An Acad Bras Cienc (2024) 96(1): e20230409 DOI 10.1590/0001-3765202420230409
Anais da Academia Brasileira de Ciências | *Annals of the Brazilian Academy of Sciences*
Printed ISSN 0001-3765 I Online ISSN 1678-2690
www.scielo.br/aabc | www.fb.com/aabcjournal

CHEMICAL SCIENCES

Clustering honey samples with unsupervised machine learning methods using FTIR data

FATIH M. AVCU

Abstract: This study utilizes Fourier transform infrared (FTIR) data from honey samples to cluster and categorize them based on their spectral characteristics. The aim is to group similar samples together, revealing patterns and aiding in classification. The process begins by determining the number of clusters using the elbow method, resulting in five distinct clusters. Principal Component Analysis (PCA) is then applied to reduce the dataset's dimensionality by capturing its significant variances. Hierarchical Cluster Analysis (HCA) further refines the sample clusters. 20% of the data, representing identified clusters, is randomly selected for testing, while the remainder serves as training data for a deep learning algorithm employing a multilayer perceptron (MLP). Following training, the test data are evaluated, revealing an impressive 96.15% accuracy. Accuracy measures the machine learning model's ability to predict class labels for new data accurately. This approach offers reliable honey sample clustering without necessitating extensive preprocessing. Moreover, its swiftness and cost-effectiveness enhance its practicality. Ultimately, by leveraging FTIR spectral data, this method successfully identifies similarities among honey samples, enabling efficient categorization and demonstrating promise in the field of spectral analysis in food science.

Key words: Fouirer transform infrared spectrophotometer, hierarchical clustering analysis, machine learning, deep Learning.

INTRODUCTION

Honey is a nutritious and energy-rich food produced by bees from the liquid, nectar of flowers. Honey has been known for centuries and is one of the oldest foods. It has never lost its importance in human nutrition and health. Honey is the first sweetener discovered by mankind. Most honey consists of carbohydrates (such as glucose, fructose, and sucrose), water, and small amounts of other substances such as proteins, ash, amino acids, enzymes, vitamins, and phenolic acids. It is known that each of these components, present in small quantities, has a characteristic nutritional value or medicinal properties (Aly et al. 2021). Although the main components of honey (carbohydrates and water) are almost the same in all honey samples, the chemical composition and physical properties of honey differ depending on the plant species from which the bees take the nectar. In this way, it is possible to determine where the honey comes from or whether foreign additives have been added to it.

Since honey has economic value, many studies have been conducted on it (Erejuwa et al. 2012, Snowdon & Cliver 1996, Meo et al. 2017, Przybylowski & Wilczynska 2001, Kwakman & Zaat 2012, Molan 1996). The aim of this study is to propose a machine learning combined with FTIR spectroscopy to cluster honey samples. After clustering, Deep Learning was applied, and it was found that accurate clustering was performed at a rate of 96.15%. In this study, I validated the

feasibility of the machine learning algorithm in combination with FTIR spectroscopy.

The main purpose of cluster analysis is to find a natural classification of data sets with complex structures in order to separate objects into homogeneous and inhomogeneous clusters. When studying unknown structures in nature, cases with natural classifications may need to be reconsidered. Cluster analysis can be used to treat situations whose natural classification is determined by a small number of variables by increasing the number of variables and investigating whether the previous classification has changed. For example, conditions such as the positive contributions of evolving technology in data collection, the development of modern measurement instruments, and the ability to collect data on new variables from units require control of traditional information. Therefore, the classification of variables and units should be reconsidered using multivariate statistical methods.

Fourier-transform infrared (FTIR) analysis is a chemical analysis method that measures the intensity of the infrared beam passing through the sample based on the wavenumber of the beam using the Fourier transform mathematical method. It is used not only for identification of microbial cells (Ojeda & Dittrich 2012), but also for structural analysis of macromolecules (Sazonova et al. 2019). The applications of FTIR spectroscopy are widespread due to its broad spectrum (Sivashanmugam & Nair 2016, Huang et al. 2018). Obtaining fast results has made the use of traditional FTIR popular. FTIR spectroscopy is a direct and reversible method. This spectroscopic method (Gómez-Ordóñez & Rupérez 2011) which provides results in a short time and with a small amount of sample, is used in the analysis of solid, liquid and gaseous samples.

Machine learning began to be developed in the 1960s (Alpaydin 2010, Samuel 1959). Contrary to popular belief, the mathematical background is not new. For deep neural networks, it can be said that classical neural networks are a special case of multilayer and multi-neural networks (Sun 2019). The most important feature of deep learning networks, which are designed in different models depending on the application domain, is that no separate study is required to extract the features suitable for the problem. In deep structure layers, the features are formed by learning the network. Deep learning networks, which can decide for themselves what information to learn instead of using the information presented to them, therefore provide more successful results than classical methods (Hinton 1989, Asadi-aghbolaghi et al. 2017, Pereira & Oliveira 2017). Deep learning consists of an advanced neural network with many hidden layers. Deep Learning is used in various applications such as image recognition (Wu & Chen 2015, Pak & Kim 2017), computer vision (Voulodimos et al. 2018, Borraz-Martínez et al. 2022), text classification (Minaee et al. 2021, Liu et al. 2017), multiple classification (Cengil & Cinar 2019, Kim et al. 2020) and regression problems (Salaken et al. 2019, Malek et al. 2018).

Related Work

Segato et al. proposed a multivariate machine learning model to understand how the physicochemical properties of honey change with heat (Segato et al. 2019). Increasing temperatures significantly modified moisture, hydroxymethylfurfural content and lightness. In their study, they used the support vector model. They found that heating honey as a pretreatment, especially raising the temperature above 39 degrees, caused a significant change in the internal structure of honey. Liu et al. used the Leave One Out Cross Validation Test (Liu et al.

2022) after obtaining the sequence of honey by DNA extraction with High Throughput Sequencing (HTS)-based metabarcoding method. Using this method, they were able to determine the geographical origin of the honey at a rate of 99%. Chien et al. (2019) classified the pretreated honey spectra using multi-layer perceptron (MLP), support vector machines (SVM) and principal component analysis (PCA) and showed that SVM and PCA gave better results than MLP (Chien et al. 2019). Also they investigated the effectiveness of several spectrum preprocessing technologies for classifying honey samples. Noviyanto A & Abdulla WH (2020) performed classification using IR spectra of honey samples. They used support vector classifier (SVC) and k-nearest neighbors (kNN) as supervised learning algorithms for classification. They used Hyperspectral imaging technique for their study. The combination of hyperspectral imaging and machine learning offers a promising, fast, automatic and non-invasive approach for honey botanical origin classification. They also reported a classification success of 90% for closed clusters and 88% for open clusters (Noviyanto & Abdulla 2020). Al-Awadhi MA and Deshmukh RR used machine learning to determine the botanical origin of honey. In their method, they used Hyperspectral imaging to get data and linear discriminant analysis (LDA) to extract features and reduce the number of dimensions, and then used SVM and KNN algorithms for classification. They reported the accuracy of the proposed model as 95.13% (Al-Awadhi & Deshmukh 2020). Batista BL et al. performed a study based on SVM (Multilayer Perceptron) and Random Forests algorithms (Batista et al. 2012). The authors used Inductively Coupled Plasma Mass Spectrometry (ICP-MS) as the experimental technique. The authors' study focused on the determination of multi-element content in Brazilian honey samples and identified forty-two chemical elements. They obtained results with 65%, 83% and 79% accuracy, respectively. They stated that they were able to quickly find the geographical origin of a honey using their proposed method. Anjos O et al. conducted a neural network-based study to determine the botanical origin of honey using moisture content, electrical conductivity, water activity, ash content, pH, and free acidity (Anjos et al. 2015). The authors found that the combination of FTIR, GC–MS, PCA, and neural networks was able to successfully discriminate the botanical origin of honey with an accuracy of 96% Deep learning, although used with large data sets, has recently been shown to be a good tool for classification examples with small data sets (Karakaplan & Avcu 2021, Avcu 2021). The authors classify some drugs by Monte Carlo sampling with a combination of Genetic algorithm (GA) and Deep neural network (DNN) due to the stochastic nature of the field, exponential number of variables and few chemical species. They also optimized the DNN parameters with GA and achieved a success rate of 93.8%.

FTIR and multivariate data processing tools are often used for the classification of honey samples. However, deep learning has many advantages that can be used for honey sample classification. Deep learning models are well-suited for modeling non-linear relationships between variables, which can be useful for honey sample classification.

MATERIALS AND METHODS

This study was carried out on honey samples from Turkey with different botanical and geographical origins. Before analysis, the samples were stored at room temperature and in a dark place. After scanning the honey samples with the Spectrum 100 (Perkin-Elmer Inc.) FTIR-ATR spectrometer, the obtained spectra were used for PCA and HCA. The dataset resulting from the clustering

process was reclassified by Deep Learning. The open source libraries Tensorflow (Abadi et al. 2021), Keras (Chollet et al. 2022) and Scikit-learn (Buitinck et al. 2022) are used for the calculations.

FTIR measurements

To dilute all honey samples with very high viscosity, 5 ml of carbon tetrachloride was added to 5 ml of honey sample volume. This process was applied equally to all honey samples. (Bailey 1965, Terrab et al. 2003, Costa et al. 2016, Verma 2020). Measurements were made in the middle range of IR (wavelength range from 4000-600 cm^{-1}). For each sample, 3 separate measurements were made with the FTIR instrument and the mean values were taken and analyzed. After each measurement was taken, the FTIR spectrometer was thoroughly cleaned to maintain its accuracy and prevent contamination from previous samples. The results obtained from the measurements were stored in Csv format without preprocessing. The spectra obtained from the measurements with the FTIR device of 65 samples are shown in Figure 1.

The differences seen in Figure 1 may be due to botanical and geographical origins, harvesting and processing methods, different physical and chemical properties or a combination of these factors and other unknown factors.

Chemometric methods

Cluster analysis, one of the multivariate statistical techniques, is used to classify ungrouped and unknown data according to their similarity. Cluster analysis is similar to discriminant analysis in that it aims to collect similar samples in the same groups, and to factor analysis in that it aims to collect similar variables in the same groups, and it also has data reduction features.

Principal component analysis (PCA) is an unsupervised learning method. The basic idea of PCA is to transform the original features into a new feature array in order of importance via a set of orthogonal vectors (Wu et al. 2018). It is commonly used to obtain a graphical representation with lower dimensions that describes the maximum variation in a data set. The first component of PCA considers the largest

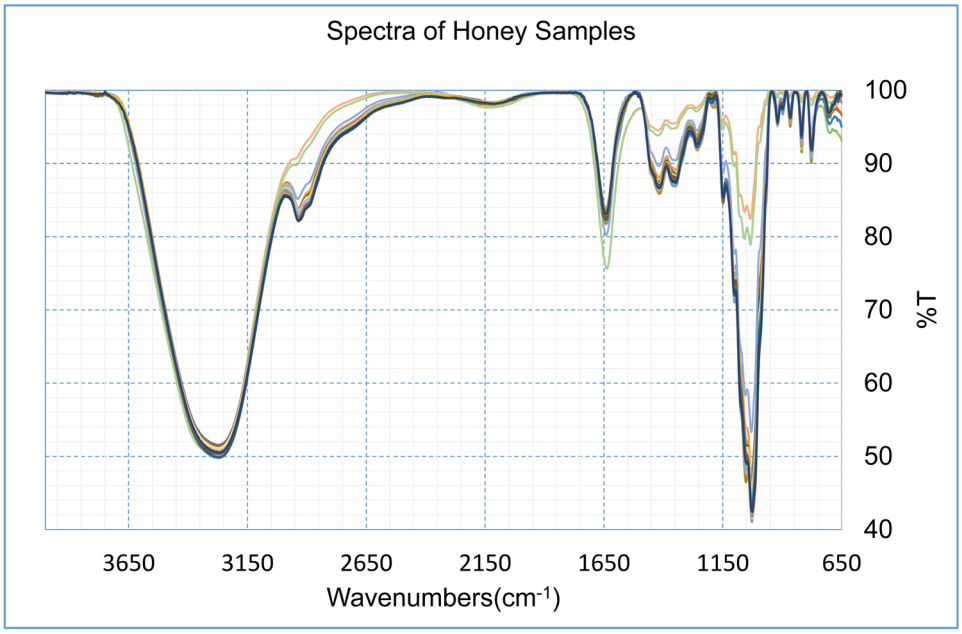

Figure 1. The FTIR-ATR Spectra of Honey Samples.

possible variability in the data, while the second component considers as much of the remaining variability as possible.

Another unsupervised learning algorithm is Hierarchical cluster analysis (HCA) (Granato et al. 2018), whose goal is to collect similar samples in the same groups or clusters, identify these clusters, and predict which group the new samples belong to based on the similarity of the samples with respect to all variables. Clustering is based on the similarity (closeness) or dissimilarity (distance) of two samples. In hierarchical clustering methods, the clusters are combined one by one, and once one group is combined with another, they are not separated in the following steps. Divisive clustering, on the other hand, is a top-down approach, where a single cluster is divided into smaller sub-clusters in each step.

There are many methods for determining the number of clusters in a data set. The oldest and most commonly used of these methods is the elbow method (Coates & Ng 2012). In the Elbow method, the number of clusters (k) is changed between 1 and 10. For each k value, the WCSS (Within Cluster Sum of Squares) value is calculated as in Equation 1.

$$WCSS = \sum_{c_k}^{c_n} \sum_{d_i}^{d_m} {}_{in\ ci} dis\tan ce(d_i, C_k)^2 \qquad \text{Eq.1}$$

C, d are cluster centroids and data point each cluster respectively.

WCSS is the sum of the square of the distance between each point and the center of a cluster. The WCSS value and the k value are plotted on a graph. As the number of clusters increases, the WCSS value decreases. When k=1, the WCSS value is the largest. In Figure 2, it can be seen that at one point the graph changes rapidly and takes the shape of an elbow. From this point on, the graph moves almost parallel to the x-axis. The k-value corresponding to this point is the optimal k-value or the optimal number of clusters. In Figure 2, the number of clusters is clearly indicated as 5.

Another method for determining the number of clusters is the dendrogram method. Ward's method and Euclidean distance were used in the calculations to create the dendrogram. L. Ferreira and D. Hitchcock stated in their study that Ward's method is the best among the other

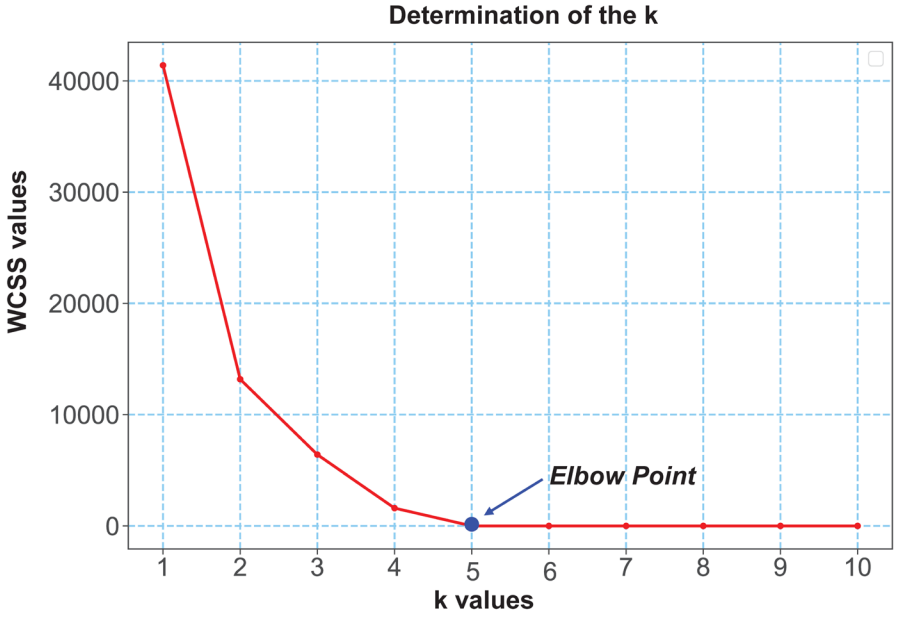

Figure 2. Identification of The Elbow Point.

methods (Ferreira & Hitchcock 2009). Ward's method is also called the method of least variance (Ward 1963). In short, Ward's method performs the merging process based on the variances. The merging process starts with the clusters with the least variance. The method is calculated as in equation 2.

$$SSE = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} = \overline{y}_i)' - (y_{ij} - \overline{y}_{ij}) \quad \text{Eq.2}$$

y_{ij} is the *j*th object in the *i*th cluster and n_i is the number of objects in the *i* th cluster.

There are many distance methods for calculating interclusters. Euclidean distance is one of them. M. Sha et al. reported that Euclidean distance gave the best results in their studies based on Raman spectroscopy (Sha et al. 2020). From Figure 3, it can be seen that the graph starts with 65 separate clusters for 65 honey samples. The number of clusters can be determined with a line drawn parallel to the x-axis from the point of sharp increase in the dendogram. The number of points where this line intersects with the dendrogram gives the number of clusters. Each point where the red line intersects in Figure 3 represents a cluster. The number of 5 clusters found confirms the number k in Figure 2.

In machine learning, when using a clustering algorithm, it requires that all features of the data have the same size (Jain et al. 1999). These differences in original features can cause problems for many machine learning models. Variables measured at different magnitudes in the data set do not contribute equally to the fit of the model and the learning function, leading to bias in the result. For example, since clustering algorithms are distance-based, a large value for a feature in our data will result in it being the dominant feature. To avoid this situation, I scale the data using standardization or Z-score normalization methods. I used the "StandardScaler" function in the "Sklearn. Preprocessing" library in Python for this process. The StandardScaler operation adjusts each column of data so that its mean is 0 and its standard deviation is 1.

For the HCA method, the score vectors obtained from PCA were used (Figure 4). Python scripting language was used for PCA and HCA analysis. In this study, using agglomerative clustering, hierarchical clustering is performed with Euclidean distance calculations, starting with the most similar samples. This process is performed using the AgglomerativeClustering

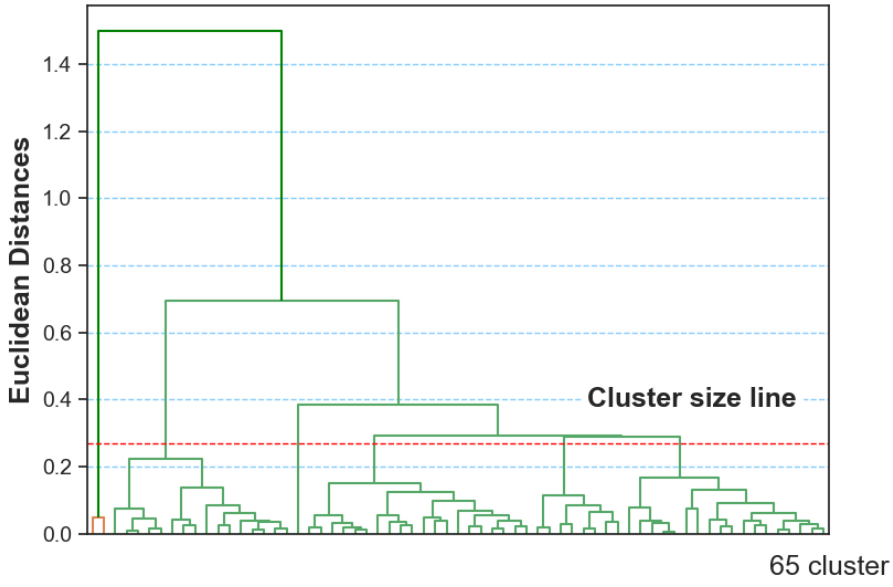

Figure 3. Cluster Numbers with Dendrogram.

function from the Sklearn.cluster library. Other hyperparameters of HCA are Euclidean and Ward parameters for affinity and linkage, respectively.

Results and Discussion

In Figure 5, the samples from Cluster 3, which are very different from the other samples, should not be considered outliers. Outlier samples are samples that cannot be assigned to any cluster as a result of the cluster analysis.

Looking at the HCA (Figure 5) and K-Means (Supplementary Material - Figure S1) clustering results, I see that there are transitions between Cluster 1 and Cluster 5, and between Cluster 2 and Cluster 5. Similarly, the HCA (Figure 5) and Gaussian Mixture (Figure S2) clustering results show a stronger transition between Cluster 1 and Cluster 5 and between Cluster 2 and Cluster 5.

The accuracy of the clustering process was tested using a Deep Learning based classification process. The outcome of the classification using Deep Learning is binary, either 0 or 1, due to its structure. Therefore, the column with the cluster number in the data file is preprocessed in the Sklearn library and LabelEncoding is applied to it.

For the classification process, 20% of the data set was randomly selected for testing and the remaining portion was used as training data. Process, the train_test_split command from the sklearn.model_selection library is used.

The structure of the deep learning model consists of the input, the hidden and the output layer. Each parameter for the hidden layer and the epoch number was tried 30 times, and the average is shown in Figure S3. The model created in this study consists of 3 layers in total. For the Deep Learning structure; FTIR data was dimensionally reduced using the boxcar algorithm and given to the input neurons. In the structure, 171 input neurons and a hidden layer structure with 338 neurons were used. The rectified linear unit (relu) function is used as the activation function of the input and hidden layers. The output value of the relu function is between zero and +∞. The reason why the relu activation function is commonly used is that inputs greater than 0 have a fixed

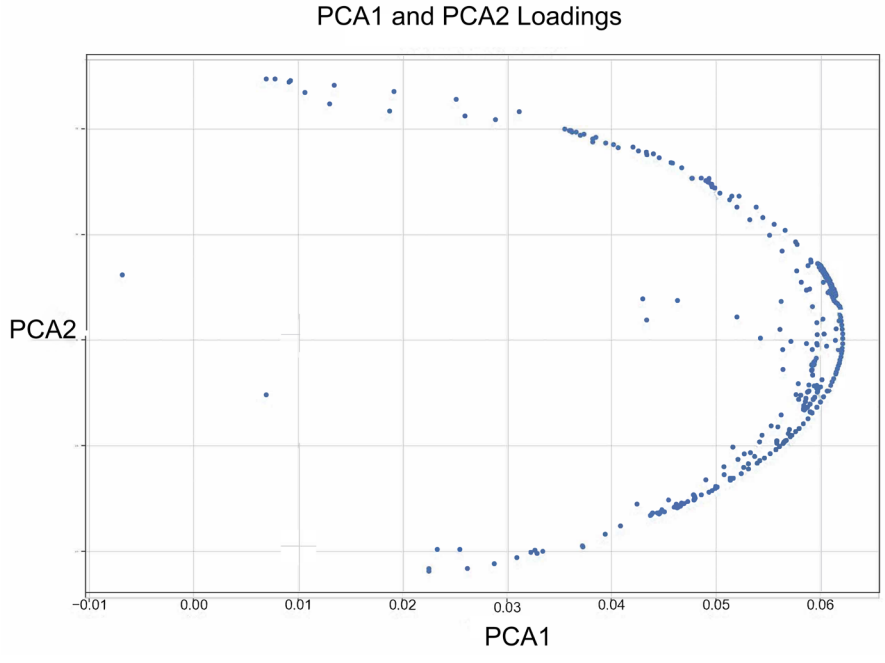

Figure 4. Loading Plot.

derivative value. The sigmoid function is used to activate the output layer. The sigmoid function is a continuous and differentiable function. It is often preferred because it is non-linear. It takes a value in the range of 0 to 1, depending on the input value. Since I have 5 classes, the output layer is set to 5.

The most commonly used function Adaptive Moment Estimation (adam) was used as an optimizer (Ruder 2017). Deep Learning is initiated with an epoch number of 150. Binary_crossentropy was used as the loss function in the structure. It is known that the lower the loss function, the better the structure is optimized.

Figure S4 shows the success of formation and losses. The loss function of the model is very close to 0.1. After the training, it can also be seen that the accuracy is above 96%. Looking at both the loss and accuracy, I can say that the training is successful.

One of the methods used to determine model accuracy is the complexity matrix. The complexity matrix created after testing the algorithm with 13 test data that were not used for training is shown in Figure S5. The rows of the matrix show the true classes and the columns of the matrix show the predicted classes. From Figure S5, it can be seen that the trained model correctly predicted the test data.

CONCLUSIONS

In this study, the focus was on developing a method for clustering honey samples based on their geographical origin without the need for a complex chemical separation process. Honey samples from different regions were analyzed using Fourier Transform Infrared (FTIR) spectroscopy, which provides a unique fingerprint for each sample. The FTIR data were used to identify clusters of samples based on similarities in their chemical composition.

To determine the optimal number of clusters, two different approaches were used. First, each frequency point of the FTIR data was evaluated as a feature and the number of features was reduced to 2 dimensions using Principal Component Analysis (PCA). These 2D data were

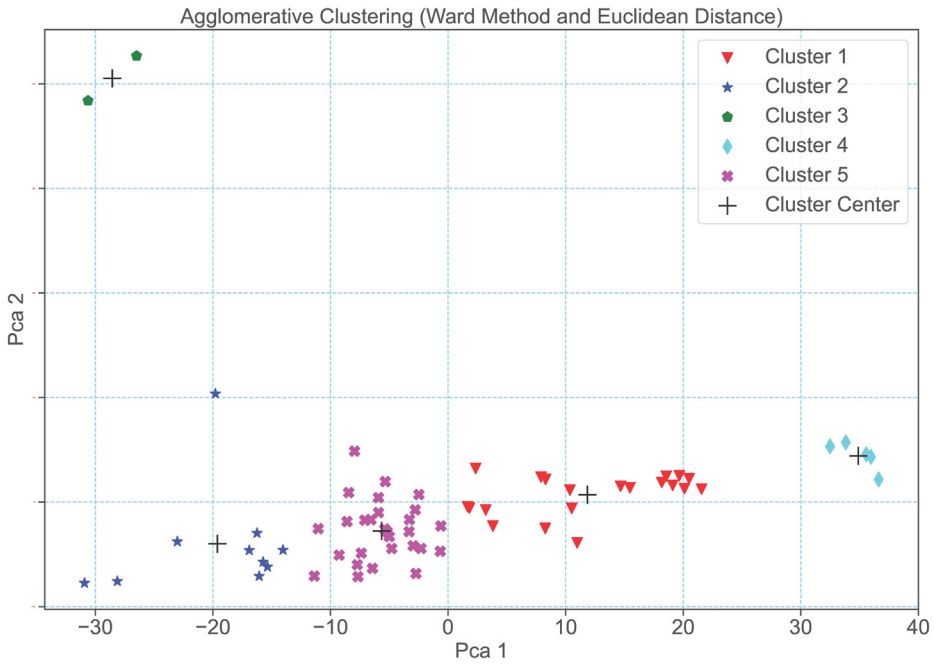

Figure 5. Clustering of Honey Samples with HCA.

then clustered using Hierarchical Clustering Analysis (HCA). Secondly, Deep Learning was used to reclassify the clustered data, and the accuracy of the model was evaluated using test data.

The results showed that the proposed model can accurately predict the geographical origin of honey samples at a rate of 96.15%, demonstrating that chemometric methods applied to FTIR data can quickly and accurately cluster honey samples without the need for sample preparation. The study also highlighted the potential of Deep Learning to automatically model complex nonlinear relationships and provide more accurate results for small data sets such as honey samples.

The movement of bees by beekeepers in the western region of Eastern Anatolia was also noted, with no clear boundaries between the regions they cover. The botanical makeup of the plants from which the bees collect pollen was mentioned as a potential factor influencing the chemical composition of honey samples, with differences potentially arising due to variations in the region's topography and altitude above sea level.

Acknowledgments

The author would like to thank Gökhan Durmaz, Department of Food Engineering, Inonu University and Tamer Arslan, Department of Food Processing, Darende Vocational School, Turgut Özal University, for support FTIR spectra.

REFERENCES

ABADI M ET AL. 2021. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. Software available from tensorflow.org.

AL-AWADHI MA & DESHMUKH RR. 2021. Honey Classification us- ing Hyperspectral Imaging and Machine Learning. in 2021 Smart Technologies, Communication and Robotics (STCR) 1-5.

ALPAYDIN E. 2010. Introduction to machine learning, 2nd ed., Cambridge, MA: MIT Press, 400 p.

ALY AA, MARAEI RW & ABD-ALLAH MM. 2021. Evaluation of physical, biochemical properties and cell viability of gamma irradiated honey. Food Measure 15: 4794-4804.

ANJOS O, IGLESIAS C, PERES F, MARTÍNEZ J, GARCÍA Á & TABOADA J. 2015. Neural networks applied to discriminate botanical origin of honeys. Food Chem 175: 128-136.

ASADI-AGHBOLAGHI M, CLAPÉS A, BELLANTONIO M, ESCALANTE HJ, PONCE-LÓPEZ V, BARÓ X, GUYON I, KASAEI S & ESCALERA S. A Survey on Deep Learning Based Approaches for Action and Gesture Recognition in Image Sequences. 2017. 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), Washington, DC, USA, 2017, p. 476-483.

AVCU FM. 2021. Az Veri Setli Çalışmalarda Derin Öğrenme ve Diğer Sınıflandırma Algoritmalarının Karşılaştırılması: Agonist ve Antagonist Ligand Örneği. İnönü Üniversitesi Sağlık Hizmetleri Meslek Yüksek Okulu Dergisi 10(1): 356-371.

BAILEY L. 1965. Paralysis of the honey bee, Apis mellifera Linnaeus. J Invertebr Pathol 7(2): 132-140.

BATISTA BL, SILVA LRS, ROCHA BA, RODRIGUES JL, BERRETTA-SILVA AA, BONATES TO, GOMES VSD, BARBOSA RM & BARBOSA F. 2012. Multi-element determination in Brazilian honey samples by inductively coupled plasma mass spectrometry and estimation of geographic origin with data mining techniques. Food Res Int 49: 209-215.

BORRAZ-MARTÍNEZ S, TARRÉS F, BOQUÉ R, MESTRE M, SIMÓ J & GRAS A. 2022. Varietal quality control in the nursery plant industry using computer vision and deep learning techniques. J Chemom 36(2): e3320. doi:10.1002/cem.3320.

BUITINCK L ET AL. 2022. API design for machine learning software: experiences from the scikit-learn project. In: 108-122. Software available from scikit-learn.org.

CENGIL E & CINAR A. 2019. Multiple Classification of Flower Im- ages Using Transfer Learning. in 2019 International Artificial Intelligence and Data Processing Symposium (IDAP) 1-6.

CHIEN H-Y, SHIH A-T, YANG B-S & HSIAO VKS. 2019. Fast honey classification using infrared spectrum and machine learning. Math Biosci Eng 16: 6874-6891.

CHOLLET F ET AL. 2022. Keras. https://keras.io.

COATES A & NG AY. 2012. Learning Feature Representations with K-Means. In: MONTAVON G, ORR GB & MÜLLER KR (Eds), Neural Networks: Tricks of the Trade, Lecture

Notes in Computer Science, vol 7700, Springer, Berlin, Heidelberg.

COSTA LS, ALBUQUERQUE BR, TEİXEİRA DM & PESSOA AM. 2016. Infrared Spectroscopy as a Tool for Monitoring Honey Authenticity. Food Chem 213: 183-188.

EREJUWA OO, SULAIMAN SA & AB WAHAB MS. 2012. Honey: A Novel Antioxidant. Molecules 17: 4400-4423. https://doi.org/10.3390/molecules17044400.

FERREIRA L & HITCHCOCK DB. 2009. A Comparison of Hierarchical Methods for Clustering Functional Data. Communications in Statistics - Simulation and Computation 38(9): 1925-1949.

GÓMEZ-ORDÓÑEZ E & RUPÉREZ P. 2011. FTIR-ATR spectroscopy as a tool for polysaccharide identification in edible brown and red seaweeds. Food Hydrocolloids 25: 1514-1520.

GRANATO D, SANTOS JS, ESCHER GB, FERREIRA BL & MAGGIO RM. 2018. Use of principal component analysis (PCA) and hierarchical cluster analysis (HCA) for multivariate association between bioactive compounds and functional properties in foods: A critical perspective. Trends Food Sci. Technol 72: 83-90.

HINTON GE. 1989. Learning distributed representations of concepts. Clarendon Press/Oxford University Press, p. 1-12.

HUANG W, ZHANG L & WU D. 2018. Application of Fourier transform infrared spectroscopy in characterization of functional groups and structures of lignin. Polymers 10(9): 1007.

JAIN AK, MURTY MN & FLYNN PJ. 1999. Data clustering: a review. ACM Computing Surveys (CSUR) 31(3): 264-323.

KARAKAPLAN M & AVCU FM. 2021. Classification of some chemical drugs by genetic algorithm and deep neural network hybrid method. Concurrency and Computation: Practice and Experience 33: e6242.

KIM J, SHIN S, YU Y, LEE J & LEE K. 2020. Multiple Classification with Split Learning. in The 9th International Conference on Smart Media and Applications 358-363.

KWAKMAN PHS & ZAAT SAJ. 2012. Antibacterial components of honey. IUBMB Life 64: 48-55.

LIU J, CHANG W-C, WU Y & YANG Y. 2017. Deep Learning for Extreme Multi-label Text Classification. in Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, p. 115-124.

LIU S, LANG D, MENG G, HU J, TANG M, ZHOU X. 2022. Tracing the origin of honey products based on metagenomics and machine learning. Food Chem 371: 131066.

MALEK S, MELGANI F & BAZI Y. 2018. One-dimensional convolutional neural networks for spectroscopic signal regression. J Chemom 32: e2977.

MEO SA, AL-ASIRI SA, MAHESAR AL & ANSARI MJ. 2017. Role of honey in modern medicine. Saudi Journal of Biological Sciences 24: 975-978.

MINAEE S, KALCHBRENNER N, CAMBRIA E, NIKZAD N, CHENAGHLU M & GAO J. 2021. Deep Learning--based Text Classification: A Comprehensive Review. ACM Comput Surv 54(3): 1-40. https://doi.org/10.1145/3439726.

MOLAN PC. 1996. The role of honey in the management of wounds. J Wound Care 8: 415-418.

NOVIYANTO A & ABDULLA WH. 2020. Honey botanical origin classification using hyperspectral imaging and machine learning. J Food Eng 265: 109684.

OJEDA JJ & DITTRICH M. 2012. Fourier Transform Infrared Spectroscopy for Molecular Analysis of Microbial Cells. In: NAVID A (Ed), Microbial Systems Biology, Methods in Molecular Biology, vol 881, Humana Press, Totowa, NJ, p. 187-211.

PAK M & KIM S. 2017. A review of deep learning in image recognition. in 2017 4th International Conference on Computer Applications and Information Processing Technology (CAIPT) p. 1-3.

PEREIRA AS & OLIVEIRA LF. 2017. Deep learning for content-based image retrieval: a comprehensive study. An Acad Bras Cienc 89(4): 2769-2785.

PRZYBYLOWSKI P & WILCZYNSKA A. 2001. Honey as an environmental marker. Food Chem 74: 289-291.

RUDER S. 2017. An overview of gradient descent optimization algorithms. Tech Rep arXiv:1609.04747, arXiv.

SALAKEN SM, KHOSRAVI A, NGUYEN T & NAHAVANDI S. 2019. Seeded transfer learning for regression problems with deep learning. Expert Systems with Applications 115: 565-577.

SAMUEL AL. 1959. Some studies in machine learning using the game of checkers. IBM Journal of research and development 3(3): 210-229.

SAZONOVA S, GRUBE M, SHVIRKSTS K, GALOBURDA R & GRAMATINA I. 2019. FTIR spectroscopy studies of high pressure-induced changes in pork macromolecular structure. J Mol Struct 1186: 377-383.

SEGATO S ET AL. 2019. Multivariate and machine learning models to assess the heat effects on honey physicochemical, colour and NIR data. Eur Food Res Technol 245: 2269-2278.

SHA M, ZHANG D, ZHANG Z, WEI J, CHEN Y, WANG M & LIU J. 2020. Improving Raman spectroscopic identification of rice varieties by feature extraction. J Raman Spectrosc 51(4): 702-710.

SIVASHANMUGAM A & NAIR BG. 2016. Fourier transform infrared spectroscopy: an evolving method for assessing soil properties. Int J Sci Res Publ 6(6): 293-298.

SNOWDON JA & CLIVER DO. 1996. Microorganisms in honey. Int J Food Microbiol: 1-26.

SUN R. 2019. Optimization for deep learning: theory and algorithms. https://doi.org/10.48550/arXiv.1912.08957.

TERRAB A, RECAMALES ÁF, HERNANZ D & HEREDİA FJ. 2003. Determination of the Organic Acid Composition of Honey by High-Performance Liquid Chromatography. J Chromatogr A 1012(1): 81-89.

VERMA ML. 2020. Biotechnological Approaches in Food Adulterants. 1st ed., CRC Press, 344 p. https://www.routledge.com/Biotechnological-Approaches-in-Food-Adulterants/Verma/p/book/9780367560676.

VOULODIMOS A, DOULAMIS N, DOULAMIS A & PROTOPAPADAKIS E. 2018. Deep Learning for Computer Vision: A Brief Review. Comput Intell Neurosci 2018: e7068349.

WARD JH. 1963. Hierarchical Grouping to Optimize an Objective Function. J Am Stat Assoc 58(301): 236-244.

WU M & CHEN L. 2015. Image recognition based on deep learning. in 2015 Chinese Automation Congress (CAC) 542-546.

WU SX, WAI HT, LI L & SCAGLIONE A. 2018. A Review of Distributed Algorithms for Principal Component Analysis. Proceedings of the IEEE 106(8): 1321-1340.

SUPPLEMENTARY MATERIAL

Figures S1-S5.

How to cite

AVCU FM. 2024. Clustering honey samples with unsupervised machine learning methods using FTIR data. An Acad Bras Cienc 96: e20230409. DOI 10.1590/0001-3765202420230409.

Manuscript received on April 11, 2023;
accepted for publication on June 14, 2023

FATIH MEHMET AVCU
https://orcid.org/0000-0002-1973-7745

Inonu University, Department of Informatics,
TR-44280 Malatya, Turkey

Correspondence to: **Fatih Mehmet Avcu**
E-mail: fatih.avcu@inonu.edu.tr

Author contributions

F. M. Avcu developed the software, drew the graphics, and wrote the manuscript.