Malaysian Journal of Chemistry, 2025, Vol. 27(1), 110-121
(Special Issue: *International Conference on Natural Products, ICNP 2024*)

Investigation of the Clustering of High Quality Stingless Bee Honeys using Unsupervised Machine Learning Models

Yusnaini Md Yusoff[1]*, Nalinah Poongavanam[2,3], Jalifah Latip[3], Mohd Razif Mamat[4], Lim Seng Joe[5], Wardah Mustafa Din[1] and Dian Indrayani Jambari[6]

[1]Pusat Pengajian Citra Universiti, Universiti Kebangsaan Malaysia, UKM, 43600 Bangi, Selangor, Malaysia
[2]Centre for Foundation and General Studies, Infrastructure University Kuala Lumpur, Unipark Suria, Jalan Ikram-Uniten, 43000 Kajang, Selangor, Malaysia
[3]Department of Chemical Sciences, Faculty of Science and Technology, Universiti Kebangsaan Malaysia, UKM, 43600 Bangi, Selangor, Malaysia
[4]Malaysia Genome and Vaccine Institute, National Institutes of Biotechnology Malaysia, Jalan Bangi, 43000 Kajang, Selangor, Malaysia
[5]Department of Food Sciences, Faculty of Science and Technology, Universiti Kebangsaan Malaysia, UKM, 43600 Bangi, Selangor, Malaysia
[6]Center for Software Technology and Management, Faculty of Information Science & Technology, Universiti Kebangsaan Malaysia, UKM, 43600 Bangi, Selangor, Malaysia
*Corresponding author (e-mail: yusnaini@ukm.edu.my)

Honey quality and authenticity are crucial due to its health benefits and rising demand, yet challenges like environmental factors and adulteration persist. This study evaluated 106 honey samples for quality, bee species distribution, and patterns using machine learning models. Unsupervised clustering techniques, including K-Means, Agglomerative, Hierarchical Clustering, and DBSCAN, were applied. Component plane analysis of the Self-Organizing Map (SOM) highlighted key clustering factors. Hierarchical clustering (unscaled dendrogram) outperformed others with a Silhouette score of 0.351, a Davies-Bouldin Index of 0.977, and a Cophenetic Correlation Coefficient of 0.709. Quality was assessed based on pH, moisture content, sugar levels, and 5-hydroxymethylfurfural (HMF) using the Malaysian Standard for stingless bee Honey (MS 2683:2017) and Codex Alimentarius guidelines. All samples met quality standards, indicating freshness and high quality. Four distinct clusters emerged with unique physicochemical properties and species distributions. The application of various unsupervised clustering techniques (e.g., K-Means, Hierarchical Clustering, DBSCAN) and a Self-Organizing Map (SOM) for analyzing honey quality and bee species distribution is innovative. While honey quality assessments are common, incorporating advanced data analytics to uncover patterns and relationships is relatively novel.

Keywords: Stingless bee honey; Malaysia; quality; unsupervised machine learning; clustering

Received: September 2024; Accepted: December 2024

As a dietary supplement and for general health maintenance, honey is used by many. Also, according to [1], honey can slow down fatigue and accelerate the speed of physical recovery. According to [2], honey is commonly used as a sweetener in both food and drinks. Honey is a common sweetener in Malaysian restaurants and cafes. It adds flavor and works as an ingredient in drinks. In addition, honey-flavored beverages can improve insulin sensitivity, blood sugar levels, and running performance [3]. Honey sold in stores may be intentionally contaminated with additives including C4 and C3 sugars or with sugar fed to bees prior to harvest. On the other hand, the genuineness of the honey can be compromised by specific circumstances. Storage, high heat, and excessive moisture can all cause honey to ferment, which in turn can alter its quality [4]. The health advantages of honey are compromised when it is adulterated. Even though honey improves health, the adulterants found in honey are harmful to humans. If a diabetic takes honey that has been contaminated with extra sugars, for instance, their blood sugar level could rise. A person's risk of obesity and type 2 diabetes are both exacerbated by the additional calories that come from eating added sugar [5]. Several factors, including the honey's processing and storage conditions, its exposure to various climates and seasons, and its geographical origin, can impact its quality, as pointed out by [6]. Also, the presence or absence of adulterants is a factor in honey quality. Geographic origin, manufacturer's packaging, and available information on the honey sold are the three main factors that influence customers' honey purchasing decisions [7]. After learning that foreign honey is often contaminated, consumers also choose to buy local honey [8]. But it's unfair to the honey farmers and their product to judge its quality just by looking at the headlines. A quality list of honey can provide transparent labeling on

†Paper presented at the International Conference on Natural Products 2024 (ICNP 2024)

honey quality, making it easier for the public to find and buy according to their purchasing power. Honey quality regulations vary by region but follow widely recognized frameworks [9]. In the EU, the Honey Directive (2001/110/EC) ensures honey is pure and unaltered, banning additives like sugar. The General Food Law (Regulation (EC) No 178/2002) mandates food safety, traceability, and contaminant-free products. Exporting countries must submit Residue Monitoring Plans (RMP) and provide health certificates for honey shipments. In the U.S., the FDA enforces strict labeling and composition standards, while the USDA grades honey based on clarity, flavor, and quality. Internationally, Codex Alimentarius guidelines [10], developed by Food and Agriculture Organization FAO and World Health Organization (WHO), set global benchmarks for honey quality and trade.

For both culinary and medicinal uses, stingless bee honey is well-known among honey connoisseurs in Malaysia. There is belief that stingless bee honey can aid in anti-aging, strengthen immune, kill bacteria, and alleviate respiratory issues like cough, sore throat, and phlegm [11]. According to several studies, the stingless bees from *Heterotrigona itama* create the vast majority of Malaysia's stingless bee honey [12 – 14]. It is vital to verify that the honey sold in the market is safe to eat because honey is widely used by the community. If the honey is grouped according to quality, buyers will have faith in it. Different species of bees, different regions, and different types of vegetation can all affect the final honey's quality. Honey quality can be assessed by looking at factors such as moisture, acidity, pH, hydroxymethylfurfural (HMF), conductivity, sugar, and proline concentration [2, 15]. The Malaysian Standard for Stingless Bee Honey (MS 2683:2017) provides comprehensive guidelines for the quality and safety of stingless bee honey [16]. It sets parameters for physicochemical properties, labeling, and safety to ensure the honey meets quality requirements. The implementation of a honey grading system in Malaysia is crucial to ensure consistent quality, enhance consumer confidence, and elevate the global competitiveness of Malaysian honey. Grading honey based on parameters such as clarity, flavor, moisture content, and purity would provide a clear benchmark for quality, helping consumers make informed choices. For example, the USDA grading system categorizes honey into grades such as Grade A, Grade B, and Grade C based on these attributes [17]. Additionally, a standardized grading system could incentivize local producers to adhere to higher quality standards, reduce the prevalence of adulteration, and support the branding of Malaysian honey as a premium product. Despite its growing industry, Malaysia has yet to adopt such a system, leaving a significant opportunity to improve product differentiation and market positioning.

Cluster analysis, in which individuals are divided into smaller groups based on the similarity between the data, is made possible by unsupervised algorithms [18]. Common data processing tasks that clustering processes aid in include pattern recognition and picture analysis. When dealing with bigger datasets, a good algorithm will do its job quickly. Research done by [19] and [20] both note that the K-Means clustering approach is quite popular because of how efficient it is and how easy it is to apply. According to [21], it can quickly process a vast amount of data. Before using the K-Mean algorithm, decide on the number of clusters. When there are multiple clusters, it means that each cluster contains unique patterns; nonetheless, there will be repercussions from making too many copies of clusters. Finding the right cluster number is essential for avoiding the inconsistencies. The K-Means clustering algorithm was enhanced by combining it with the Elbow approach. If you have a lot of data, the elbow approach can tell you how many clusters to use [21]. There are two main types of hierarchical clustering: agglomerative and divisive. Analysis in the agglomerative approach begins with individual element clusters (N clusters) and works its way up to larger groups of clusters. In contrast, the Divisive technique employs top-down analysis, starting with a single cluster and thereafter splitting into numerous clusters as it descends the hierarchical dendrogram [18, 22]. The dendogram's partition count is used to determine the number of clusters. Due to the difficulty of this approach, the Agglomerative clustering algorithm makes use of two well-known methods—the silhouette width and the gap statistic—to determine the total number of clusters [18].

The increasing global demand for honey, coupled with its culinary and medicinal benefits, highlights the critical need for reliable quality assurance measures. In Malaysia, stingless bee honey is highly valued for its health benefits, yet the lack of a standardized grading system and consistent quality control undermines consumer trust and limits its global market potential. Additionally, honey adulteration with sugars, improper storage conditions, and environmental factors such as climate and vegetation affect its authenticity and quality. These issues necessitate robust methods to assess and categorize honey based on physicochemical properties and safety standards. This study aims to address these challenges by evaluating the quality of stingless bee honey, employing machine learning clustering techniques to identify patterns and variations, and emphasizing the importance of implementing a comprehensive grading system to enhance consumer confidence and industry competitiveness.

Table 1. Legend for the Codes of Sample.

State	Code	Location	Code	Species	Code
Negeri Sembilan	01	Terachi	01	*Heterotrigona itama*	01
		Seri Menanti	02	*Geniotrigona thoracica*	02
		Simpang Durian	03	*Tetrigona binghami* *Lophotrigona conifrons*	 03
Pahang	02	Chenur	01	*Tetrigona apicalis*	04
		Simpang Pelangai 1	02	*Tetrigona melanoleuca*	
		Simpang Pelangai 2	03	*Heterotrigona*	05
Wilayah Persekutuan Putrajaya	03	Laman Stingless bee Taman Botani	01	*erythogastra*	06
		Kebun Stingless bee D'Rimba Desa P9	02		07
		Kebun Stingless bee D'Putra Rimba P15	03		
		Kebun Stingless bee Komuniti Presint 16	04		
Selangor	04	Gombak	01		
Perak	05	Tanjung Malim	01		

K01010101 represents a sample of stingless bee (K) from Negeri Sembilan (01), specifically from Terachi (01). The species of the bees is *Heterotrigona itama* (01), and this particular sample is numbered 1 (01).

EXPERIMENTAL

Chemicals and Materials

Stingless Bee Honey Sampling

The total of 106 stingless bee honey samples were collected from five states: Negeri Sembilan, Pahang, Selangor, Perak, and Wilayah Persekutuan Putrajaya. The specimens were gathered between April and May 2024. The stingless bee species collected included *Heterotrigona itama*, *Geniotrigona thoracica*, *Tetrigona binghami*, *Lophotrigona conifrons*, *Tetrigona apicalis*, *Tetrigona melanoleuca* and *Heterotrigona erythrogastra*. The bees were raised in man-made hives constructed from wooden boxes or wooden stakes. The samples were labeled using a combination code based on four criteria: state, location, species, and replicate of samples (**Table 1**).

Physicochemical Methods

Analysis of Hydroxymethylfurfural (HMF)

The HMF concentration in the stingless bee honey samples was determined using a reflectometer (Model RQflex 20) with a measurement range of 1.0–60.0 mg/L. The reflectometer was calibrated using a high-molecular-weight (HMF) barcode strip. Each stingless bee honey sample, weighing 2.5 g, was diluted with 10 mL of distilled water. The HMF test strip was submerged in the prepared solution for two seconds. The strip was then separated and cleaned to remove any excess solution before being inserted into the device.

The Presence of Lipids and Proteins

Lipids and proteins typically found in raw honey, was identified using the RapidRAW method by the Malaysia Genome and Vaccine Institute (MGVI) [23]. Five drops of normal saline pH 7.0 (Reagent 1) were mixed with three drops of stingless bee honey sample and stirred until a homogeneous solution was obtained. Subsequently, seven drops of RapidRAW reagent (Reagent 2) were added and thoroughly blended. The mixture was then left undisturbed for two minutes to allow the formation of precipitate.

Measurement of Moisture Content

The moisture content was evaluated using a portable refractometer (Model ATC). Three drops of stingless bee honey sample were placed onto the prism coating and then covered with the prism cover. The moisture content of the honey was measured and recorded through the eyepiece after the sample was applied to the prism screen.

The Sugar Content

The sugar concentration was quantified using a portable refractometer (Model ATC). The same methods were employed as in the moisture content evaluation.

pH

A 2.5 g portion of stingless bee honey was mixed with 19 mL of distilled water in a 100 mL container. The pH of the sample was measured using a pH meter (Trans Instrument BP3001 model). Prior to analysis, the pH meter was calibrated with standard solutions of pH 4, 7, and 10.

Unsupervised Machine Learning

The 106 stingless bee honey samples were clustered using unsupervised machine learning techniques in the Python programming language. The clustering algorithms employed included K-Means, Agglomerative, Hierarchical, and DBSCAN clustering. The elbow method was used to determine the optimal number of clusters for these stingless bee honey samples. Additionally, the ideal number of principal components was identified through cumulative explained variance plotting. Two types of scalers, specifically Standard Scaler and MinMax Scaler, were applied to normalize the scale, alongside an unscaled dataset in which refers to data that has not been normalized to adjust its scale or range, to perform the clustering methods. The Silhouette Score, Davies-Bouldin Index, and Cophenetic Correlation Coefficient Score were used to identify the most suitable unsupervised machine learning model for the tested stingless bee honey.

RESULTS AND DISCUSSION

The data in **Table 2** illustrates the spatial distribution of diverse bee species across distinct regions in multiple states. *Heterotrigona itama* is a commonly found species in Negeri Sembilan, specifically in Seri Menanti, Simpang Durian, and Terachi. Other species such as *Geniotrigona thoracica* and *Tetrigona binghami* are also present in the area. *Heterotrigona itama* remains the most often documented species in Pahang, specifically in Chenur and Simpang Pelangai, among a small number of other species. Perak exhibits a notable abundance of *Heterotrigona itama*, although Putrajaya predominantly documented the occurrence of this species in several areas. The Gombak area in Selangor exhibited the highest species diversity, encompassing *Geniotrigona thoracica*, *Heterotrigona itama*, and *Tetrigona binghami*. In general, *Heterotrigona itama* is the species that is found in the greatest number of states.

Table 2. The distribution of raw stingless bee honey used.

State	Area	Species	No. of samples
Negeri Sembilan	Seri Menanti	*Heterotrigona itama*	4
Negeri Sembilan	Simpang Durian	*Heterotrigona itama*	10
Negeri Sembilan	Terachi	*Geniotrigona thoracica*	2
Negeri Sembilan	Terachi	*Heterotrigona itama*	6
Negeri Sembilan	Terachi	*Tetrigona binghami*	3
			25
Pahang	Chenur	*Geniotrigona thoracica*	1
Pahang	Chenur	*Heterotrigona itama*	5
Pahang	Chenur	*Lophotrigona conifrons*	2
Pahang	Simpang Pelangai 1	*Heterotrigona itama*	7
Pahang	Simpang Pelangai 2	*Heterotrigona itama*	7
			22
Perak	Tanjung Malim	*Geniotrigona thoracica*	4
Perak	Tanjung Malim	*Heterotrigona itama*	11
			15
Putrajaya	Kebun Stingless bee D'Putra Rimba P15	*Heterotrigona itama*	5
Putrajaya	Kebun Stingless bee Desa P9	*Heterotrigona itama*	5
Putrajaya	Kebun Komuniti P16	*Heterotrigona itama*	5
Putrajaya	Taman Stingless bee Taman Botani	*Geniotrigona thoracica*	1
Putrajaya	Taman Stingless bee Taman Botani	*Heterotrigona itama*	15
			31
Selangor	Gombak	*Geniotrigona thoracica*	2
Selangor	Gombak	*Heterotrigona itama*	4
Selangor	Gombak	*Heterotrigona erythogastra*	1
Selangor	Gombak	*Tetrigona apicalis*	2
Selangor	Gombak	*Tetrigona binghami*	2
Selangor	Gombak	*Tetrigona melanoleuca*	1
			13

Table 3. The range of Five Type of Physicochemical Parameter Analysis According to the locations.

	Physicochemical Parameter					
State	**Number of sample (N = 106)**	**Rapidraw_pH solution**	**Sugar Content (%)**	**pH**	**HMF (mg/L)**	**Moisture Content (%)**
Negeri Sembilan	25	3.267 - 3.906	65.0 - 74.0	2.96 - 3.84	1.4 - 3.8	24.2 - 32.5
Pahang	22	3.172 - 3.971	66.0 - 74.0	2.93 - 3.81	1.3 - 4.0	24.2 - 32.0
Perak	15	2.863 - 3.766	68.0 - 74.0	2.72 - 3.56	1.0 - 4.2	24.2 - 30.0
Putrajaya	31	3.104 - 4.220	68.0 - 74.0	2.93 - 3.86	1.6 - 6.4	24.2 - 30.0
Selangor	13	3.080 - 3.586	62.7 - 73.0	2.75 - 3.28	1.0 - 2.7	25.0 - 35.3

Table 3 presents data obtained from measurements conducted under different conditions for several physicochemical parameters. The honey is efficiently identified through a screening process. A solid precipitate can be generated to detect the presence of pollen and microorganisms in honey. A study conducted in 2019 by [23] revealed that solutions containing pure honey can exhibit a range of colors, ranging from pale green to a greenish blue. All 106 samples produced a solution that ranged in hue from green to blue, resulting in the development of precipitate (data not shown). These products create the illusion of being genuine honey. The sugar concentration in honey is regulated by the Malaysian Standard for stingless bee honey Specification: MS 2683 (2017) and Codex Alimentarius. These guidelines specify that the total sugar concentration in honey should fall within the range of 60% to 85%. It is worth noting that the findings from this study remain valid. The findings suggest that the honey is of high quality, as the study maintained rigorous standards in collecting all 106 samples on site. Various studies have indicated that honey tends to have an acidic nature, as evidenced by research conducted by [24 – 26]. In this study, the pH readings of stingless bee honey, as determined by the pH meter and RapidRAW method, ranged from 2.75 to 4.23. These values align with the pH range specified by MS 2683 and CODEX standards. It is essential to monitor the HMF levels to maintain the quality and freshness of Malaysian honey, including stingless bee honey. Greater amounts of HMF can suggest potential quality issues, serving as a dependable indicator of the honey's age and storage conditions. According to the standards set in 2017, the permissible limit for 5-HMF in honey is 30.0 mg per kilogram [27]. However, according to Codex (2001), it is recommended to limit the intake of HMF to no more than 80 mg/kg. According to the results of this study, it seems that all the samples under investigation are fresh and unadulterated. All samples must have a maximum moisture content of 35.0% or less, and in this case, they all fall within this range from MS 2683.

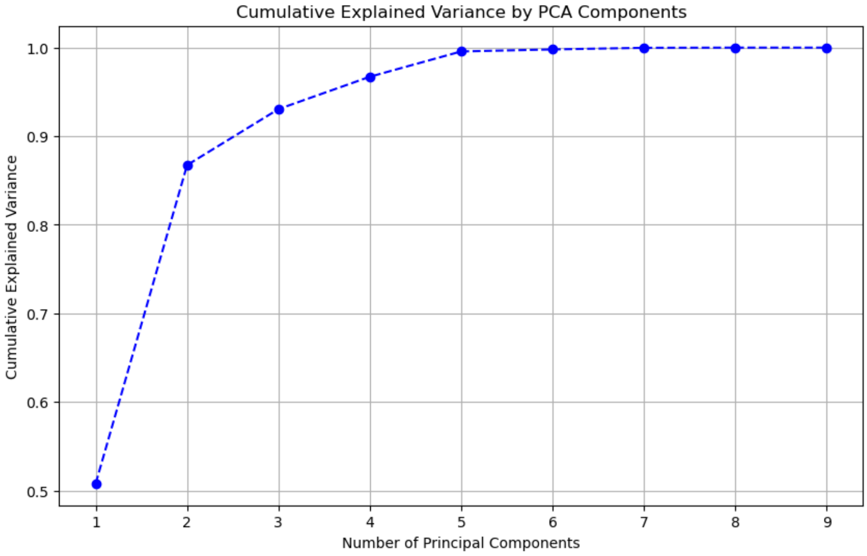

Figure 1. The cumulative explained variance plot. The optimal number of principal components for a minimum of 90% of variance, (n= 3).

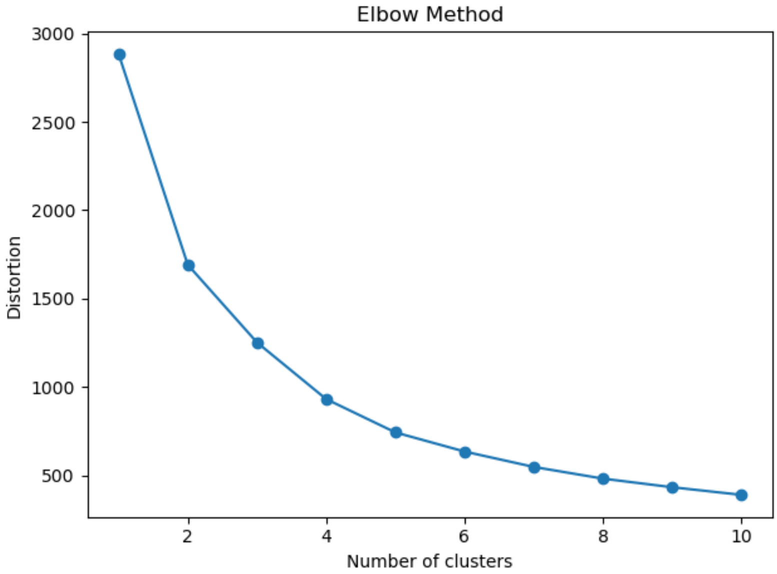

Figure 2. The Elbow method provides an efficient and meaningful number of clusters to balance complexity and performance in their clustering tasks. The optimal k is typically at the elbow point, representing the best balance between compact clusters and minimal redundancy, (k=3).

The unique physicochemical characteristics and potential health benefits of stingless bee honey, produced by stingless bees, have attracted considerable attention. Understanding these traits is crucial for classifying and ensuring the quality of stingless bee honey. Machine learning, particularly unsupervised learning methods such as clustering, offers a powerful tool for comprehending complex data patterns in this scenario. Unsupervised learning, as opposed to supervised learning, does not rely on labelled data. This makes it ideal for exploratory data analysis, where the goal is to uncover hidden patterns within the data. For achieving optimal clustering of stingless bee honey, several unsupervised machine learning algorithms were selected, including K-Means, Agglomerative, Hierarchical, and DBSCAN. These algorithms were chosen based on their ability to analyze the physicochemical parameters of the honey. The accuracy of unsupervised machine learning for clustering was maximized by identifying the optimal number of clusters and principal components using the Elbow method and cumulative explained variance plotting.

The eigenvalues are utilized to assess the extent to which the initial principal components capture the overall variance. The plot of cumulative explained variance (**Figure 1**) is useful for determining the optimal number of principal components to keep for clustering purposes. As an example, one approach could be to retain components that account for a minimum of 90% of the variance. In this scenario, this would result in retaining three principal components.

One approach is to perform K-Means clustering with varying values of (k), representing the number of clusters, and then compute the within-cluster sum of squares (WCSS) for each (k). The within-cluster sum of squares (WCSS) is calculated by summing the squared distances between each point and the centroid of its cluster. At a certain juncture, the rate of decline begins to decelerate, resulting in a distinct "elbow" formation. This point signifies the ideal number of clusters. Adding additional clusters beyond this point does not have a significant impact on reducing the within-cluster sum of squares (WCSS). The elbow plot demonstrates a significant decrease in WCSS until reaching a point of inflection at k = 3, after which the decline becomes more gradual (**Figure 2**). It is probable that k = 3 represents the optimal number of clusters [28]. Ensuring that machine learning algorithms perform optimally requires careful attention to scaling, treating each feature consistently regardless of its original scale. Scaling is crucial for achieving optimal model performance as it guarantees that every feature has an equal impact on the model. Without proper scaling, models may place too much emphasis on features with larger numerical values, which can result in subpar generalization and performance. In this study, two common scalers, StandardScaler and MinMaxScaler, were emphasized. StandardScaler and MinMaxScaler are two common methods for normalizing data in machine learning [29]. Standard Scaler transforms data to have a mean of 0 and a standard deviation of 1, making features with different scales comparable. MinMaxScaler, on the other hand, scales data to a specified range, usually between 0 and 1, preserving the relative relationships between values. Both methods ensure that no single feature dominates due to its scale, improving the performance of machine learning algorithms sensitive to data magnitude.

Table 4. The metrics used on multiple unsupervised machine learning modelling.

Unsupervised Machine Learning Modelling	Silhouette Score	Davies-Bouldin Index	Cophenetic Correlation Coefficient
K-Means Not Scaled	0.414	0.822	
K-Means StandardScaler	0.252	1.336	
K-Means MinMaxScaler	0.275	1.340	
Agglomerative Not Scaled	0.418	0.812	
Agglomerative StandardScaler	0.202	1.499	
Agglomerative MinMaxScaler	0.223	1.534	
Hierachical Dendogram Not Scaled	0.351	0.977	0.709
Hierachical Dendogram StandardScaler	0.202	1.499	0.650
Hierachical Dendogram MinMaxScaler	0.223	1.534	0.508
DBSCAN Not Scaled	0.305	1.186	
DBSCAN StandardScaler	0.428	0.965	
DBSCAN MinMaxScaler	-	-	

*Value highlighted in grey were the selected modelling for giving better clustering

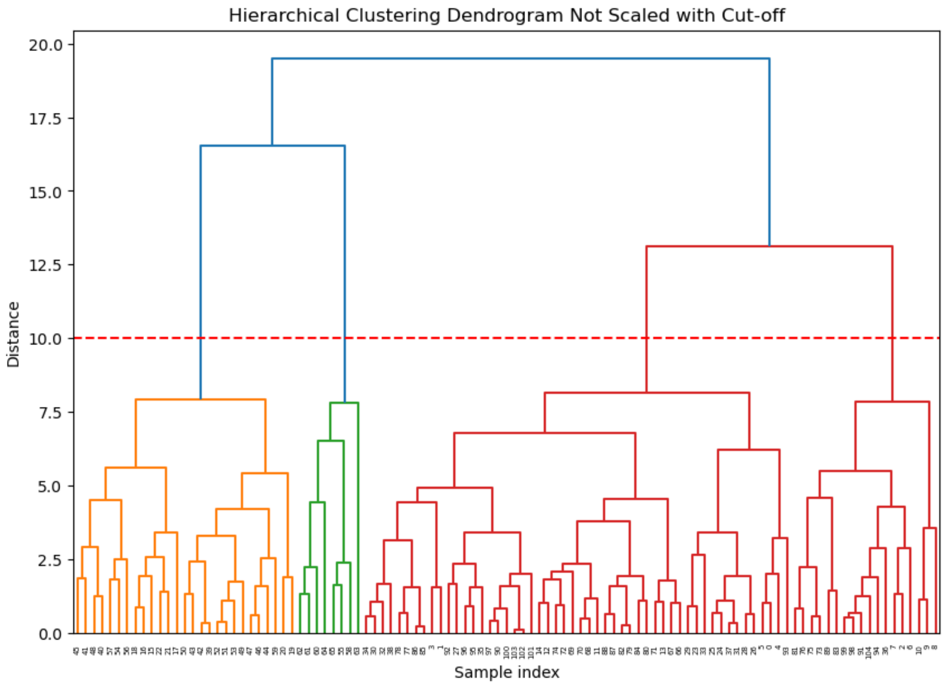

Figure 3. The cutoff distance defines the threshold for merging clusters and determines the number of clusters formed. Hierachical Dendogram Not Scaled with the cut off red line at a distance:10 which suggest up to four separate cluster.

In addition, three criteria were assessed to evaluate the clustering model's performance. The Cophenetic Correlation Coefficient measures how well the dendrogram reflects the distances between the original data points. The Silhouette Score measures the resemblance of a data point to its own cluster compared to other clusters. A higher score indicates a greater level of distinctness in the clusters. The Davies-Bouldin Index computes the mean similarity ratio between each cluster and the cluster that has the highest resemblance to it. A lower number indicates better grouping. Furthermore, to gain a more comprehensive understanding of the optimal clustering model, either a scatter plot or a hierarchical clustering dendrogram was constructed.

Among the unsupervised machine learning models tested, including K-Means, Agglomerative, Hierarchical, and DBSCAN, both the unscaled datasets and the DBSCAN with StandardScaler produced strong metric scores, making them contenders for the best clustering model (**Table 4**). After evaluating the visualizations from each model, the hierarchical dendrogram (unscaled) was selected as the final model (**Figure 3**). This model achieved notable metric scores: 0.351 for the Silhouette score, 0.977 for the Davies-Bouldin Index, and 0.709 for the Cophenetic Correlation Coefficient. Ultimately, four clusters were identified at a linkage distance of 10, a point where clear separations between clusters were observed (**Figure 3**).

Table 5. The clustering, as shown in **Figure 3**, is based on bee species and their corresponding states.

Cluster	Bee Species	By State
1	*Heterotrigona itama* *Geniotrigona thoracica* *Tetrigona binghami*	Negeri Sembilan Putrajaya Pahang Perak
2	*Heterotrigona itama* *Geniotrigona thoracica*	Negeri Sembilan Putrajaya Pahang Perak
3	*Heterotrigona itama* *Geniotrigona thoracica* *Tetrigona binghami* *Tetrigona apicalis* *Tetrigona melanoleuca* *Heterotrigona erythogastra*	Selangor
4	*Heterotrigona itama* *Geniotrigona thoracica* *Lophotrigona conifrons*	Pahang Putrajaya Selangor

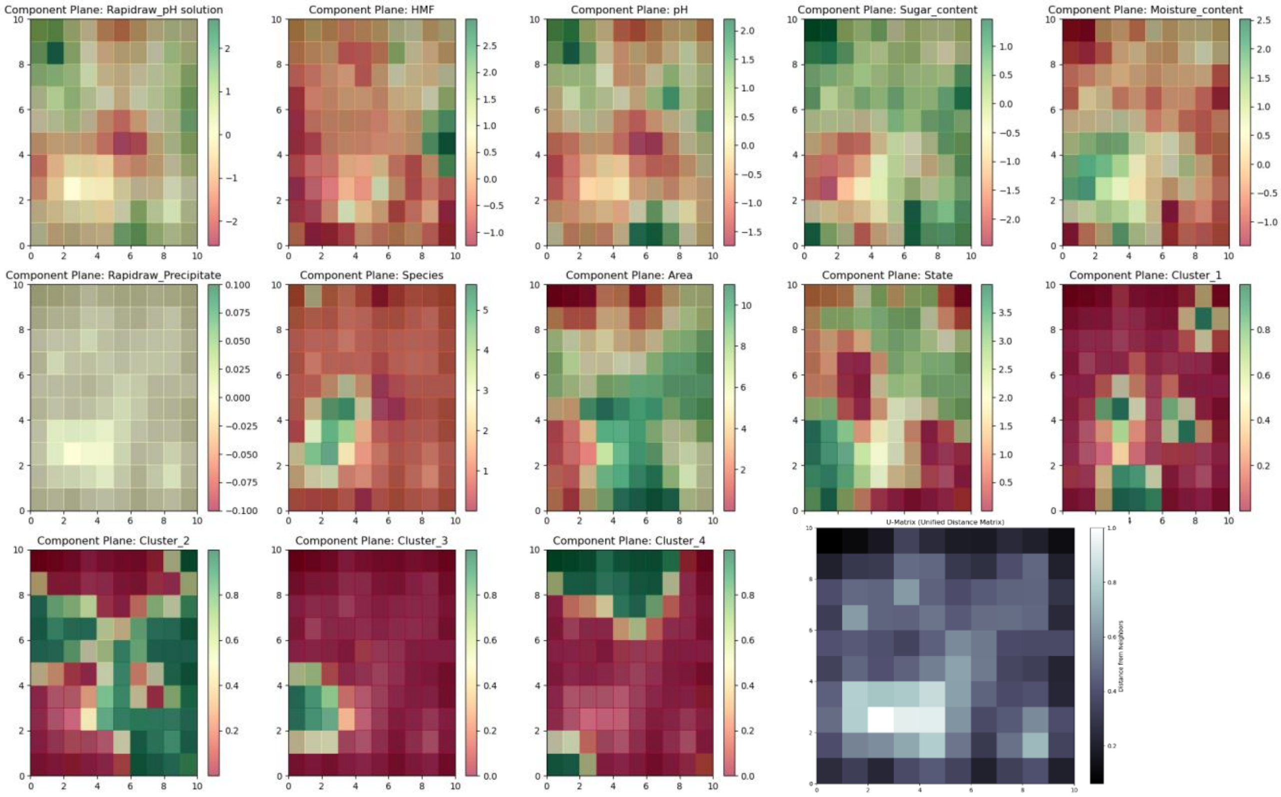

Figure 4: SOM representing the relationship of all variables against the cluster formed in Figure 3. The relationship between SOM component plane forms clusters that are represented in U-matrix. Each component plane in the SOM represents variables used in the study which are Species, Rapidraw_Precipitate, Rapidraw_pH solution, HMF, pH, Sugar_content, Moisture_content, Area, State, Cluster_1, Cluster_2, Cluster_3 and Cluster_4.

The clustering of stingless bee honey samples reveals distinct groupings based on bee species and geographic distribution, highlighting unique characteristics of each cluster (**Table 5**). Cluster 1 encompasses *Heterotrigona itama, Geniotrigona thoracica, and Tetrigona binghami* and spans the states of Negeri Sembilan, Putrajaya, Pahang, and Perak, suggesting a diverse yet specific combination of bee species within these regions. Cluster 2, though geographically identical to Cluster 1, includes only *Heterotrigona itama* and *Geniotrigona thoracica*, differentiating itself by the absence of *Tetrigona binghami*. Cluster 3, centered in Selangor, features the most diverse range of species, including *Heterotrigona erythogastra, Tetrigona apicalis, and Tetrigona melanoleuca*, in addition to the common *Heterotrigona itama* and *Geniotrigona thoracica*. This unique diversity makes it a notable outlier. Cluster 4 includes *Heterotrigona itama*, *Geniotrigona thoracica*, and *Lophotrigona conifrons*, with a distribution spanning Pahang, Putrajaya, and Selangor. The inclusion of *Lophotrigona conifrons* sets Cluster 4 apart, reflecting its distinct ecological niche. These clusters illustrate how bee species composition and geographic distribution interplay to create unique profiles for each cluster.

The distribution of different characteristics across the SOM grid is represented by individual planes, such as Species, Rapidraw_Precipitate, Rapidraw_pH solution, HMF, pH, Sugar_content, Moisture_content, Area, State, and Cluster. These planes are superimposed on the U-Matrix to show how they influence the development of clusters (**Figure 4**). By combining the U-Matrix with individual planes, it becomes possible to identify which features influence the boundaries or similarities between clusters. For example, areas with distinct feature distributions across planes might align with darker regions on the U-Matrix, indicating strong boundaries between clusters driven by those features. Conversely, homogeneous distributions across planes often correspond to lighter regions, suggesting closely related clusters. This relationship between the U-Matrix and feature planes provides a comprehensive view of the clustering process, linking feature variations to cluster formations in a visually interpretable manner. In **Figure 4**, it is also possible to identify the relationship between different component plane of clusters with each component plane related to the physicochemical variable.

Table 6: The legend for categorical data for machine learning approaches.

Legend for Species		**Legend for Rapidraw_Precipitate**		**Legend for Area**		**Legend for State**		**Cluster**	
Geniotrigona thoracica	0	YES	0	Chenur	0	Negeri Sembilan	0	No	0
Heterotrigona itama	1			Gombak	1	Pahang	1	Yes	1
Heterotrigona erythogastra	2			Kebun Stingless bee D'Putra Rimba P15	2	Perak	2		
Lophotrigona conifrons	3			Kebun Stingless bee Desa P9	3	Putrajaya	3		
Tetrigona apicalis	4			Kebun Komuniti P16	4	Selangor	4		
Tetrigona binghami	5			Seri Menanti	5				
Tetrigona melanoleuca	6			Simpang Durian	6				
				Simpang Pelangai 1	7				
				Simpang Pelangai 2	8				
				Taman Stingless bee Taman Botani	9				
				Tanjung Malim	10				
				Terachi	11				

Table 6 highlights the legend for categorical data used to interpret the Self-Organizing Maps (SOMs) based on clusters. By combining the illustration from **Figure 4** and information from **Table 6,** this study reveals that Cluster 1 exhibits a prevalence of lower pH value, lower HMF, higher moisture content, higher sugar content, and is in the area encompassing Taman stingless bee Taman Botani, Tanjung Malim, and Terachi. Cluster 2 is characterized by a higher pH value, lower HMF concentration, higher sugar content, and moderate to low moisture content. The species found in this cluster include *Geniotrigona thoracica*, *Heterotrigona itama*, and *Heterotrigona erythogastra*. These species are found in the areas of Simpang Durian, Simpang Pelangai 1, Simpang Pelangai 2, Taman stingless bee Taman Botani, Tanjung Malim, and Terachi. Cluster 3 is characterized by a lower pH value, lower HMF, lower sugar content, higher moisture content, and the presence of species such as *Tetrigona apicalis*, *Tetrigona binghami*, and *Tetrigona melanoleuca*. These species are found in the areas of Chenur, Gombak, and Kebun stingless bee D'Putra Rimba P15, which are in Putrajaya and Selangor. Finally, Cluster 4 is characterized by a moderate pH value, mostly with higher HMF, higher sugar content, lower moisture content, and the presence of species such as *Geniotrigona thoracica* and *Heterotrigona itama*. This cluster is found in the areas of Chenur, Gombak, and Kebun stingless bee D'Putra Rimba P15 in Putrajaya and Selangor.

The study conducted by [30] examined the variability of sucrose, reducing sugars, acidity, and HMF content in honey samples within the jurisdiction of the honey cluster. A study conducted by [31] further supports the notion that moisture content plays a significant role in influencing cluster variation. Based on present study, species and HMF are likely to have a significant impact on cluster differentiation. A previous study by [32] identified K, Ca, Mg, and Na as the primary factors influencing the differentiation of honeys. [33] also incorporates additional features, such as mineral contents and colour parameters, to enhance the clustering of honey. It is worth noting that the physicochemical properties analyzed in current study fell within the acceptable range according to the Malaysian Standard for stingless bee honey Specification: MS 2683 (2017) and Codex Alimentarius.

CONCLUSION

Overall, the study analyzed the spatial distribution of different bee species in various regions across multiple states. It was found that *Heterotrigona itama* was the most observed species, especially in Negeri Sembilan, Pahang, Perak, and Putrajaya. Selangor displayed the highest species diversity, with the presence of *Geniotrigona thoracica, Heterotrigona itama*, and *Tetrigona binghami*. The physicochemical parameters of honey samples were evaluated in accordance with Malaysian standards and Codex Alimentarius. These parameters included pH, moisture content, sugar content, and HMF levels. The findings indicated that the 106 stingless bee honey samples exhibited exceptional quality, freshness, and adherence to the acceptable ranges for crucial parameters, thus implying their genuineness. When it comes to clustering, different unsupervised machine learning models were assessed. The hierarchical dendrogram (unscaled) was ultimately chosen as the final model, successfully identifying four distinct clusters. Through component plane analysis, it was observed that each cluster displayed distinct characteristics related to pH, HMF, moisture content, sugar content, and species distribution across various regions. The findings presented in this study highlight the significance of closely monitoring physicochemical parameters and spatial distribution to gain a comprehensive understanding of honey quality and the ecological distribution of bee species in different regions. Further research should prioritize the expansion of the dataset to encompass a wider range of regions, botanical sources and species which also can provide better insight for introducing grading system of stingless honey clustering. Additionally, the application of more sophisticated machine learning techniques can greatly enhance our comprehension of honey quality in various environmental conditions.

ACKNOWLEDGEMENTS

This study was funded by GGPM-2023-044 from Universiti Kebangsaan Malaysia. The authors also would like to express their heartfelt gratitude to Amrina Dynamic Ventures, Sinar Alam Enterprise, Botany Department, Landscape and Park Department, Putrajaya, Koperasi Komuniti Penduduk Putrajaya Bhd, Komuniti Presint 15H4 and Komuniti Bandar Presint 16 for their invaluable support and contributions to this study.

REFERENCES

1. Tang, H. (2022) [Retracted] Honey on Basketball Players’ Physical Recovery and Nutritional Supplement. *Computational Intelligence and Neuroscience*, **2022(1)**, 6953568.

2. Zapata-Vahos, I. C., Henao-Rojas, J. C., Yepes-Betancur, D. P., Marín-Henao, D., Giraldo Sánchez, C. E., Calvo-Cardona, S. J., David, D. and Quijano-Abril, M. (2023) Physicochemical parameters, antioxidant capacity, and antimicrobial activity of honeys from tropical forests of Colombia: Apis mellifera and Melipona eburnea. *Foods*, **12(5)**, 1001.

3. Ahmad, N. S., Ooi, F. K., Ismail, M. S. and Mohamed, M. (2015) Effects of post-exercise honey drink ingestion on blood glucose and subsequent running performance in the heat. *Asian Journal of Sports Medicine*, **6(3)**.

4. Rachineni, K., Kakita, V. M. R., Awasthi, N. P., Shirke, V. S., Hosur, R. V. and Shukla, S. C. (2022) Identifying type of sugar adulterants in honey: Combined application of NMR spectroscopy and supervised machine learning classification. *Current Research in Food Science*, **5**, 272–277.

5. Goran, M. I., Ulijaszek, S. J. and Ventura, E. E. (2013) High fructose corn syrup and diabetes prevalence: a global perspective. *Global Public Health*, **8(1)**, 55–64.

6. Sharma, R., Thakur, M., Rana, K., Devi, D. and Bajiya, M. R. (2023) Honey, its quality and composition and their responsible factors. *International Journal of Bio-resource and Stress Management*, **14(1)**, 178–189.

7. Wu, S., Fooks, J. R., Messer, K. D. and Delaney, D. (2015) Consumer demand for local honey. *Applied Economics*, **47(41)**, 4377–4394.

8. Al-Waili, N., Salom, K., Al-Ghamdi, A. and Ansari, M. J. (2012) Antibiotic, pesticide, and microbial contaminants of honey: human health hazards. *The Scientific World Journal*, **2012(1)**, 930849.

9. Thrasyvoulou, A., Tananaki, C., Goras, G., Karazafiris, E., Dimou, M., Liolios, V., Kanelis, D. and Gounari, S. (2018) Legislation of honey criteria and standards. *Journal of Apicultural Research*, **57(1)**, 88–96.

10. CODEX Alimentarius Commission (1987) Revised CODEX Standard for Honey 12-1981. Revised 1987, 2001. *World Health Organization (WHO)*, **3**.

11. Rashid, M. R., Nor Aripin, K. N., Syed Mohideen, F. B., Baharom, N., Omar, K., Md Taujuddin, N. M. S., Mohd Yusof, H. H. and Addnan, F. H. (2019) The effect of kelulut honey on fasting blood glucose and metabolic parameters in patients with impaired fasting glucose. *Journal of Nutrition and Metabolism*, **2019(1)**, 3176018.

12. Haron, H., Talib, R. A., Subramaniam, P., Arifen, Z. N. Z. and Ibrahim, M. (2022) A comparison of chemical compositions in Kelulut honey from different regions. *Malaysian Journal of Analytical Sciences*, **26**, 447–456.

13. Dan, P. N. S. M., Omar, S. and Ismail, W. I. W. (2018) Physicochemical analysis of several natural Malaysian honeys and adulterated honey. In *IOP Conference Series: Materials Science and Engineering*, **440(1)**, 012049.

14. Yap, S. K., Chin, N. L., Yusof, Y. A. and Chong, K. Y. (2019) Quality characteristics of dehydrated raw Kelulut honey. *International Journal of Food Properties*, **22(1)**, 556–571.

15. Alaerjani, W. M. A. and Mohammed, M. E. A. (2024) Impact of floral and geographical origins on honey quality parameters in Saudi Arabian regions. *Scientific Reports*, **14(1)**, 8720.

16. Malaysian Standard MS 2683 (2017) Stingless Bee Honey Specification.

17. Keskin, M., Keskin, Ş. and Kolaylı, S. (2021) Health-promoting benefits of honey. In *Preparation of Phytopharmaceuticals for the Management of Disorders, Academic Press, London*, 303-306.

18. Burghardt, E., Sewell, D. and Cavanaugh, J. (2022) Agglomerative and divisive hierarchical Bayesian clustering. *Computational Statistics & Data Analysis*, **176**, 107566.

19. Huang, S., Kang, Z., Xu, Z. and Liu, Q. (2021) Robust deep k-means: An effective and simple method for data clustering. *Pattern Recognition*, **117**, 107996.

20. Wang, S., Li, M., Hu, N., Zhu, E., Hu, J., Liu, X. and Yin, J. (2019) K-means clustering with incomplete data. *IEEE Access*, **7**, 69162–69171.

21. Syakur, M. A., Khotimah, B. K., Rochman, E. M. S. and Satoto, B. D. (2018) Integration k-means clustering method and elbow method for identification of the best customer profile cluster. In *IOP Conference Series: Materials Science and Engineering, IOP Publishing*, **336**, 012017.

22. Widia Sembiring, R., Mohamad Zain, J. and Embong, A. (2011) A Comparative Agglomerative Hierarchical Clustering Method to Cluster Implemented Course. *arXiv e-prints*, pp.arXiv, **1101**.

23. Johari, N. A., Ashaari, N. S., Mamat, M. R. and Muhamad, A. (2019) Simple and Rapid Screening Test to Detect Fake Honey Product. *Journal of Agricultural Science and Technology*, **9**, 330–338.

24. Omar, S., Enchang, F. K., Nazri, M. U. I. A., Ismail, M. M. and Ismail, W. I. W. (2019) Physicochemical profiles of honey harvested from four major species of stingless bee (Kelulut) in north east peninsular of Malaysia. *Malaysian Applied Biology*, **48(1)**, 111–116.

25. Moniruzzaman, M., Sulaiman, S. A., Azlan, S. A. M. and Gan, S. H. (2013) Two-year variations of phenolics, flavonoids and antioxidant contents in acacia honey. *Molecules*, **18(12)**, 14694–14710.

26. Lim, A. R., Sam, L. M., Gobilik, J., Ador, K., Choon, J. L. N., Majampan, J. and Benedick, S. (2022) Physicochemical properties of honey from contract beekeepers, street vendors and branded honey in Sabah, Malaysia. *Tropical Life Sciences Research*, **33(3)**, 61.

27. Ng, W. J., Sit, N. W., Ooi, P. A. C., Ee, K. Y. and Lim, T. M. (2021) Botanical origin differentiation of Malaysian stingless bee honey produced by Heterotrigona itama and Geniotrigona thoracica using chemometrics. *Molecules*, **26(24)**, 7628.

28. Maheswari, K. (2019) Finding best possible number of clusters using k-means algorithm. *International Journal of Engineering and Advanced Technology*, **9(1S4)**, 533-538.

29. Ahsan, M. M., Mahmud, M. P., Saha, P. K., Gupta, K. D. and Siddique, Z. (2021) Effect of data scaling methods on machine learning algorithms and model performance. *Technologies*, **9(3)**, 52.

30. Muhati, G. L. and Warui, M. W. (2022) Physicochemical properties and floral sources of honey produced in Marsabit Forest Reserve, Northern Kenya. *Journal of Food Quality*, **2022(1)**, 3841184.

31. Zhang, X., Zhang, S., Qing, X. and Lu, Z. (2019) A new strategy for rapid classification of honeys by simple cluster analysis method based on combination of various physicochemical parameters. *Chemical Research in Chinese Universities*, **35(3)**, 390–394.

32. Brugnerotto, P., Silva, B., Seraglio, S. K. T., Schulz, M., Blainski, E., Dortzbach, D., Gonzaga, L. V., Fett, R. and Costa, A. C. O. (2021) Physicochemical characterization of honeys from Brazilian monitored beehives. *European Food Research and Technology*, **247**, 2709–2719.

33. Sakač, M. B., Jovanov, P. T., Marić, A. Z., Pezo, L. L., Kevrešan, Ž. S., Novaković, A. R. and Nedeljković, N. M. (2019) Physicochemical properties and mineral content of honey samples from Vojvodina (Republic of Serbia). *Food chemistry*, **276**, 15–21.