (TP)	1 (FN)
Negative	0 (FP)	16 (FN)

5. Conclusion

In this study, the modelling of decision tree for agarwood oil quality grading using and branch node was successfully developed. This technique was chosen to accommodate with the issue because currently, there are no decision tree models available for the agarwood oil industry. It was found that branch node resulted in decision tree model with 94.74% accuracy and similar confusion matrices, as they used the same agarwood oil dataset to model the decision tree. The results of the cross-validation for both nodes were also similar, with slightly different values for the Kloss. The pruning levels for each node differed, with some compounds in the branch node having higher pruning levels than the leaf node. The results showed that decision trees are capable of distinguishing between high- and low-quality agarwood oil grades, which is significant and useful for the agarwood oil industry and related research areas. As decision tree techniques are relatively new for classifying agarwood oil, there is potential for further exploration in this area. As a future direction, it is recommended to test decision tree models that can categorize more than two grades of agarwood oil quality and comparing the results of these models.

Acknowledgment

The authors would like to thank Dr. Mohamad Hushnie Haron for the data collection as well as Forest Research Institute Malaysia (FRIM) and BioAromatic Research Centre of Excellence (BARCE), Universiti Malaysia Pahang (UMP). The gratitude also is expressed to all Advanced Signal Processing research members and staff involved from both FRIM and UMP. The financial acknowledgement is dedicated to Institute of Research Management and Innovation (IRMI), Universiti Teknologi MARA, Cawangan Johor, Kampus Pasir Gudang, Malaysia and College of Engineering, Universiti Teknologi MARA (UiTM), Shah Alam, Selangor.

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,” J. Teknol. (Sciences Eng., vol. 68, no. 1, pp. 37–42, 2014.

[2] H. Chhipa and N. Kaushik, “Fungal and bacterial diversity isolated from Aquilaria malaccensis tree and soil, induces agarospirol formation within 3 months after artificial infection,” Front. Microbiol., vol. 8, no. JUL, 2017.

[3] S. Suharti, P. Pratiwi, E. Santoso, and M. Turjaman, “Feasibility Study of Business in Agarwood Inoculation at Different Stem Diameters and Inoculation Periods,” Indones. J. For. Res., vol. 8, no. 2, pp. 114–129, 2011.

[4] Y. Liu, J. Wei, Z. Gao, Z. Zhang, and J. Lyu, “A Review of Quality Assessment and Grading for Agarwood,” Chinese Herbal Medicines, vol. 9, no. 1, pp. 22–30. Jan. 2017.

[5] M. H. Haron, M. N. Taib, N. Ismail, N. A. Mohd Ali, and S. N. Tajuddin, “Statistical analysis of agarwood oil compounds based on GC-MS data,” 2018 9th IEEE Control Syst. Grad. Res. Colloquium, ICSGRC 2018 – Proceeding, no. August, pp. 27–30, 2019.

[6] A. F. M. Amidon et al., “Agarwood Oil Quality Grading using OVO Multiclass Support Vector Machine,” 2021 2nd International Conference on Artificial Intelligence and Data Sciences (AiDAS), pp. 1-5. 2021.

[7] N. Ismail, M. H. F. Rahiman, M. N. Taib, N. A. M. Ali, M. Jamil and S. N. Tajuddin, "The grading of agarwood oil quality using k-Nearest Neighbor (k-NN)," 2013 IEEE Conference on Systems, Process & Control (ICSPC), pp. 1-5. 2013.

[8] Ishihara, M., Tsuneya, T., & Uneyama, K. "Components of the Agarwood Smoke on Heating". Journal of Essential Oil Research, 5(4), 419– 423. 1993.

[9] Ishihara, M., Tsuneya, T., & Uneyama, K. "Fragrant sesquiterpenes from agarwood. Phytochemistry", 33(5), 1147–1155. 1993.

[10] Chippa H, Chowdhary K & Kaushik N, "Artificial production of agarwood oil in Aquilaria sp. by fungi: a review," Phytochemistry Reviews, 16(5), 835-860, 2017.

[11] Y. Song and Y. Lu, "Decision tree methods: applications for classification and prediction," vol. 27, no. 2, pp. 130–135, 2015.

[12] H. Patel and P. Prajapati, "Study and Analysis of Decision Tree Based Classification Algorithms", IJCSE Open Access, no. July 2019, 2018.

[13] Chiu Mei-Hung, Yu Yuh-Ru, Liaw Hongming, Hao Lin, "The Use Of Facial Micro-Expression State And Tree-Forest Model For Predicting Conceptual-Conflict Based Conceptual", ESERA E-Proceeding, Science Education Research: Engaging learners for a sustainable future, 2016.

[14] J. Brownlee, "Master Machine Learning Algorithms," in Discover How They Work and Implement Them From Scratch, Melbourne, 2019, pp. 73- 82.

[15] N. Z. Mahabob et al., "A Study on ANN Performance Towards Three S ignificant Compounds of High-Quality Agarwood Oil," no. May, pp. 116– 120, 2022.

[16] A. Fawwaz et al., "The Grading of Agarwood Oil Quality Based on Multiclass Support Vector Machine (MSVM) Model," no. May, pp. 91–95, 2022.

[17] N. Ismail, M. H. F. Rahiman, M. N. Taib, N. A. M. Ali, M. Jamil and S. N. Tajuddin, "Application of ANN in agarwood oil grade classification," 2014 IEEE 10th International Colloquium on Signal Processing and its Applications, pp. 216-220. 2014.