Classification of Chili Plant Origin by Using Multilayer Perceptron Neural Network

Dyah Kurniawati Agustika
School of Engineering
The University of Warwick
Coventry, United Kingdom
Dept. of Physics Education
Universitas Negeri Yogyakarta
Sleman, Indonesia
Dyah.Agustika@warwick.ac.uk

Nur Aeni Ariyanti
Dept. of Biology Education
Universitas Negeri Yogyakarta
Sleman, Indonesia
nuraeni@uny.ac.id

I Nyoman Kusuma Wardana
Dept. of Electrical Engineering
Politeknik Negeri Bali
kusumawardana@pnb.ac.id

Doina Daciana Iliescu
School of Engineering
The University of Warwick
Coventry, United Kingdom
D.D.Iliescu@warwick.ac.uk

Mark Stephen Leeson
School of Engineering
The University of Warwick
Coventry, United Kingdom
Mark.Leeson@warwick.ac.uk

***Abstract*— The geographical origin of the plants can affect the growth and hence the quality of the plants. In this research, the origin of the chili plants has been investigated by using Fourier transform infrared (FTIR) spectroscopy. The spectroscopy generated 3734 data with a wavenumber range from 4000-400 cm^{-1}. The pre-processing of the spectra was done by using baseline correction and vector normalization. The analysis was then taken in the biofingerprint area of 1800-900 cm^{-1} range which has 934 data points. Feature extraction for dimension reduction was achieved using principal component analysis (PCA). The PC scores from PCA were then fed into a k-means and a multilayer perceptron neural network (MLPNN). The k-means clustering shows that the samples can be distinguished into three different groups. Meanwhile, for the MLPNN, the number of the hidden layer's neurons and the learning rate of the system were optimized to get the best classification result. A hidden layer with twenty neurons had the highest accuracy, while a learning rate of 0.001 had the highest value of 100%.**

Keywords—Fourier transform infrared spectroscopy, origin of plants, multilayer perceptron neural network, k-means

I. INTRODUCTION

Environmental conditions such as water content, light and temperature can greatly affect the optimal growth of plants. Elements in the soil that are absorbed by plants also affect their growth [1]. The environmental conditions in which plants grow can also influence their metabolites quality [2]. Therefore, different geographical origin with different environmental condition could affect the growth and quality of the plants. Changes in environmental conditions such as changes in temperature, availability of water and nutrients can cause stress on plants [3]. The external environmental could rapidly changing and potentially damaging the plants that interact with it [2]. Stressed plants will release secondary metabolites in the form of Volatile Organic Compounds (VOCs), which are different from plants under normal conditions [3].

The emission of VOCs can be detected by Fourier Transform Infrared (FTIR) spectroscopy. This has been used to identify honey based on its geographic and botanical origin [4]. Polysaccharide analysis using FTIR spectroscopy was used to identify Dendrobiums [5]. Moreover, Christou *et al.* used FTIR to discriminate the origins of Carob flesh and seeds [6].

The output of FTIR is in the form of spectral data with large dimensions; therefore, it must be processed first in order to recognize the sample. The spectral data processing begins with pre-processing and feature extraction and then proceeds by analysis with a pattern recognition machine. The common pre-processing techniques are baseline correction, normalization, smoothing with Savitzky-Golay or Discrete Fourier Transform (DWT). Meanwhile, feature extraction for dimension reduction can be carried out with Principal Component Analysis (PCA) [7] and Fast Fourier Transform (FFT) [8]. Spectra processing for sample recognition is then carried out using pattern recognition machines such artificial neural network, Random Forest, and Linear Discriminant Analysis [9].

In this study, we used FTIR spectroscopy to discriminate chili plants from three different areas with different geographical conditions, which are chili plants from the Cangkringan, Sleman, D.I. Yogyakarta, which is a mountainous area, chili plants from the Bantul, D.I. Yogyakarta, which is a coastal area and Purworejo, Central Java which is a lowland area. Chili plants are known as plants that are rich in vitamins and beneficial for health [10]. The origin of chili plants from plant leaves were tested by using FTIR spectroscopy. After going through the pre-processing and feature extraction stages, the spectral data were then clustered by using k-means clustering method and classified by using the multilayer perceptron model of artificial neural network. The main objective of this research was to find the optimum configuration and parameters of the multilayer perceptron neural network model (MLPNN). The model was then used for the future test.

II. Material and Methods

A. Samples

The samples used in this study were chili leaves taken from three different locations, namely Bantul, D.I. Yogyakarta, Sleman, D.I. Yogyakarta, and Purworejo, Central Java. For one location, four different samples were taken, so that a total of twelve samples were taken from three different areas. For the FTIR test, three leaves were taken from each sample. The leaves were then grounded into pellets and then tested by FTIR spectroscopy.

B. Fourier transform infrared spectroscopy

The FTIR test was then carried out at the Integrated Research and Testing Laboratory, Gadjah Mada University, and the FTIR spectroscopy were in transmittance mode. The results of FTIR are spectra in the mid-infrared region (4000–400 cm^{-1}) and produced 3734 data points.

C. Pre-processing and feature extraction

The pre-processing stage can help improve the accuracy of the pattern recognition system, because this stage aims to eliminate noise and highlight the desired signal information [11]. Here, in the pre-processing stage, baseline correction of spectral data was carried out in the range (4000–400 cm^{-1}), after which vector normalization was performed by dividing each data point in the spectra by the norm [12] in equation (1),

$$Norm = \sqrt{s_1^2 + s_2^2 + s_3^2 + \cdots + s_{3734}^2} \qquad (1)$$

The next step was to cut the data in the biofingerprint area in the range of 1800 to 900 cm^{-1} which was used for further analysis. In this area there was important sample information that was a representation of the sample composition [13].

Spectra in the range of 1800-900 cm^{-1} consisted of 934 data points. These data then became the input for the pattern recognition system. However, the large amount of data 934 × 12 (934 data points by 12 samples) could complicate the pattern recognition process. Therefore, it was necessary to perform feature extraction. One of the feature extraction techniques used in FTIR analysis is PCA and this was employed here. PCA highlights important data information and works to reduce dimensions by transforming data into variables in a new basis [14]. This variable is a linear combination of the original data and is calculated and sorted by order of importance [15].

The spectral data in the form of a 934 × 12 dimension matrix were converted into a covariance matrix which was then manipulated by calculating the eigenvalues and eigenvectors of the covariance matrix and to obtain the data's important characteristics. The eigenvectors were then sorted based on the largest eigenvalues (the principal components [14]).

D. K-Means

The matrix form of the spectral data after preprocessing and PCA in the previous section (with dimensions 11 × 12) could be grouped based on k clusters, where $k \geq 2$▌ This method is called k-means clustering, which is a type of iterative method that partitions a set of samples into k clusters. Samples in a cluster have the same characteristics but are different to those in a different group. By using the distance function, the similarity between two objects can be determined [16], [17].

E. Multilayer Perceptron Neural Network

ANNs can perform complex computational tasks by imitating the work of the human biological nervous system [18]. This study used the MLPNN architecture which consisted of one input layer, a hidden layer and one output layer, and each layer had neurons for processing the data. Spectral data in the range 1800 – 900 cm^{-1} which had been feature extracted with PCA were divided into training and testing data by using the Kennard-Stone algorithm. Training data and training data were split 50%:50%.

The training data in the form of an 11 × 6 matrix became input to the neurons in the input layer. Neurons in the input layer was connected to a node in the hidden layer, where the connections between these neurons had a certain weight. At the beginning of the training process the weights between the layers were randomly initiated, and the input x_i in the input layer was multiplied by the weight w_{ij} that exists between the input and hidden layer.

The neurons of x_i in the input layer were multiplied by the weight w_{ij} to form a preactivation function $y_j = b_j + \sum_i x_i w_{ij}$, where b_j was biased for each neuron in the hidden layer. Furthermore, y_j was the input for the activation function f_j in the hidden layer. The product of f_j and y_j was the output of the hidden layer, h_j, and $h_j = f_j\left(b_j + \sum_i x_i w_{ij}\right)$.). Then h_j were connected to neurons in the output layer by a connection that had weight w_{jk} and formed a preactivation function $y_k = b_k + \sum_j f_j\left(b_j + \sum_i x_i w_{ij}\right) w_{jk}$, where b_k was the bias of each neuron in the outer layer. The system output coming from the output layer was defined as

$$o_k = f_k\left(b_k + \sum_j f_j\left(b_j + \sum_i x_i w_{ij}\right)\right) \qquad (2)$$

The error function then calculated the difference between the system output and the expected output thus

$$E = \left(\frac{1}{2}\right) \sum_k (t_k - o_k)^2 \qquad (3)$$

with t_k being the target or the expected result of the network.

Backpropagation algorithm minimized the value of E by adjusting the weights in the previous layers. The modification of the weight on the connector that connected the input layer to the hidden layer was achieved using

$$\boldsymbol{w_{ij}(t) = w_{ij}(t-1) - \eta\left(\frac{\partial E}{\partial w_{ij}}\right)} \tag{4}$$

Where $\boldsymbol{\eta}$ was the learning rate and t was the number of iterations. Meanwhile the modification of w_{jk}, which was the weight on the connector connecting the hidden layer with the outer layer, was obtained by an equation in the same form as the equation but with $\boldsymbol{w_{ij}}$ replaced by $w_{jk.}$[18].

In this research, to obtain the best result, the parameters of MLPNN were optimized. First, the number of neurons from 1 to 20 in the hidden layer were optimized, then the learning rate of the best architecture were also be optimized.

III. RESULTS AND DISCUSSION

The FTIR spectra of the measurement results were in the range of 4000 – 400 cm^{-1} and produced 3734 data points. To find out whether FTIR could distinguish chili leaf samples based on geographical origin, the data was classified by a pattern recognition machine. However, due to the large amount of data and the possibility of noise in the sampling process, before being processed by the pattern recognition machine, pre-processing was carried out first. Baseline correction was performed on the spectral data and then vector normalization was performed.

The pre-processed data hads dimensions of 3734 × 12 and would require a very large number of neurons to be processed by MLPNN. Therefore, it was necessary to reduce the dimensions by using PCA, which resulted in 11 PC scores which were ordered based on their value (PC1 had the largest value). PCA analysis output data measuring 11 × 12 was then used for further analysis.

After PCA, the data was then analyzed by using the K-means method to prove that different groups could be distinguished. The K-means analysis result are depicted in Fig.1

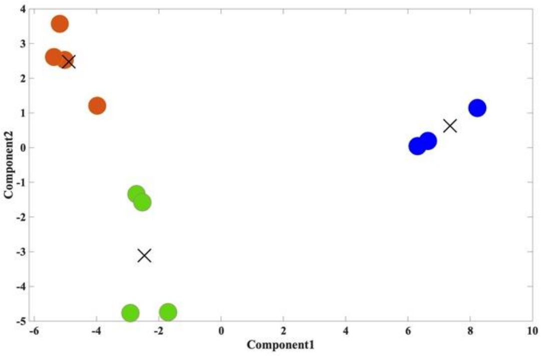

Fig.1. The K-means analysis of the spectra

K-means alone was not enough to measure the performance of the system. Therefore, the MLPNN was used. The PCA data were divided into training and testing data by 50%:50% using the Kennard-Stone algorithm. The MLPNN system architecture used in the analysis is depicted in Fig. 2. The input was training data in the matrix form of 11 × 6 (11 PCs from 6 samples). In this process, a hyperbolic tangent was the hidden layer's activation function, and the log-sigmoid was the one in the output layer. The system output classified samples into three types (based on three different sampling locations), so that two neurons were needed, with a target of 00 for the Purworejo, Central Java sample, 01 for the Sleman, D.I. Yogyakarta sample and 10 for the Bantul, D.I. Yogyakarta sample leaves.

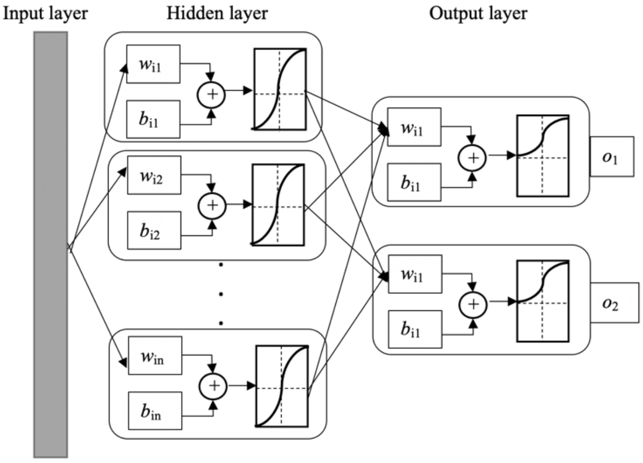

Fig. 2. The architecture of MLPNN

To get the best results, it was necessary to optimize the MLPNN parameters. In this research, the first thing to do was to optimize the number of neurons in the hidden layer. The number of neurons was varied from 1 to 20 and the results searched using a learning rate of 0.001. At the beginning of the training process, the system used random weights, and the training was repeated up to 10 times, then the accuracy was averaged. The accuracies were then compared and the neuron with the highest accuracy was used for the next process. After that, the network architecture with each neuron in the hidden layer was tested using test data. The results of the average network accuracy for the 20 neurons of both training and testing data are presented in Table 1.

TABLE 1. THE MODEL ACCURACY OF DIFFERENT NUMBER OF NEURONS IN THE HIDDEN LAYER

Number of neurons	**Model Accuracy**
1	56%
2	61.33%
3	60%
4	60.67%
5	66%
6	66.67%

Number of neurons	**Model Accuracy**
7	64%
8	66.67%
9	64%
10	61.33%
11	65.33%
12	64%
13	61.33%
14	68%
15	72%
16	56%
17	61.33%
18	62.67%
19	61.33%
20	76%

From Table 1, the highest value of accuracy is seen to be for twenty neurons. An MLPNN with twenty neurons in the hidden layer was thus used for further processing.

The next step was to optimize the learning rate of the network with eight neurons in the hidden layer. The learning rates used were 0.1, 0.01, 0.001, 0.0001, 0.00001. The training process with each learning rate was carried out 10 times then the accuracies were averaged; and the results are presented in Table 2.

TABLE 2. THE MODEL ACCURACY OF DIFFERENT LEARNING RATES

Learning rate	**Model Accuracy**
0.1	55.33%
0.01	46.67%
0.001	76%
0.0001	60.67%
0.00001	60.67%

From table 2 it can be seen that the highest accuracy value is obtained for a learning rate of 0.001. The learning rate and network architecture values with twenty neurons in the hidden layer were then used for further analysis. This system was run 10 times to find the one with the highest accuracy then the weights were stored for the introduction of new samples. From the running results. As a result, the highest accuracy reached 100%, which demonstrated that the system can distinguish the samples.

IV. Conclusion

The discrimination of the chili plants origin from three different areas in D.I. Yogyakarta and Central Java has been investigated. The research used FTIR spectroscopy to detect the VOCs emitted by the plants. The spectral results were then pre-processed by using baseline correction and vector normalization. The resulting big data set was then reduced by using PCA. The PC scores from PCA were then fed into K-means clustering and an MLPNN. The K-means analysis showed that the samples could be divided into three different groups. Meanwhile, for the MLPNN, the parameters, such as the number of neurons and the learning rate, were optimized to obtain the best classification results. From the analysis the best architecture had twenty neurons in the hidden layer, and 0.001 was the learning rates with the highest accuracy that reached 100%.

Acknowledgment

This work is supported by the Indonesia Endowment Fund for Education (LPDP).

References

[1] E. Ceyhan, A. Kahraman, and M. Onder, “The Impacts of Environment on Plant Products,” *Int. J. Biosci. Biochem. Bioinforma.*, vol. 2, no. 1, pp. 48–51, 2012, doi: 10.7763/ijbbb.2012.v2.68.

[2] B. Ncube, J. F. Finnie, and J. Van Staden, “Quality from the field: The impact of environmental factors as quality determinants in medicinal plants,” *South African J. Bot.*, vol. 82, pp. 11–20, 2012, doi: 10.1016/j.sajb.2012.05.009.

[3] Y. L. Dorokhov, T. V. Komarova, and E. V. Sheshukova, *Volatile organic compounds and plant virus-host interaction*. Elsevier, 2014.

[4] F. Guyon, E. Logodin, D. A. Magdas, and L. Gaillard, “Potential of FTIR- ATR diamond in discriminating geographical and botanical origins of honeys from France and Romania,” *Talanta Open*, vol. 3, no. September 2020, 2021, doi: 10.1016/j.talo.2020.100022.

[5] N. D. Chen, N. F. Chen, J. Li, C. Y. Cao, J. M. Wang, and H. P. Huang, “Similarity Evaluation of Different Origins and Species of Dendrobiums by GC-MS and FTIR Analysis of Polysaccharides,” *Int. J. Anal. Chem.*, vol. 2015, 2015, doi: 10.1155/2015/713410.

[6] C. Christou, A. Agapiou, and R. Kokkinofta, “Use of FTIR spectroscopy and chemometrics for the classification of carobs origin,” *J. Adv. Res.*, vol. 10, pp. 1–8, 2018, doi: 10.1016/j.jare.2017.12.001.

[7] G. E. De Benedetto, B. Fabbri, S. Gualtieri, L. Sabbatini, and P. G. Zambonin, “FTIR-chemometric tools as aids for data reduction and classification of pre-Roman ceramics,” *J. Cult. Herit.*, vol. 6, no. 3, pp. 205–211, 2005, doi: 10.1016/j.culher.2005.06.004.

[8] R. Chaber *et al.*, “Distinguishing Ewing sarcoma and osteomyelitis using FTIR spectroscopy,” *Sci. Rep.*, vol. 8, no. 1, pp. 1–8, 2018, doi: 10.1038/s41598-018-33470-3.

[9] J. Zeng *et al.*, “A review of the discriminant analysis methods for food quality based on near-infrared spectroscopy and pattern recognition,” *Molecules*, vol. 26, no. 3, 2021, doi: 10.3390/molecules26030749.

[10] D. K. Agustika *et al.*, “Gas Chromatography-Mass Spectrometry Analysis of Compounds Emitted by Pepper Yellow Leaf Curl Virus-Infected Chili Plants : A Preliminary Study,” *Sep. MDPI*, vol. 8, no. 9, p. 136, 2021.

[11] L. Bai and Y. Liu, “Classification of FTIR cancer data using wavelets and fuzzy C-means clustering,” *Wavelet Appl. Ind. Process. III*, vol. 6001, no. November 2005, p. 60010B, 2005, doi: 10.1117/12.629946.

[12] R. Gautam, S. Vanga, F. Ariese, and S. Umapathy, “Review of multidimensional data processing approaches for Raman and infrared spectroscopy,” *EPJ Tech. Instrum.*, vol. 2, no. 1, 2015, doi: 10.1140/epjti/s40485-015-0018-6.

[13] M. C. D. Santos, C. L. M. Morais, and K. M. G. Lima, “ATR-FTIR spectroscopy for virus identification: A powerful alternative,” *Biomed. Spectrosc. Imaging*, vol. 9, no. 3–4, pp. 103–118, 2020, doi: 10.3233/bsi-200203.

[14] D. K. Agustika, S. N. Hidayat, K. Triyana, D. D. Iliescu, and M. S. Leeson, “Steady-state response feature extraction optimization to enhance electronic nose performance,” *Int. Conf. Electr. Eng. Comput. Sci. Informatics*, vol. 2020-Octob, pp. 144–149, 2020, doi: 10.23919/EECSI50503.2020.9251887.

[15] B. Liu, Y. Li, L. Zhang, and J. Wang, “Interpretation of FTIR spectra by principal components - Artificial neural networks,” *Spectrosc. Lett.*, vol. 39, no. 4, pp. 373–385, 2006, doi: 10.1080/00387010600803664.

[16] J. Pérez-Ortega, N. Nely Almanza-Ortega, A. Vega-Villalobos, R. Pazos-Rangel, C. Zavala-Díaz, and A. Martínez-Rebollar, “ The K -Means Algorithm Evolution ,” *Introd. to Data Sci. Mach. Learn.*, 2020, doi: 10.5772/intechopen.85447.

[17] J. Wu, *Advances in K-means Clustering: A Data Mining Thinking*, vol. 53, no. 9. Springer-Verlag Berlin Heidelberg, 2013.

[18] H. Akbas and G. Özdemir, “An integrated prediction and optimization model of a thermal energy production system in a factory producing furniture components,” *Energies*, vol. 13, no. 22, 2020, doi: 10.3390/en13225999.