Applied Mathematics and Computational Intelligence
Volume 13, No. 4, 2024 [72-89]

Exploring Diversity and Abundance of Stingless Bee using Clustering Approach

Nur Maziah Jalilah Jamil[1], Chin Ying Liew[2*], Min Leong Yii[3], Lee Hung Liew[4], Mohd Fahimee Jaapar[5], Jane Labadin[6]

[1,2,3,4] College of Computing, Informatics and Mathematics, Universiti Teknologi MARA (UiTM), Sarawak Branch, 94300 Kota Samarahan, Sarawak, Malaysia
[5] Malaysian Agricultural Research and Development Institute (MARDI), 43400 Serdang, Selangor, Malaysia
[6] Faculty of Computer Science and Information Technology, Universiti Malaysia Sarawak, 94300 Kota Samarahan, Sarawak, Malaysia

[*] Corresponding author : cyliew@uitm.edu.my

Received: 14 October 2024
Revised: 23 October 2024
Accepted: 30 October 2024

ABSTRACT

Stingless bees are paramount in food chain as they are important pollinators of field crops. Recent studies revealed that these bees are seriously threatened by climate change and rapid urbanization across the world. It is thus important to study the relationship between the stingless bee's diversity and the characteristics of the locations they inhibit. At the same time, clustering algorithms is a powerful machine learning approach in exploring unsupervised data. Consequently, this study aims to explore the stingless bee diversity in Malaysia through hierarchical, k-means and DBSCAN clustering. The dataset of this study consists of individual stingless bees collected from 12 locations. It comprises 14 environmental features, 3 physical characteristics, 35 species count, 12 genera counts and 3 diversity-and-abundance weights. A four-stage methodology is employed in the study. The results show that DBSCAN effectively groups data into clusters that are well-defined, but the results are less informative. In contrast, hierarchical and k-means clustering are found producing results that provide clearer insights, with hierarchical clustering delivering notably richer results.

Keywords: DBSCAN, Hierarchical clustering, High dimensional dataset, *k*-means, Meliponine

1 INTRODUCTION

The diversity and abundance of stingless bees serve as an indicator of the health of nature. Tropical, Afrotropical, and Neotropical regions are rich in honey reserves produced by bees [1][2]. The stingless bee, also known as meliponine, is a type of bee that is without stinger. Approximately 600 species of stingless bees have been identified worldwide through the study of their morphology [3] and chemical ecology [4]. Their importance is paramount in the food chain, particularly in crop pollination. Field research has confirmed the importance of stingless bees as pollinators of field crops, particularly in tropical countries [5][6]. The coexistence of stingless bees and other life species relies on the exploitation of standardized resources. Therefore, it is essential to

systematically compile the distribution and abundance of colonies in tropical rainforests in Malaysia and similar climate locations to understand their phylogeny for grouping similarities in their preferences that will ensure their optimal survival. Linking biological diversity, chemical diversity, and environmental factors is crucial for developing an inventory approach to preserving stingless bees.

The status of extinction on a rare stingless bee species compared with a common stingless bee is challenging to establish with a lack of information on the International Union for Conservation of Nature's (IUCN) Red List of Threatened Species, indicating the necessity of further research into the precious creatures. At this juncture, lies the problem diversified in multiple sources and views for data mining in the following studies for all the stingless bees, in areas such as their morphological features [7], foraging distances [8], distribution [9][10] and emerging ecological relationship with other species [11]. The future of stingless bees is at stake with lack of information as the bees are known to fail to return to their colonies [8] when external conditions influence their flying activities by distance and time spent [12][13]. Forests as the natural habitat of stingless bees [14][15] are the resource points where biodiversity assessment can be performed through cluster analysis on satellite images [16]. Thus, the problem identified a direct possible satellite images application on bee diversity where it can be implied that a healthy forest or any location may have greater abundance of stingless bee species. Therefore, the equivalent is using its counterpart, satellite data in place of images. Relating the satellite data that captures the environmental characteristics of a location with the stingless bee species abundance data of the location is the issue that need to be investigated. The study will add value to gaining sufficient information for the study area about stingless bee ecology centred at their habitat of preference. The lack of findings on preferred habitat characteristics for stingless bee is the problem identified, while the corresponding research question is: "What groups can be identified in the locations where stingless bees are sampled, through the environmental properties, physical characteristics and diversity and abundance data at the locations, using clustering approach?"

In this study, three clustering algorithms – hierarchical, *k*-means and Density-Based Spatial Clustering Application with Noise (DBSCAN) – are employed and compared to explore the diversity and abundance data of stingless bee in Malaysia. The organization of this paper is as follows. The first section discusses the introduction of the study, the second section presents the related literature review, the third section puts forth the methodology employed and the limitations of the study, and the fourth section gives the results and provides the corresponding discussion. The last section concludes the study, including the recommendation for future studies.

2 LITERATURE

The literature includes discussion on stingless bees, clustering algorithm, hierarchical clustering, *k*-means clustering and DBSCAN.

2.1 Stingless bees

More than 25 species have been found to play a significant role in pollinating 14 economically significant crops such as coconut *(Cocos nucifera)*, coffee *(Coffea arabica)* and rambutan *(Nephelium lappaceum)* [10][17][18]. The stingless bee species are mostly from genera *Melipona* and *Trigona* [17]. However, the varying characteristics of stingless bees surviving in different environmental

properties pose a challenge for taxonomic analysis or naming for their identification, which directly impacts study on the diversity and abundance of these bees. The importance of proper nomenclature for taxonomy is emphasized in clinical practice about human viruses [19][20]. Key taxonomic characters are those indicative of reproductive isolation or limited gene flow in which the correct matching of mutualistic, antagonistic and symbiotic relationships is a critical information in conservation and management [21]. Taxon-specific responses of stingless bees to habitat changes can be traced from degradation of native indigenous forests [22]. Such study is built on inventory of stingless bees that revealed their spatial distribution across habitat types, including the dispersion of wild and managed colonies [23][24].

Stingless bees are examined through the study of their morphology [3] and chemical ecology [4]. Example morphological traits are body size, hind leg pattern, malar space on their head and wings formation. Available data on identification of stingless bees are limited [25]. The most predominant stingless bee genus found in Peninsular Malaysia are *Geniotrigona spp.*, *Heterotrigona spp*., *Homotrigona spp.*, *Lepidotrigona spp.*, *Tetragonilla spp., Tetragonula spp.*, and *Tetrigona spp* [14]. In Malaysia, research on the diversity and abundance of stingless bees is crucial due to the changes in land use from forested areas to agricultural and developmental purposes [26]. Based on the findings in [26], the overall abundance, evenness and richness of stingless bees decreases from north to south of Peninsular Malaysia, as in Australia. The relation between ecological variation in Peninsular Malaysia and the diversity and abundance of stingless bee population is recommended for future assessment. The findings of the topmost dominant species recorded are: *Heterotrigona itama*, *Geniotrigona thoracica, Tetragonula laeviceps, Lepidotrigona terminata* and *Tetrigona apicalis.*

Stingless bees have been documented to prefer warm and dry climate [15], nesting in hollowed cavities, favoring native tree species [27], underground nests at ridges, slopes, and valleys at virgin jungle reserves [14]. They are also known to be subjected to seasonal availability of mass flowering plants for collection of pollen and nectar on rainy and dry seasons respectively [28]. The degradation of natural habitats in tropical rainforest leads to the reduction or disappearance of stingless bee species [22]. These stingless bees have varying behaviour, ecology, nest architecture, colony size, and worker morphology [29]. Stingless bees exhibit various territoriality and foraging ranges [12][17]. Some stingless bees also become lost during daily foraging and fail to return to their nests prompting studies on homing ability [8]. The state of human disturbance to stingless bee habitats [22] has prompted efforts to assist beekeepers in preserving colonies they rear in meliponine farms. Overall, the realization is that stingless bees are precious workers of the earth, along with other pollinator animals such as birds and bats, which are sensitive to habitat disturbances and climate change [30]. The distribution of features at study areas in stingless bee diversity and abundance in literature [26] led to various representations of the findings, when the composition of species affects ecosystem as found by researchers [31][32].

2.2 Clustering algorithm

Cluster analysis is an unsupervised machine learning method that automatically groups similar data without any intentional outcome or in other words, without any information about the desired output while the system finds structure in the data on its own [33][34]. Hierarchical clustering and partitional clustering are two broad categories of clustering algorithms [35][36]. Both hierarchical clustering and partitional clustering are bound by limitations in clustering data using a global threshold for measuring distances and similarity between two points, two sets and two vectors.

Distance calculation between two points, two sets or two vectors using distance metrics such as Manhattan distance and Euclidean distance or similarity metrics such as Cosine Similarity and Jaccard Similarity are the common vital component in the clustering algorithms, defining the different forms of the global threshold that determine the cluster assignment [36][37][38].

The theory of measurement in clustering refers to the mathematical framework used to quantify the similarity or dissimilarity between objects. The fundamental principle of classification involves the identification of objects that exhibit similarities within a pattern space, based on a quantifiable measure of distance or dissimilarity. Consequently, objects characterized by a minimal distance or dissimilarity are deemed to belong to the same cluster. A commonly used measure of distance or dissimilarity is Euclidean distance in Equation 1 [39]. The Euclidean distance denoted by d_{ij}, where n represents number of vectors, x_{ik} represents input image vector and x_{jk} represents comparison image vector, and i, j, k are indices denoting the dimensions of the vectors. It is the square root of the sum of differences of the two vectors, measuring nearest distance, which is typically applied in image processing [39]. An alternative distance measurement is rectilinear or Manhattan distance which is the sum of absolute differences between points across all the dimensions. The distance formula is given in Equation 2, where $d_{man(x,y)}$ denotes Manhattan distance, x and y represent two vectors being compared, d represents number of dimensions in the vector, i, j and k are indices representing dimension of vectors [39].

$$d_{ij} = \sqrt{\sum_{k=1}^{n} \left(x_{ik} - x_{jk}\right)^2} \quad (1)$$

$$d_{man^{(x,y)}} = \sum_{k=1}^{d} \left|x_j - y_j\right| \quad (2)$$

2.3 Hierarchical clustering

Hierarchical clustering methods begin by treating each data point as a single cluster, and then iteratively merge or split to form a clustering structure [40][41]. Hierarchical clustering can be divided into two sub-categories, agglomerative hierarchical and divisive hierarchical [34], where the groups are assigned points by their distance using a linkage metric in the creating of large datasets whether agglomeratively or divisively [38]. The three common linkage measures are single linkage, complete linkage and average linkage [42]. Additional linkages are weighted linkage and Ward linkage. The single linkage computes the distance between the closest elements of the two clusters. Complete linkage computes the distance between the clusters' most distant elements and average linkage computes the average distance between elements of the two clusters. Weighted linkage uses the Weighted Pair Group Method with Arithmetic Mean (WPGMA). Ward linkage computes the increase of the error sum of squares after merging two clusters into a single cluster. The hierarchical clustering algorithm relies on an N x N connectivity matrix, which is built by calculating the similarity between each pair of data points. This similarity matrix is then used to determine the distance between clusters, with the linkage criterion calculated based on pairwise distances between clusters. The similarity metric plays a crucial role in shaping the clusters and measuring the distance between them [43].

2.4 *K*-means clustering

K-means clustering displays the data arranged into a nested sequence of groups without any hierarchical structure but with partitional clustering algorithm [44][45]. Specifically, *k*-means clustering is a centroid-based technique within the partitional clustering algorithm, where data objects are grouped into *k* clusters, clustering using the center point of the cluster. Euclidean distance and Manhattan distance are commonly used in the *k*-means algorithm. *K*-means clustering is a widely employed unsupervised technique for partitioning data into distinct groups. The algorithm requires the specification of the number of clusters, denoted by the parameter *k*. For instance, when $k = 2$, the data is divided into two clusters. The iterative process continues until the centroids of the newly formed clusters converge or a predetermined maximum number of iterations is reached. The *k*-means algorithm aims to produce clusters of the same sizes, thereby facilitating the identification of patterns in the data [45]. Determining optimum number of clusters is through using Silhouette Method. The Silhouette score, ranging from -1 to $+1$, provides a measure of how well an object fits within its assigned cluster. Scores approaching $+1$ indicate that an object shares similar characteristics with its cluster, while scores approaching -1 suggest that the object possesses distinct characteristics that differentiate it from its cluster. A clustering configuration is considered suitable when most objects exhibit high silhouette scores, indicating a strong similarity between objects within clusters. Conversely, a clustering configuration is deemed unsuitable when many objects have low or negative Silhouette scores, indicating a poor fit between objects and their assigned clusters.

2.5 DBSCAN clustering

Density-Based Spatial Clustering Application with Noise (DBSCAN) is a density-based clustering algorithm that systematically finds all core points and expands each to all density-reachable points to form clusters [46][47]. It identifies objects with numerous nearby neighbours as clusters, while objects isolated in sparse regions are labelled as outliers or noise. These outliers are characterized by having nearest neighbours that are too distant from them [47]. The algorithm can find any type of shape in the data. Prior to using the DBSCAN algorithm, the user needs to set two important settings that are fixed. The distance between points (or neighbourhood distance) and the minimum number of points (or threshold) needed to form a group. The distance setting determines the radius of neighbourhood range of the same group. The minimum number of points, or threshold setting, determines the number of data points required to form a cluster. The threshold is established by adjusting the minimum number of sample points within the neighbourhood radius needed to identify a point as a core point. The DBSCAN algorithm is good at ignoring noisy data and can handle datasets with errors. It can find and remove noisy data points from the results. The algorithm looks at how the data is spread out and uses this information to identify different groups and mark the clustering results.

3 MATERIAL AND METHODS

The study aims to identify clusters based on the diversity and abundance of stingless bees in the environment characteristic of the habitat. Discussion on each stage of the methodology is given in the following subsections.

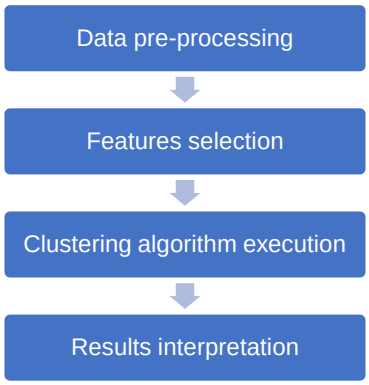

Figure 1 : Study methodology flowchart

3.1 Data pre-processing

The data used in this study are the diversity and abundance of stingless bees obtained from the Malaysian Agricultural Research and Development Institute (MARDI) [26]. The diversity and abundance data consists of the abundance data for each genus of stingless bees identified and the abundance of each species recognized. It is collected from June 2013 to June 2015. The 12 study areas where the samples are collected are given in Table 1. There are 35 species, and 12 genera identified, with 1,599 individuals sampled throughout the 12 locations in Malaysia. For each location, the environmental data and geographical characteristics are needed. In this stage, data pre-processing includes gathering the environmental and geographical characteristics data that are available, labelling the data, and normalization of the data. The study areas are referred to as the location (L) and labelled as L1 to L12 in this study.

Table 1 : Study Area

Study area	Label
Perlis State Park, Perlis	L1
Titi Hayun Recreational Forest, Kedah	L2
Lata Kekabu Recreational Forest, Perak	L3
Kampung Tengah Teluk Bahang, Penang	L4
Gunung Nuang Recreational Forest, Selangor	L5
Kampung Pagi National Park, Pahang	L6
Pulau Tekak, Tasik Kenyir, Terengganu	L7
Jelawang, Kelantan	L8
Sungai Udang Recreational Forest, Melaka	L9
Gunung Datuk Recreational Forest, Negeri Sembilan	L10

Gunung Ledang Recreational Forest, Johor	L11
Bukit Nanas Recreational Forest, Federal Territory	L12

The environmental data, including the satellite data, is retrieved based on the Global Positioning System (GPS) of each location, from the POWER Project's Daily v2.2.16 version obtained from the National Aeronautics and Space Administration (NASA) Langley Research Center (LaRC) Prediction of Worldwide Energy Resource (POWER) Project funded through the NASA Earth Science/Applied Science Program. The data retrieved targets on the period from 9 to 11 AM on the data collection date by MARDI at all 12 locations. The environmental and satellite data retrieved comprised 14 variables. They are the minimum and difference of temperature at 2 meters (T2M MIN and T2M delta), dew point temperature at 2 meters (T2MDEW MIN and T2MDEW Delta), relative humidity at 2 meters (RH2M MIN and RH2M Delta), specific humidity at 2 meters (QV2M MIN and QV2M Delta), ultraviolet A radiation (UVA MIN and UVA Delta), ultraviolet B radiation (UVB MIN and UVB Delta) and light intensity (LX MIN and LX Delta). Elevation at the location and the latitude and longitude of the locations are the geographical characteristic variables used in this study and are extracted from Google Earth Pro 7.3.6.9796.

The number of individuals for each species is extracted for each location. Since there are 35 species identified, the number of variables regarding species (S) for each location totalled 35. The variables are labelled as S1 to S35. Similarly, the number of individuals identified under each genus (M) are determined for each location. The 12 genera reported resulting in 12 variables regarding genus for each location. The variables are labelled as M1 to M12. On top of this, the abundance weightages of each species and genus each location are incorporated in this study too. The weightage of genera abundance, M_gdv_{Li}, sampled at a location i is expressed in Equation 3. This is defined as the sum of the ratios between the count of species occurrences in a genus at a specific location and the total count of species occurrences in the same genus across all 12 locations. Equation 4 shows an example of the genera abundance calculation for L1, M_gdv_{L1}. The value of $M_{1\_gdv\_L1} = \frac{1}{20}$ implies that there is 1 species occurrence in the genus *Heterotrigona* (M1) at L1, out of a total count of 20 species occurrences of the same genus across all 12 locations.

$$M\_gdv_{Li} = \sum_{n=1}^{12} M_{n\_gdv\_Li} \quad (3)$$

$$\text{where } M_{n\_gdv\_Li} = \frac{\text{count of species occurences within a genus at a specific location}}{\text{total count of species occurrences within the same genus across all 12 locations}}$$

$$M\_gdv_{L1} = \frac{1}{20} + \frac{1}{12} + \frac{3}{30} + \frac{0}{3} + \frac{1}{3} + \frac{2}{23} + \frac{0}{5} + \frac{3}{36} + \frac{8}{91} + \frac{0}{7} + \frac{0}{14} = 0.8249 \quad (4)$$

The weightage of individual bee abundance in a genus, M_cpg_{Li}, found at a location i is given in Equation 5. The calculation for genus M1 at L1 returns a value of 0.0593, which is given by $M\_cpg_{L1} = \frac{20}{337}$ where there are 20 individual bees in genus M1 sampled at L1 and a total of 337 individual bees in genus M1 sampled at all 12 locations. Similarly, the weightage of individual bees in a species found at a location, S_cpg_{Li}, is given in Equation 6. The calculation for species S1 at L1

returns a value of 0.0647, which is given by $S\_cpg_{L1} = \frac{20}{309}$ where there are 20 individual bees in species S1 sampled at L1 and a total of 309 individual bees in species S1 sampled at all 12 locations.

$$M\_cpg_{Li} = \frac{\text{number of individual bees of a genus at a specific location}}{\text{total number of individual bees across all genera at all 12 locations}} \quad (5)$$

$$S\_cpg_{Li} = \frac{\text{number of individual bees of a species at a specific location}}{\text{total number of individual bees of the same species across all 12 locations}} \quad (6)$$

Consequently, there are 67 variables collected for each of the 12 locations. They are made up of 14 environment variables, 3 geographical characteristic variables, 35 species count variables, 12 genera count variables, and 3 ratio variables. The values of each variable are normalized so that it ranges from 0 to 1 so that all features contribute equally to the distance calculations when executing the clustering algorithm.

3.2 Features selection

The number of features a dataset has is commonly referred to as the dimension of the dataset. However, the dimensionality of a dataset depends on the ratio of the number of features, *f*, and the size, *N*, of the dataset. A high dimensional dataset is one where *f* is larger than *N*, denoted as $f >> N$ [48][49]. For a high dimensional data that carries a vast amount of information to store and analyze, dimensionality reduction is a crucial step in pre-processing of data that aims to shrink the dataset while preserving its accuracy [50]. Since there are 67 features in the dataset with 12 data points in this study, it is a small and high-dimensional dataset. Subsequently, features selection or to be specific, features extraction or data reduction needs to be carried out. Principal component analysis (PCA) is employed in this study to reduce the dimensionality of the dataset. It transforms a set of correlated variables into uncorrelated variables termed as principal components (PC). It is suitable to be used as a feature extraction technique, supporting the features selection.

PCA is selected as an aggregator enabling inclusion of all elements of the descriptor matrices by [51] where they are combining it with hierarchical clustering approaches. Past studies show that several PC that accounted for a percentage ranging between more than 80% [52] to at least 90% of the total variance [53] are selected. Thus, this study resolves to set a threshold value of more than 90% of explained variance in the PCA conducted. Figure 2 shows that the PCA conducted suggests 10 principal components, which contribute to a total of 94.3% explained variance, to be included for clustering algorithm execution in the next stage.

3.3 Clustering algorithm execution

In this step, hierarchical clustering, *k*-means clustering and DBSCAN are employed using the data that have been pre-processed. The optimal principal components are input to the respective clustering algorithm. This study uses Orange software (version 3.37.00), a powerful and widely used free software for data mining [54], for the execution of all the clustering algorithms. For hierarchical clustering, the agglomerative approach with single linkage is used as it is relatively more suitable for small dataset. The agglomerative approach with Ward linkage as it is recommended for data with high dimensionality, which may cause overlapping clustering [42]. Both hierarchical clustering approaches use Euclidean distance metric. For *k*-means, standard

Euclidean distance metric is used, with suitable value of *k*. Based on the studies on the preference of stingless bees for their habitat and the observation of the geographical features of the locations of this study and their surrounding land area, the value of *k* is pre-set as 3. For DBSCAN, Manhattan distance metric is used as it is relatively more suitable for dataset with high dimensionality [55]. Two other standard parameters for DBSCAN, the neighbourhood distance or more commonly referred to as epsilon (ε) and minimum number of points are tuned to obtain an optimum result. In this study, the distance value of 186.926 is employed based on the 'elbow' identified in the *k*-distance plot of DBSCAN as presented in Figure 3. Due to a small dataset, the minimum number of points are tuned to 3 in this study.

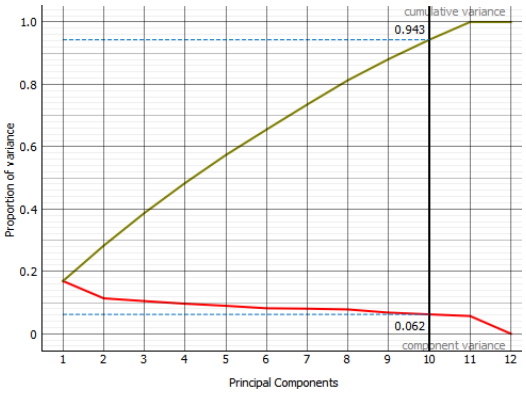

Figure 2 : Scree plot of PCA

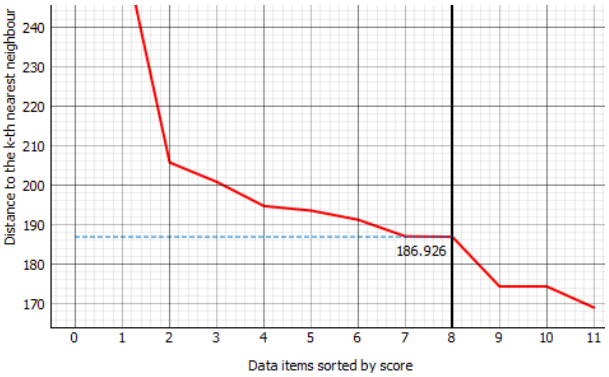

Figure 3 : *k*-distance plot of DBSCAN

3.4 Results interpretation

The interpretation of results refers to comparing the results with the data available and literature review. Visual representations such as dendrograms and scatter plots of the clustering results are generated to compare.

4 RESULTS AND DISCUSSION

In this section, the results obtained from the methodology stages are presented and discussed. The interpretation of the results is subject to the data used. The dataset used in this study is of small size and high dimensionality. When hierarchical clustering with single linkage is employed, the locations are grouped into 1 cluster (L11, L5, L1, L4, L6, L7, L8, L9, L10, L12) only, with two ungrouped locations (L3 and L2) at 89.0% similarity as presented in Figure 4. The clustering of the locations from the hierarchical clustering with Ward linkage resulted in 4 clusters at 73.6% similarity as shown by the dendrogram in Figure 5. The first cluster consists of L2 and L3. The second cluster consists of 5 locations in two sub-clusters: L12, L9 and L10, and L6, L7 and L8. The third cluster has two members, L1 and L4 while the fourth cluster groups two locations together, L5 and L11. As for the clusters produced from *k*-means clustering algorithm, the grouping of the locations is displayed in Figure 7. Five locations (L5, L6, L7 and L8) are grouped in one cluster. L1, L4, L9, L10, L11 and L12 are group in another cluster. One more cluster comprised L2 and L3. Clustering using DBSCAN produced two clusters in Figure 8. One cluster grouped L2, L3, L5, L6, L7 and L8 together. Another cluster grouped the rest of six locations (L1, L4, L9, L10, L11 and L12) together. The Euclidean distance map for hierarchical clustering is shown in Figure 6. It agrees with

the clustering results shown in Figure 4 and Figure 5. It reveals the grouping of the locations into 4 clusters (as shown by the locations grouped with red colour rectangles) that agrees with Figure 5. The locations grouped with blue colour rectangle, together with locations L2 and L3, agree with the clustering shown in Figure 4.

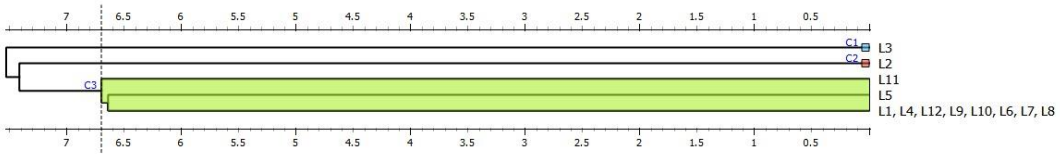

Figure 4 : Hierarchical clustering of locations with single linkage at 89.0% similarity

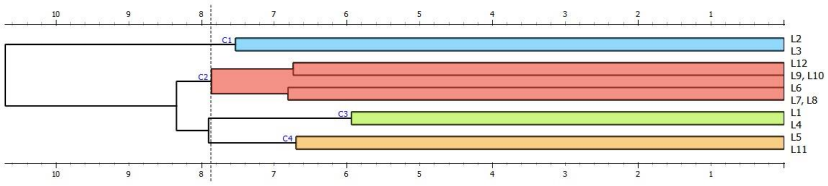

Figure 5 : Hierarchical clustering of locations with Ward linkage at 73.6% similarity

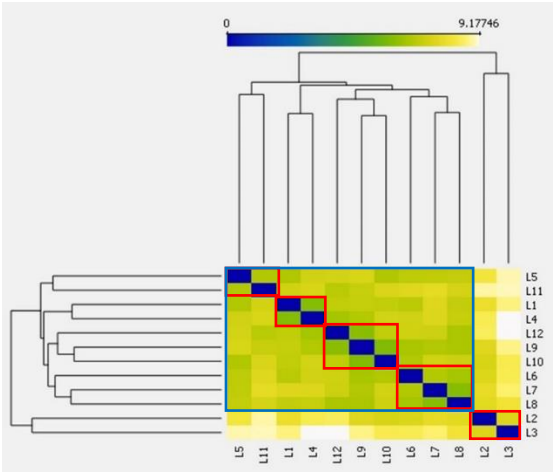

Figure 6 : Distance map measured in Euclidean distance

The two ungrouped locations under hierarchical clustering with single linkage are L2 and L3, but they are grouped as a cluster by hierarchical clustering with Ward linkage and also *k*-means clustering. Though not grouped in a distinct cluster by DBSCAN, L2 and L3 are together in a clustering with other locations. This implies L2 and L3 share certain similar characteristics. This is supported by the past studies and the data collected. Table 2(a) shows that L3 and L2 are the two

top-ranked location in terms of the ecological indices of stingless bees [26], signifying the two locations as the most abundance in terms of the stingless bee's diversity in this study. Based on the data summarized in Table 2(b), L3 and L2 are also the two locations recorded with the highest number of individual bees sampled. Out of 35 species identified [26], L3 and L2 are recorded with 34 and 33 species, respectively as revealed in Table 2(c). They are the two top ranked locations in terms of number of species found, supporting the fact that stingless bee's species at L3 and L2 are the most diverse, and they are suitable for the stingless bees. When the land area and its location geographical features are further investigated, there are waterfall at L2 and L3 and they belong to lowland and hill type of forest, respectively.

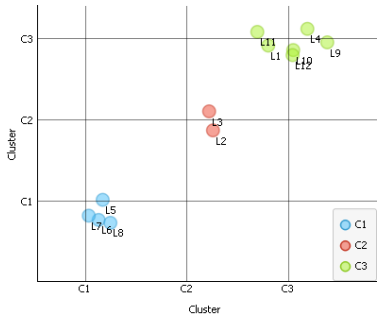

Figure 7 : *K*-means cluster

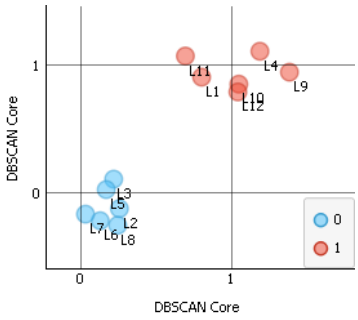

Figure 8 : DBSCAN cluster

On the other hand, locations L6, L7 and L8 are grouped together as a branch or sub-cluster of cluster two, as shown in Figure 5, whereas they are grouped with all the rest of locations shown in Figure 4. These three locations are closely grouped together in terms of the number of individual bees sampled as presented in Table 2(b). They are reported with 21, 22 and 23 types of species, respectively. With *k*-means clustering algorithm, L6, L7 and L8 are grouped together with L5, implying they are in the states that are geographically neighbouring to one another. With DBSCAN, L6, L7 and L8 are grouped together with L2, L3 and L5, in Figure 8. These six locations are the six top ranked locations according to the Shannon-Weiner index in the decreasing order of L3, L2, L7, L5, L6 and L8, as revealed by Table 2(a).

From Figure 5, L9, L10 and L12 are grouped under another branch or sub-cluster of cluster two while locations L1 and L4 are grouped in a different cluster. L4 and L12 are in area that are highly developed with heavy urbanization as presented in Table 4 [26]. Interestingly, these five locations are ranked at the bottom in terms of the Shannon-Weiner index as presented in Table 2(a), with decreasing order of L10, L1, L9, L4 and L12. From Figure 7, L9, L10 and L12 are grouped with L1, L4 and L11 within a cluster where all these six locations are ranked as the bottom six locations in terms of the Shannon-Weiner index. Except L9, the rest of the locations (L1, L4, L10, L11 and L12) are ranked the bottom five locations in terms of the number of individual bees sampled. They are also at the bottom in the descending ranking in terms of the number of species identified at the locations. According to Table 3(c), the habitat disturbance of L1, L4, L10, L11 and L12 is categorized as fragmented type. Figure 8 reveals the same results as Figure 7, where these six locations (L1, L4, L9, L10, L11 and L12) are grouped under the same cluster by DBSCAN. All these six locations are ranked at the bottom six locations in terms of the Shannon-Weiner index.

Surprisingly, the cluster with L5 and L11 as members in Figure 5 are different with the results shown in Figure 7 from *k*-means clustering and Figure 8 from DBSCAN clustering, where they are

grouped in different clusters. The result is not in line with the ranking of diversity and abundance based on the Shannon-Weiner index. From Table 2(a), L5 is ranked fourth but L11 is ranked tenth. The same goes to the ranking based on the number of the individual bees and number of species identified at the locations. Nonetheless, both L5 and L11 are grouped together as shown in Table 4. Both are commented as located in the two most industrialized states where habitat disturbance is obvious [26]. Urbanization has been reported as a threat towards the sustainability of stingless bees' existence [22]. Urbanization also contributes negatively to climate change, which affect the wellbeing of stingless bees [30]. Besides that, both L5 and L11 share the same forest type (hill type), geographical feature (mountain), and habitat disturbance type (fragmented) as revealed in Table 3(a), (b) and (c).

Table 2 : Ranking of locations based on (a) ecological indices of stingless bees [26, p.3]; (b) number of individual stingless bees sampled; (c) number of stingless bees' species identified

(a)		(b)		(c)	
Location	Shannon-Weiner index, H'	Location	Number of individual bees	Location	Number of species
L3	2.64	L3	209	L3	34
L2	2.56	L2	173	L2	33
L7	2.46	L9	163	L8	23
L5	2.37	L6	160	L7	22
L6	2.34	L7	156	L6	21
L8	2.30	L8	139	L10	21
L10	2.24	L5	127	L5	20
L1	2.10	L10	127	L1	19
L9	2.01	L1	114	L9	15
L11	1.75	L4	101	L4	14
L4	1.74	L11	89	L11	13
L12	1.24	L12	41	L12	9

Table 3 : Grouping of locations according to (a) type of forest; (b) geographical feature; (c) habitat disturbance type

(a)		(b)		(c)	
Location	Forest type	Location	Geographical feature	Location	Habitat disturbance type
L1	Lowland hill semi-deciduous	L1, L12	Hill	L1, L2 L3, L5, L6 L10, L11 L12	Fragmented
L2, L4 L6, L7, L9	Lowland	L2, L3 L4, L8	Waterfall	L4, L8, L9	Logged
L3, L5, L8, L10, L11, L12	Hill	L5, L10 L11	Mountain	L7	Agriculture
		L6, L7, L9	River		

Table 4 : Grouping of locations according to habitat type [26]

Location	Habitat type
L5, L11	They are in the two most industrialized states, which experience significant habitat disturbance.
L6, L10	They are in two states that are neighbouring to each other and share similar geographical characteristics to a certain extent.
L3, L9	The diversity rate of L3 is higher than that of L9. However, the total number of species identified between the two locations does not show much differences.
L1, L2, L7, L8	They are locations with not more than 20% habitat disturbance. They are in a land area that is green, with a relatively lower degree of urbanization, suggesting it suitable for the stingless bees to build their nest. They are location in four states that are situated along three major mountain ranges, which are extremely rich in biodiversity.
L4	It is located at highly developed land area, where habitat disturbance affects the viability of stingless bees.
L12	It is located at highly developed land area, where habitat disturbance affects the viability of stingless bees.

5 CONCLUSION

This study intends to explore using clustering approaches on the groups that can be identified at the locations where stingless bees are sampled, through the environmental properties, physical characteristics, and diversity and abundance data of the locations. Three clustering algorithms are used, hierarchical clustering, *k*-means clustering, and DBSCAN. DBSCAN has managed to group the locations into clusters that are tally with the diversity and abundance in terms of the Shannon-Weiner index. Nevertheless, with two clusters produced, the results given by DBSCAN are less informative. Both hierarchical clustering and *k*-means clustering are relatively more useful as the results produced are richer. Hierarchical clustering particularly stands out as the results agrees with most of the past studies and the observations of the physical characteristics of the locations. Subsequently, this indicates that the variables incorporated in the study are well-supported. It is recommended that a small-size dataset that is coupled with high-dimensionality may consider particularly hierarchical clustering with Ward linkage, followed by *k*-means clustering algorithms. Recommendation for future studies includes considering the emergent clustering algorithms that are targeted for small-size dataset and innovative methodology in analysing study with high-dimensional dataset, particularly in considering another machine learning approach or a combination of more than one machine learning approach.

ACKNOWLEDGEMENT

This study was funded by Ministry of Higher Education (MOHe) through Fundamental Research Grant Scheme (FRGS) code (FRGS/1/2023/ICT04/UITM/02/1).

REFERENCES

[1] A. A. Kidane, F. M. Tegegne, and A. J. M. Tack, "Indigenous knowledge of ground-nesting stingless bees in southwestern Ethiopia," *Int J Trop Insect Sci*, vol. 41, no. 4, pp. 2617–2626, 2021, doi: 10.1007/s42690-021-00442-6.

[2] B. Tesfaye Dubale and T. Gelgelu Desha, "Assessment of stingless bee (Apidae: Meliponini) & production practices and indigenous knowledge in West Arsi and Bale Zones of South-Eastern Oromia, Ethiopia," *World Journal of Agricultural Science and Technology*, 2023, doi: 10.11648/j.wjast.20230101.11.

[3] M. S. Engel, C. Rasmussen, R. Ayala, and F. F. de Oliveira, "Stingless bee classification and biology (Hymenoptera, Apidae): a review, with an updated key to genera and subgenera," *Zookeys*, vol. 1172, pp. 239–312, 27AD. doi: https://doi.org/10.3897/zookeys.1172.104944

[4] S. D. Leonhardt, "Chemical ecology of stingless bees," *J Chem Ecol*, vol. 43, no. 4, pp. 385–402, 2017, doi: 10.1007/s10886-017-0837-9.

[5] W. A. Azmi *et al.*, "Effects of pollination by the Indo-Malaya stingless bee (Hymenoptera: Apidae) on the quality of greenhouse-produced rockmelon," *J Econ Entomol*, vol. 112, no. 1, pp. 20–24, Feb. 2019, doi: 10.1093/jee/toy290.

[6] S. A. M. Khalifa *et al.*, "Overview of bee pollination and its economic value for crop production," *Insects*, vol. 12, no. 8, 688, Jul. 2021, doi: 10.3390/insects12080688.

[7] J. Yoo, M. Z. Hossain, and K. A. Ahmed, "A machine learning based approach to study morphological features of bees," in *Proceedings of The 1st International Electronic Conference on Entomology*, Basel, Switzerland: MDPI, Jul. 2021, 10607. doi: 10.3390/IECE-10607.

[8] W. Kasiera, S. Kariuki, M. Musonye, K. Krausa, and N. Kiatoko, "Influence of landscape on foraging range and homing ability of afrotropical stingless bees," *Insectes Soc*, vol. 70, no. 1, pp. 59–67, 2023, doi: 10.1007/s00040-023-00899-3.

[9] C. L. Yurrita, M. A. Ortega-Huerta, and R. Ayala, "Distributional analysis of melipona stingless bees (Apidae: Meliponini) in Central America and Mexico: setting baseline information for their conservation," *Apidologie*, vol. 48, no. 2, pp. 247–258, 2017, doi: 10.1007/s13592-016-0469-z.

[10] K. Wayo, T. Sritongchuay, B. Chuttong, K. Attasopa, and S. Bumrungsri, "Local and landscape compositions influence stingless bee communities and pollination networks in tropical mixed fruit orchards, Thailand," *Diversity (Basel)*, vol. 12, no. 12, 482, 2020, doi: 10.3390/d12120482.

[11] G. T. de Paula *et al.*, "Further evidences of an emerging stingless bee-yeast symbiosis," *Front Microbiol*, vol. 14, Aug. 2023, doi: 10.3389/fmicb.2023.1221724.

[12] S. Benedick, J. A. Gansau, and A. H. Ahmad, "Foraging behaviour of *Heterotrigona itama* (Apidae: Meliponini) in residential areas," *Pertanika J Trop Agric Sci*, vol. 44, no. 2, 2021, doi: 10.47836/pjtas.44.2.13.

[13] N. Thonhual, "Machine learning to examine the foraging periods of bees," in *2023 18th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP)*, IEEE, Nov. 2023, pp. 1–6. doi: 10.1109/iSAI-NLP60301.2023.10354611.

[14] H. Salim, A. Dzulkiply, R. Harrison, C. Fletcher, A. R. Kassim, and M. Potts, "Stingless bee (hymenoptera: Apidae: Meliponini) diversity in dipterocarp forest reserves in Peninsular Malaysia," *Raffles Bull Zool*, vol. 60, no. 1, pp. 213–219, 2012.

[15] A. H. Jalil, *Beescape for Meliponines: Conservation of Indo-Malayan stingless bees*, 1st ed., vol. 1. Partridge Publishing Singapore, 2014.

[16] J. Talaga and P. Netzel, "Assessment of forest biodiversity in Poland using unsupervised learning and satellite data," in *IGARSS 2024 - 2024 IEEE International Geoscience and Remote Sensing Symposium*, IEEE, Jul. 2024, pp. 4294–4298. doi: 10.1109/IGARSS53475.2024.10642676.

[17] V. Meléndez Ramírez, R. Ayala, and H. Delfín González, "Crop pollination by stingless bees," in *Pot-Pollen in Stingless Bee Melittology*, Cham: Springer International Publishing, 2018, pp. 139–153. doi: 10.1007/978-3-319-61839-5_11.

[18] C. Vaidya, G. Fitch, G. H. D. Martinez, A. M. Oana, and J. Vandermeer, "Management practices and seasonality affect stingless bee colony growth, foraging activity, and pollen diet in coffee agroecosystems," *Agric Ecosyst Environ*, vol. 353, 108552, Sep. 2023, doi: 10.1016/j.agee.2023.108552.

[19] B. C. Bennett and M. J. Balick, "Does the name really matter? The importance of botanical nomenclature and plant taxonomy in biomedical research," *J Ethnopharmacol*, vol. 152, no. 3, pp. 387–392, Mar. 2014, doi: 10.1016/j.jep.2013.11.042.

[20] B. Barratt, M. Cock, and R. Oberprieler, "Weevils as targets for biological control, and the importance of taxonomy and phylogeny for efficacy and biosafety," *Diversity (Basel)*, vol. 10, no. 3, 73, Jul. 2018, doi: 10.3390/d10030073.

[21] S. M. Jackson *et al.*, "The importance of appropriate taxonomy in Australian mammalogy," *Aust Mammal*, vol. 45, no. 1, pp. 13–23, Oct. 2022, doi: 10.1071/AM22016.

[22] N. Kiatoko, S. K. Raina, and F. van Langevelde, "Impact of habitat degradation on species diversity and nest abundance of five African stingless bee species in a tropical rainforest of Kenya," *Int J Trop Insect Sci*, vol. 37, no. 03, pp. 189–197, Sep. 2017, doi: 10.1017/S174275841700011X.

[23] C. E. Ruano Iraheta, M. Á. Hernández Martínez, L. A. Alas Romero, M. E. Claros Álvarez, D. R. Arévalo, and V. A. Rodríguez González, "Stingless bee distribution and richness in El Salvador," *J Apic Res*, vol. 54, no. 1, pp. 1–10, Jan. 2015, doi: 10.1080/00218839.2015.1029783.

[24] A. Assefa and M. Lemma, "Ecological niche modeling for stingless bees (genus Melipona) in Waghemira and North Wollo zones of Amhara Regional State, Ethiopia," *Sci Afr*, vol. 15, e01102, 2022, doi: 10.1016/j.sciaf.2022.e01102.

[25] A. H. Abdul Jalil and I. Shuib, "Pictorial key to Indo-Malayan stingless bee genera By Abu Hassan Jalil and Ibrahim Shuib," 2014, *Cairns*.

[26] M. F. Jaapar *et al.*, "The diversity and abundance of stingless bee (Hymenoptera: Meliponini) in Peninsular Malaysia," *American-Eurasian Network for Scientific Information*, vol. 10, no. 9, pp. 1–7, 2016.

[27] C. Maia-Silva, M. Hrncir, C. I. da Silva, and V. L. Imperatriz-Fonseca, "Survival strategies of stingless bees (Melipona subnitida) in an unpredictable environment, the Brazilian tropical dry forest," *Apidologie*, vol. 46, no. 5, pp. 631–643, 2015, doi: 10.1007/s13592-015-0354-1.

[28] K. P. Aleixo, C. Menezes, V. L. Imperatriz Fonseca, and C. I. da Silva, "Seasonal availability of floral resources and ambient temperature shape stingless bee foraging behavior (Scaptotrigona aff. depilis)," *Apidologie*, vol. 48, no. 1, pp. 117–127, 2017, doi: 10.1007/s13592-016-0456-4.

[29] F. H. I. D. Segers, C. Grüter, C. Menezes, S. Mateus, and F. L. W. Ratnieks, "Correlated expression of phenotypic and extended phenotypic traits across stingless bee species: worker eye morphology, foraging behaviour, and nest entrance architecture," *J Apic Res*, vol. 61, no. 5, pp. 598–608, Oct. 2022, doi: 10.1080/00218839.2022.2114711.

[30] E. Rahimi and C. Jung, "Global trends in climate suitability of bees: Ups and downs in a warming world," *Insects*, vol. 15, no. 2, 2024, doi:10.3390/insects15020127.

[31] K. S. Prendergast, "The influence of plant species, origin and color of garden nursery flowers on the number and composition of pollinating insect visitors," *J Agric Urban Entomol*, vol. 38, no. 1, 2022, doi: 10.3954/JAUE22-07.

[32] D. Liu, P. S. Chang, S. A. Power, J. N. B. Bell, and P. Manning, "Changes in plant species abundance alter the multifunctionality and functional space of heathland ecosystems," *New Phytologist*, vol. 232, no. 3, pp. 1238–1249, Nov. 2021, doi: 10.1111/nph.17667.

[33] I. H. Sarker, "Machine learning: Algorithms, real-world applications and research directions," *SN Comput Sci*, vol. 2, no. 3, 160, May 2021, doi: 10.1007/s42979-021-00592-x.

[34] X. Ran, Y. Xi, Y. Lu, X. Wang, and Z. Lu, "Comprehensive survey on hierarchical clustering algorithms and the recent developments," *Artif Intell Rev*, vol. 56, no. 8, pp. 8219–8264, Aug. 2023, doi: 10.1007/s10462-022-10366-3.

[35] I. A. Venkatkumar and S. J. K. Shardaben, "Comparative study of data mining clustering algorithms," in *2016 International Conference on Data Science and Engineering (ICDSE)*, IEEE, Aug. 2016, pp. 1–7. doi: 10.1109/ICDSE.2016.7823946.

[36] P. Mulinka, P. Casas, K. Fukuda, and L. Kencl, "HUMAN - Hierarchical clustering for unsupervised anomaly detection & interpretation," in *2020 11th International Conference on Network of the Future (NoF)*, IEEE, Oct. 2020, pp. 132–140. doi: 10.1109/NoF50125.2020.9249194.

[37] S. Pandey and P. Khanna, "A hierarchical clustering approach for image datasets," in *2014 9th International Conference on Industrial and Information Systems (ICIIS)*, IEEE, Dec. 2014, pp. 1–6. doi: 10.1109/ICIINFS.2014.7036504.

[38] J. Irani, N. Pise, and M. Phatak, "Clustering techniques and the similarity measures used in clustering: A survey," *Int J Comput Appl*, vol. 134, no. 7, pp. 9–14, 2016, doi: 10.5120/IJCA2016907841.

[39] R. Suwanda, Z. Syahputra, and E. M. Zamzami, "Analysis of Euclidean distance and Manhattan distance in the *k*-means algorithm for variations number of centroid *k*," *J Phys Conf Ser*, vol. 1566, no. 1, 012058, Jun. 2020, doi: 10.1088/1742-6596/1566/1/012058.

[40] V. Cohen-addad, V. Kanade, F. Mallmann-trenn, and C. Mathieu, "Hierarchical clustering," *Journal of the ACM*, vol. 66, no. 4, pp. 1–42, Aug. 2019, doi: 10.1145/3321386.

[41] P. Shetty and S. Singh, "Hierarchical clustering: A survey," *International Journal of Applied Research*, vol. 7, no. 4, pp. 178–181, Apr. 2021, doi: 10.22271/allresearch.2021.v7.i4c.8484.

[42] Vijaya, S. Sharma, and N. Batra, "Comparative study of single linkage, complete linkage, and Ward method of agglomerative clustering," in *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, IEEE, Feb. 2019, pp. 568–573. doi: 10.1109/COMITCon.2019.8862232.

[43] A. E. Ezugwu *et al.*, "A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects," *Eng Appl Artif Intell*, vol. 110, 104743, Apr. 2022, doi: 10.1016/j.engappai.2022.104743.

[44] Aastha Gupta, Himanshu Sharma, and Anas Akhtar, "A comparative analysis of *k*-means and hierarchical clustering," *EPRA International Journal of Multidisciplinary Research (IJMR)*, pp. 412–418, Sep. 2021, doi: 10.36713/epra8308.

[45] P. B. Maruthi and P. Bilas, "Comparative analysis of *k*-means and hierarchical clustering in bigdata environment," in *2022 6th International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS)*, 2022, pp. 1–6. doi: 10.1109/CSITSS57437.2022.10026370.

[46] M. Hahsler, M. Piekenbrock, and D. Doran, "DBSCAN : Fast density-based clustering with *R*," *J Stat Softw*, vol. 91, no. 1, 2019, doi: 10.18637/jss.v091.i01.

[47] D. Deng, "DBSCAN clustering algorithm based on density," in *2020 7th International Forum on Electrical Engineering and Automation (IFEEA)*, IEEE, Sep. 2020, pp. 949–953. doi: 10.1109/IFEEA51475.2020.00199.

[48] T. Hastie, R. Tibshirani, and J. Friedman, "High-dimensional problems: p N," 2009, pp. 649–698. doi: 10.1007/978-0-387-84858-7_18.

[49] S. Tahvili and L. Hatvani, "Transformation, vectorization, and optimization," in *Artificial Intelligence Methods for Optimization of the Software Testing Process*, Elsevier, 2022, pp. 35–84. doi: 10.1016/B978-0-32-391913-5.00014-2.

[50] W. Jia, M. Sun, J. Lian, and S. Hou, "Feature dimensionality reduction: a review," *Complex & Intelligent Systems*, vol. 8, no. 3, pp. 2663–2693, Jun. 2022, doi: 10.1007/s40747-021-00637-x.

[51] M. Jafarzadegan, F. Safi-Esfahani, and Z. Beheshti, "Combining hierarchical clustering approaches using the PCA method," *Expert Syst Appl*, vol. 137, pp. 1–10, Dec. 2019, doi: 10.1016/j.eswa.2019.06.064.

[52] V. V Shah, D. Muzyka, C. Guidarelli, K. Sowalsky, F. B. Horak, and K. M. Winters-Stone, "Chemotherapy-induced peripheral neuropathy and falls in cancer survivors relate to digital balance and gait impairments," *JCO Precis Oncol*, vol. 8, 2024, doi: 10.1200/PO.23.00312.

[53] L. Rocchi, L. Chiari, and A. Cappello, "Feature selection of stabilometric parameters based on principal component analysis," *Med Biol Eng Comput*, vol. 42, no. 1, pp. 71–79, Jan. 2004, doi: 10.1007/BF02351013.

[54] D. Tebala and D. Marino, "Companies and artificial intelligence: An example of clustering with Orange," 2023, pp. 1–12. doi: 10.1007/978-3-031-33461-0_1.

[55] G. H. Shah, "An improved DBSCAN, a density-based clustering algorithm with parameter selection for high dimensional data sets," in *2012 Nirma University International Conference on Engineering (NUiCONE)*, IEEE, Dec. 2012, pp. 1–6. doi: 10.1109/NUICONE.2012.6493211.