



## Model for Identification and Prediction of Leaf Patterns: Preliminary Study for Improvement

Ari Muzakir<sup>1\*</sup>, Usman Ependi<sup>2</sup>

<sup>1,2</sup>Faculty of Computer Science, Universitas Bina Darma, Indonesia

### Abstract.

**Purpose:** Many studies have conducted studies related to automation for image-based plant species identification recently. Types of plants, in general, can be identified by looking at the shape of the leaves, colors, stems, flowers, and others. Not everyone can immediately recognize the types of plants scattered around the environment. In Indonesia, herbal plants thrive and are abundantly found and used as a concoction of traditional medicine known for its medicinal properties from generation to generation. In the current Z-generation era, children lack an understanding of the types of plants that benefit life. This study identifies and predicts the pattern of the leaf shape of herbal plants.

**Methods:** The dataset used in this study used 15 types of herbal plants with 30 leaf data for each plant to obtain 450 data used. The extraction process uses the GLCM algorithm, and classification uses the K-NN algorithm.

**Result:** The results carried out through the testing process in this study showed that the accuracy rate of the leaf pattern prediction process was 74% of the total 15 types of plants used.

**Value:** Process of identifying and predicting leaf patterns of herbal plants can be applied using the K-NN classification algorithm combined with GLCM with the level of accuracy obtained.

**Keywords:** Leaf pattern prediction, GLCM algorithm, K-NN algorithm, identification of herbal plants

**Received** April 2021 / **Revised** May 2021 / **Accepted** November 2021

*This work is licensed under a [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/).*



### INTRODUCTION

Currently, classifying plant types into taxonomic forms that are very suitable is a concern for professionals from various fields such as agronomy, environment, forestry, and others. Typically, the plant species identification process is carried out by a botanist based on a visual inspection of the plant leaves. Leaf-based identification is more reliable for the classification process than identifying the flower and fruit parts because the process is relatively short.

According to Ni'mah and Sutojo [1], the number of herbal plants in their research states that about 38 thousand plant species grow in Indonesia, and more than 2039 species are included in herbal plants. The existence of several plant species or species, especially in Indonesia, makes it quite a challenge and a challenging task to do. It is not arbitrary for people to identify plants by looking at the shape of the leaves, especially herbal plants, which are widely used in Indonesia as traditional medicine. In Indonesia, herbal plants thrive in almost all regions. Identification of this herbal plant is usually made by observing the plant's unique leaf texture, shape, architecture, and color.

The process is carried out by analyzing leaf parameters. Several algorithms identify forms and diseases in unique plants needed to detect leaf diseases [2]. The importance of image processing and machine learning methods is helpful for accurately identifying plant species and leaf diseases. Image segmentation with this model provides more accurate results using an experimentally optimized cluster size [3].

The process of identifying ideal plant species is through leaves. Although in the end, the leaves will change color shape. Leaf color can change in several scenarios, such as changes in weather, changes in light levels, and leaf disease [4]. Image processing is a solution that can be done to solve problems in identifying the types and shapes of leaves from herbal plants. Waliyansyah [5] said that the problems often experienced by humans are fatigue and limitations in vision, especially in this case, distinguishing the shape of the leaves from the types of herbal plants.

---

\* Corresponding Author

Email: [arimuzakir@binadarma.ac.id](mailto:arimuzakir@binadarma.ac.id) (Muzakir), [u.ependi@binadarma.ac.id](mailto:u.ependi@binadarma.ac.id) (Ependi)

DOI: [10.15294/sji.v8i2.30024](https://doi.org/10.15294/sji.v8i2.30024)

The selection of the features and shape of the suitable leaves is the most important aspect of the automatic plant identification process. Recognizing the visual form of this type of herbal plant is an easy job for a botanist, but if a machine does it it is a very complex and computationally expensive process [4]. Furthermore, Wu et al. [6] used image processing methods to formulate an automatic system for introducing plants through leaves in their research. Initially, 12 leaf features were extracted and then reduced to only five variables to form a feature for entering training data and class testing. This technique is used to produce an accuracy rate of 90.3%.

In several studies that have been carried out, recognition and diagnosis methods have been carried out using standard image segmentation, extraction, and pattern recognition procedures [7]. However, recognition methods based on pipelined procedures have made progress subject to two aspects of the problem [8], namely accuracy that depends on feature extraction and relatively complicated pipelined procedures [9].

In this study, the focus of identification was carried out on the leaves of herbal plants to identify and predict the pattern of the leaf shape of herbal plants so that they can be recognized accurately. The pattern recognition is done by extracting the texture features using the GLCM algorithm. Then the classification process is carried out by the K-NN algorithm. For the collection of the dataset, each leaf object from herbal plants will be taken pictures from several different sides and different lighting levels as many as 30 samples. Furthermore, with the GLCM algorithm, the extraction process will be carried out by converting the image to grayscale.

The formation of GLCM in an image is shown as four degrees of gray direction, namely 0, 45, 90, and 135 degrees of gray (grayscale level) [10].

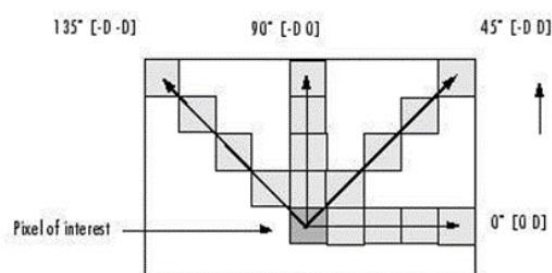


Figure 1. Four grayscale levels [10]

This study modified a dataset to identify, detect, and predict the leaf shape pattern of herbal plants. This research is a preliminary study to conduct experiments in finding a good algorithm in classifying and predicting with machine learning algorithms. There are two contributions to this research, namely:

- 1) We modified a new dataset to identify, detect, and predict the leaf shape pattern of herbal plants.
- 2) Experiment with datasets to measure the best accuracy of machine learning algorithms with the K-NN algorithm.

This paper is divided into four sessions. Session 2 discusses the methodology and data used. The discussion about experiments and test results is in session 3. In session 4, we conclude the results of our work and further research plans.

## METHODS

This section will discuss the processes carried out in research, such as data collection, preprocessing, feature extraction, and object classification.

### Data Collection

This research focuses on experimental studies by making a sample of leaf datasets on herbal plants in Indonesia. The types of plants used in this study amounted to 15 types of herbal plants, namely Red Spinach Leaf, Red Binahong Leaf, Insulin Leaf, Guava Leaf, Castor Leaf, Sauropus androgynus, Hibiscus Leaf, Orthosiphon Aristatus, Bowl Leaf, Bay leaf, Betel leaf, Cherry Leaf, Soursop leaf, Madagascar Periwinkle, and False daisy (Figure 2).

The data sampling process was carried out from January to February 2021 for taking 15 types of herbal plants. Of the 15 herbal plants, 30 pictures of each leaf object were taken from different sides and conditions so that the detection level of accuracy was greater. The total dataset of the 15 types of plants is 450 image data.



Figure 2. Samples of 15 types of leaves from herbal plants

### Data Preprocessing

Preprocessing is an essential step for the leaf object recognition system. Pre-processing applied to image objects is in the form of grayscale and resize. The image object that becomes the input may have different parameter values depending on the adequacy of the light and leaf color. The properties of each image object produce varying contrast and noise [11]. Each image processing in this study is stored in an RGB color profile in PNG format. In Figure 3, the preprocessing data section will go through several stages: image collection, cropping process, the object conversion process to grayscale, resizing process, and GLCM extraction results.

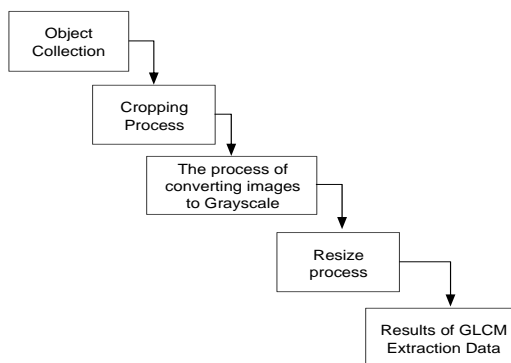


Figure 3. Data Preprocessing Stages

### Extraction Features

Feature extraction is retrieving content or content to simplify the number of resources required [12]. Feature extraction used in this study uses texture feature extraction using the Gray Level Cooccurrence Matrix (GLCM) algorithm. Textural feature extraction contains information about the structural arrangement of the surface and its relationship to its environment [13]. A total of five texture-based feature parameters used in this study are contrast, correlation, energy or angular second moment (ASM), homogeneity, and dissimilarity [14].

Energy or so-called angular second moment (ASM) is the same condition with a little grayscale value but high pixels (Equation 1). The experimental results can be seen in Table 4 below.

$$ASM = \sum_i \sum_j p^{2[i,j]} \tag{1}$$

Contrast (CON) is useful for measuring the spatial frequency of the image and the difference of several moments from the GLCM (Equation 2).

$$CON = \sum_i \sum_j (i - j)^2 P[i, j] \quad (2)$$

Correlation (COR) measures the linear dependence of the gray color on the image (can be calculated by Equation 3).

$$COR = \frac{\sum_i \sum_j (i, j) \cdot p(i, j) - \mu_x \mu_y}{\sigma_x \sigma_y} \quad (3)$$

Homogeneity or what is also known as inverse different moment (IDM). This describes the similarity of pixels. The homogeneous image GLCM matrix gives a value of 1. Very low if the image texture requires minimal modification (see Equation 4).

$$IDM = \sum_i \sum_j \frac{p[i, j]}{1 + |i - j|} \quad (4)$$

Although features are important in the data extraction process, several problems result from the magnitude of this feature extraction process, namely computational problems and complexity, a relatively large amount of waste of storage memory, so that the classification process becomes more complicated [15]. In addition, too many features in the extraction process cause over-fitting problems, leading to poor accuracy as many are removed [11].

### Object Classification

In this study, the method of classification using the K-Nearest Neighbors algorithm. This classification algorithm is based on the class of the nearest neighbors. This classification makes it possible to take more than one neighbor. This technique is called the K-Nearest Neighbors (K-NN) classification [16]. Through a more modern approach, K-NN classification looks for a group of k objects in the training set that is closest to the object being tested and is based on the class dominance in its environment [17]. This approach has three key elements: a group of labeled objects, a distance or mathematical equation for calculating the distance between objects, and k (a value that indicates the number of neighbors).

The K-NN algorithm works on the shortest distance from the query instance to the training sample to determine its neighbors [18]. The stages are carried out using the following steps [19]:

- a. Specifies the k parameter.
- b. Calculate the distance between the data to be evaluated with all training.
- c. Sort the distances formed.
- d. Determines the closest distance to the order k.
- e. Pair the appropriate class.
- f. Find the number of classes from the closest neighbors and assign the class as the data class to be evaluated (Equation 5).

$$d_i = \sqrt{\sum_{i=1}^p (X_{2i} - X_{1i})^2} \quad (5)$$

Description:

$X_1$  = Sample Data

$X_2$  = Test Data

$i$  = Data Variables

$d$  = Distance

$p$  = Data Dimensions

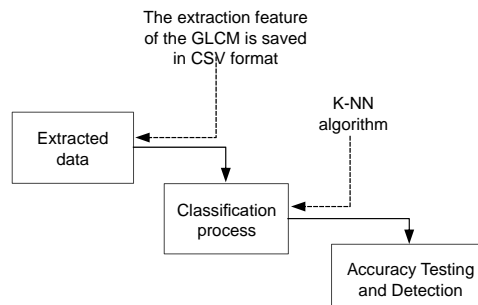


Figure 4. The classification process using the K-NN algorithm

### RESULT AND DISCUSSION

The original image data in data collection for identification will be resized using a size of 250 \* 230 pixels. The process is carried out as shown in Figure 2 previously, namely through image processing and feature

extraction using python 3.8 programmings and Jupiter Notebook IDE. The devices used for the configuration in this study use Windows 10 64bit, Intel i5 processor with 2.3 GHz and 8GB RAM.

The training data used in the identification process amounted to 15 types of herbal plant leaves with 1 type of leaf successfully collected as many as 30 objects, so that the total was 450 leaf samples. The GLCM algorithm is used to find the matrix pixel value, which has a value, and the process in the image is calculated based on the object's property as a feature. The data from the GLCM extraction were then stored in comma-separated values (CSV) format for further analysis. From the leaf data that has been collected as many as 450 samples are then only used as many as 66 leaf data in GLCM. In table 1, an example of the extraction process results from the GLCM algorithm on Binahong Merah leaves.

Table 1. Examples of Red Binahong Leaf Extraction Results with the GLCM Algorithm Based on Gray Degrees

GLCM parameters	Degree			
	0	45	90	135
Contrast	3.993.377.557	4.126.724.361	3.198.491.108	3.663.198.395
Energy	0.514846019	0.512931615	0.518713389	0.513496657
Correlation	0.960331342	0.959037191	0.968212115	0.963638256
Homogeneity	0.56486454	0.563284384	0.57337698	0.564946244
Dissimilarity	7.635.684.399	7.749.328.027	7.683.683.421	7.404.729.906

Based on the data that has been extracted with the previous GLCM algorithm, the following process is to test the accuracy of the K value, which is to see the lowest value or the value that has the most minor error. Figure 5 shows the K value error rate or prediction error from the previous herbal plant dataset.

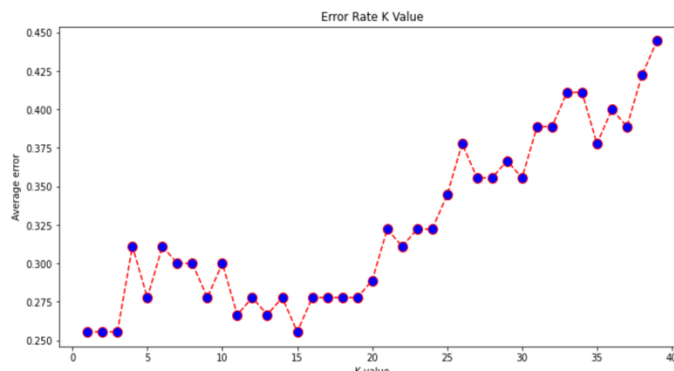


Figure 5. The results of the K-Value Error Rate analysis

Table 2. Accuracy Results

Name	Precision	Recall	f1-score	Support
Red Spinach Leaf	0.50	0.71	0.59	7
red binahong Leaf	1.00	0.88	0.93	8
insulin Leaf	0.50	0.50	0.50	4
guava Leaf	0.00	0.00	0.00	3
Castor Leaf	0.57	0.80	0.67	5
Sauropus androgynus	1.00	1.00	1.00	4
Hibiscus Leaf	0.67	0.67	0.67	6
Orthosiphon aristatus	0.80	1.00	0.89	4
Bowl Leaf	0.83	1.00	0.91	5
Bay leaf	0.73	0.89	0.80	9
Cherry Leaf	0.80	0.57	0.67	7
Betel leaf	1.00	0.75	0.86	8
Soursop leaf	1.00	0.17	0.29	6
Madagascar Periwinkle	1.00	1.00	1.00	7
False daisy	0.75	0.86	0.80	7
<b>Accuracy</b>		<b>0.74</b>		<b>90</b>

Furthermore, to see the testing of each leaf of a herbal plant, a classification was carried out to see the success of the leaf identification process. Table 2 shows that almost all the types of samples used in the leaf identification process have been detected well with good lighting conditions as well, but when the lighting conditions are less good (darker image conditions), the detection process is increasingly difficult to recognize. From the classification model carried out by machine learning, several types of leaves such as Sauropus androgynous and Madagascar Periwinkle get precision, recall, and f1-scores with a value of 100 %. While Guava Leaf can not be detected properly, this is likely the impact of the quality of the images taken are not good.

Furthermore, the leaf pattern identification testing process is also carried out through the condition of the image results obtained based on the energy and contrast parameters of the object, namely sufficient light and low light conditions. This test was carried out using four leaf samples, where for each type of leaf data there were three samples or classes used, namely three samples with sufficient light and three samples with low light. Figure 6 shows the relationship and predictions of the leaf type classification process. For example, the prediction results of the Red Spinach Leaf have a match with the Hibiscus Leaf.

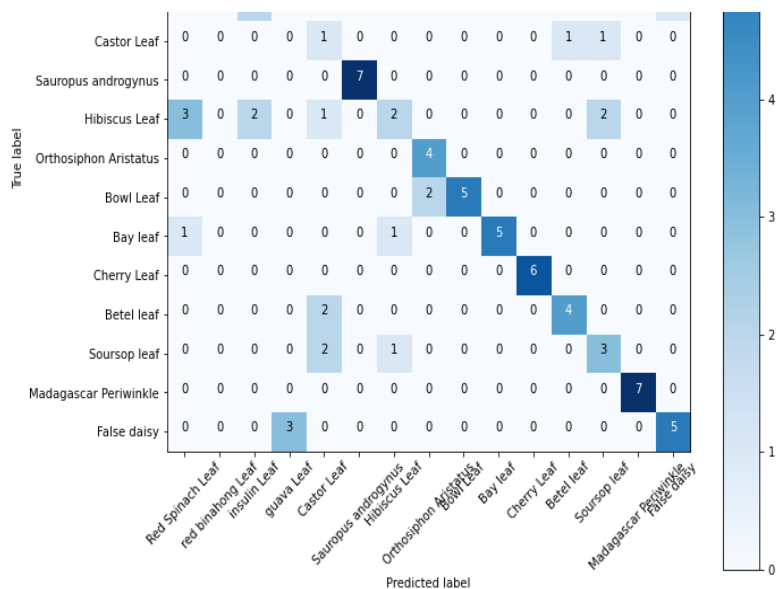


Figure 6. Label Correlation for Prediction

From image objects that failed to be recognized due to several factors, namely the conditions for taking the objects that were not right and the distance was too far, and the number of recognition classes available in the dataset. The processing results are obtained through the formula in Equations 1 to 4 above. For that, it is necessary to take objects for the correct dataset and increase the number of detection object classes better by paying attention to the right parameters such as lighting composition and contrast of the captured objects.

## CONCLUSION

Herbal plants are currently starting to be seen by the public as alternative herbal medicinal ingredients. However, if there is an understanding and mistakes in identifying the types of plants to be used, it will have a fatal impact on health. Based on the objectives of this study discussed previously, the process of identifying and predicting leaf patterns of herbal plants can be applied using the K-NN classification algorithm combined with GLCM. Based on the results of processing and testing data, it was found that the level of accuracy obtained by this method was 74 % of the total 15 samples of plant species used. This is of course, still very little for the population of herbal plants in Indonesia. For further research, we will develop a mobile-based application to more easily assist herbal plants' identification and detection process. In addition, we will improve the number of samples of the plant dataset used. For this reason, a database is needed that can store a collection of datasets so that the detection process is better, and deep learning-based learning is required at this step.

## REFERENCES

- [1] F. S. Ni'mah and T. Sutojo, "Identifikasi tumbuhan obat herbal berdasarkan citra daun menggunakan algoritma gray level co-occurrence matrix dan k-nearest neighbor," *Jurnal Teknologi dan Sistem Komputer*, vol. 6, no. 2, pp. 51–56, 2018.
- [2] M. E. Chowdhury *et al.*, "Automatic and reliable leaf disease detection using deep learning techniques," *AgriEngineering*, vol. 3, no. 2, pp. 294–312, 2021.
- [3] S. Nandhini, S. Parthasarathy, A. Bharadwaj, and K. H. Vardhan, "Analysis on classification and prediction of leaf disease using deep neural network and image segmentation technique," *Ann. Rom. Soc. Cell Biol.*, vol. 25, no. 6, pp. 9035–9041, 2021.
- [4] G. Saleem, M. Akhtar, N. Ahmed, and W. S. Qureshi, "Automated analysis of visual leaf shape features for plant classification," *Comput. Electron. Agric.*, vol. 157, pp. 270–280, Feb. 2019.
- [5] R. R. Waliyansyah, "Identifikasi jenis biji kedelai (*glycine max l*) menggunakan gray level coocurance matrix (glcm) dan k-means clustering," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 7, no. 1, pp. 17–26, 2020.
- [6] S. G. Wu, F. S. Bao, E. Y. Xu, Y.-X. Wang, Y.-F. Chang, and Q.-L. Xiang, "A leaf recognition algorithm for plant classification using probabilistic neural network," in *2007 IEEE Int. Symp. Signal Process. Inf. Technol.*, Giza, Egypt, Dec. 2007, pp. 11–16.
- [7] S. Sabzi, R. Pourdarbani, and J. I. Arribas, "A computer vision system for the automatic classification of five varieties of tree leaf images," *Computers*, vol. 9, no. 1, p. 6, 2020.
- [8] K. Yang, W. Zhong, and F. Li, "Leaf segmentation and classification with a complicated background using deep learning," *Agronomy*, vol. 10, no. 11, p. 1721, 2020.
- [9] R. J. Khanjar, "DNN based plant diseases recognition using classification of leaf images," *Medico Legal Update*, vol. 20, no. 4, pp. 1933–1942, 2020.
- [10] K. R. Castleman, "Digital image processing," *Pearson Education, Canada*, 1993.
- [11] D. Kumar, "Feature extraction and selection of kidney ultrasound images using GLCM and PCA," *Procedia Comput. Sci.*, vol. 167, pp. 1722–1731, 2020.
- [12] P. Mohanaiah, P. Sathyanarayana, and L. GuruKumar, "Image texture feature extraction using GLCM approach," *Int. J. Sci. Res. Publ.s*, vol. 3, no. 5, pp. 1–5, 2013.
- [13] R. M. Haralick, K. Shanmugam, and I. H. Dinstein, "Textural features for image classification," *IEEE Trans. Syst. Man Cybern.*, no. 6, pp. 610–621, 1973.
- [14] G. Saleem, M. Akhtar, N. Ahmed, and W. Qureshi, "Automated analysis of visual leaf shape features for plant classification," *Comput. Electron. Agric.*, vol. 157, pp. 270–280, 2019.
- [15] R. T. Collins, Y. Liu, and M. Leordeanu, "Online selection of discriminative tracking features," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 10, pp. 1631–1643, 2005.
- [16] A. Budianto, R. Ariyuana, and D. Maryono, "Perbandingan k-nearest neighbor (k-nn) dan support vector machine (svm) dalam pengenalan karakter plat kendaraan bermotor," *J. Ilm. Pendidik. Tek. dan Kejuru.*, vol. 11, no. 1, pp. 27–35, 2018.
- [17] Vinita Chandani, Romi Satria Wahono, and Purwanto Purwanto, "Komparasi algoritma klasifikasi machine learning dan feature selection pada analisis sentimen review film," *J. Intell. Syst.*, vol. 1, no. 1, pp. 56–60, 2015.
- [18] M. Lestari, "Penerapan algoritma klasifikasi Nearest Neighbor (K-NN) untuk mendeteksi penyakit jantung," *Faktor Exacta*, vol. 7, no. 4, pp. 366–371, 2015.
- [19] L. A. R. Hakim, A. A. Rizal, and D. Ratnasari, "Aplikasi prediksi kelulusan mahasiswa berbasis k-nearest neighbor (k-nn)," *JTIM: J. Teknol. Inf. Dan Multimed.*, vol. 1, no. 1, pp. 30–36, 2019.