IDENTIFICATION OF HERBAL PLANT BASED ON LEAF IMAGE USING GLCM FEATURE AND K-MEANS

Recha Abriana Anggraini[1]; Fanny Fatma Wati[2]; Muhammad Ja'far Shidiq[3]; Ade Suryadi[4]; Haerul Fatah[5]; Desiana Nur Kholifah[6]

[1,2,4,5]Information Systems, [6]Software Engineering
Universitas Bina Sarana Informatika
www.bsi.ac.id
[1]recha.rcb@bsi.ac.id, [2]fanny.ffw@bsi.ac.id, [4]ade.axd@bsi.ac.id, [5]haerul.hef@bsi.ac.id, desiana.dfh@bsi.ac.id

[3]Computer Science
STMIK Nusa Mandiri, Jakarta, Indonesia
www.nusamandiri.ac.id
ash.shidiq.mj@gmail.com

Abstract—*Medicinal plants are one of the groups of plants that have enormous benefits for humans because they can help the medical process for healing disease. Herbal plants can be used as ingredients for medicines, medicines produced from herbal plants are also natural. Lack of knowledge of herbal plants causes people to prefer chemical-based medicines to help cure their diseases, even though chemical-based drugs have side effects on human health. This study aims to identify types of herbal plants based on the extraction of contrast, correlation, energy, and homogeneity features as well as shape recognition based on metric and eccentricity values. The method used in this research is GLCM features and K-means clustering. In this study, the data used consisted of 352 data divided into 320 training data and 32 testing data. This research succeeded in identifying and classifying herbal plant species using GLCM features and K-means clustering segmentation with an average accuracy value of 85.94%.*

Keywords*:* Leaf Image, GLCM, Herbal Plants, K-Means Clustering, Identification

Abstrak—*Tumbuhan obat adalah salah satu kelompok tumbuhan yang manfaatnya sangat besar bagi manusia karena dapat membantu para medis untuk proses penyembuhan penyakit. Tumbuhan herbal dapat dimanfaatkan sebagai bahan pembuat obat, obat yang dihasilkan dari tumbuhan herbal juga bersifat alami. Kurangnya pengetahuan akan tumbuhan herbal menyebabkan masyarakat cenderung lebih memilih obat berbahan kimia untuk membantu penyembuhan penyakitnya, padahal obat berbahan kimia memiliki efek samping bagi kesehatan manusia. Penelitian ini bertujuan untuk mengidentifikasi jenis tumbuhan herbal berdasarkan ekstraksi fitur kontras, korelasi, energi, dan homogenitas serta pengenalan bentuk berdasarkan nilai metric dan eccentricity. Metode yang digunakan dalam penelitian ini adalah GLCM fitur dan K-means clustering. Dalam penelitian ini, data yang digunakan terdiri dari 352 data yang dibagi menjadi 320 data training dan 32 data testing. Penelitian ini berhasil mengidentifikasi dan mengklasifikasikan jenis tumbuhan herbal menggunakan GLCM fitur dan segmentasi K-means clustering dengan nilai rata-rata akurasi sebesar 85,94%.*

Kata kunci*: Gambar Daun, GLCM, Tanaman Herbal, K-Means Clustering, Identifikasi*

INTRODUCTION

Plants are the most important part of life on earth. Plants are useful as a supplier of oxygen for breathing, as food, fuel, medicines, cosmetics, and much more. The process of grouping plants can be done by identifying the leaf shape image from the plant itself. The process of identifying the leaf image of the plant can be done by recognizing the leaf pattern by recognizing the structural characteristics of the leaf such as the shape and texture of the leaf (Wu et al., 2007)(Chaki & Parekh, 2011)

Medicinal plants or commonly known as herbal plants are one group of plants that are of enormous benefit to humans. This plant can be used as an ingredient in medicine to help cure various diseases in the medical world. Besides, drugs produced from herbal plants are also safer to consume than drugs made from chemicals. However, because of the many types of herbal plants and limited information (Hidanti et al., 2016) and the knowledge of herbal plants that cause disease treatment choices always fall on drugs that contain chemicals (Hidanti et al., 2016) whereas chemicals are inorganic and pure,

whereas the human body is organic and complex so chemical drugs are often not suitable and effective (Ni'mah et al., 2018a) to cure certain diseases even some chemical drugs are only symptomatic or temporary and must be taken for life by the patient (Ni'mah et al., 2018b). To provide information to the public as well as medical personnel about herbal plants, we need a system that can identify and recognize herbal plants based on one part of the plant. One of the identification processes can be done by analyzing digital images in the form of leaf images of these herbal plants and making recognition of a pattern or its characteristics.

Research on the identification of an image has been developed, one of them by differentiating the texture in the image. Image textures can be distinguished based on density, regularity, uniformity, and roughness (Ganis et al., n.d.). Because computers cannot distinguish textures like human vision, texture analysis is used to determine patterns of digital images. Analysis of the texture will produce a value of the texture characteristics or characteristics which can then be processed by the computer for the classification process (Purnamasari & Sutojo, 2017). One method of texture analysis that can be used is the Gray Level Co-occurrence Matrix (GLCM).

This study aims to identify herbal plants by analyzing the image of the shape and texture of the leaves of these herbal plants. Image texture can be distinguished by uniformity, hardness, density, and regularity. To distinguish the texture of the image used texture analysis to determine the value of the characteristics or characteristics of the texture which is then processed by the computer for the segmentation process to determine the type of plant. The method used in this study is the GLCM (Gray Level Co-Occurrence Matrix) and K-Means Clustering method.

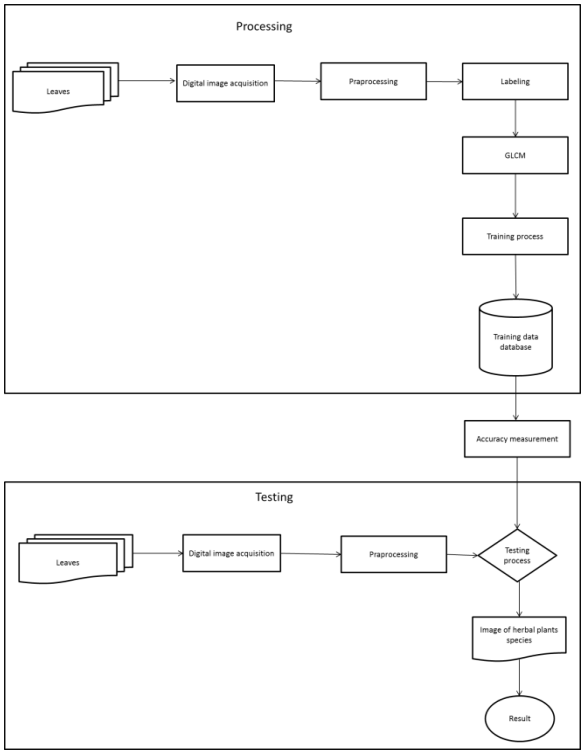
Research on the identification of plants using leaf images has been widely developed. Ni'mah in 2018 (Ni'mah et al., 2018b) conduct research with the identification of herbal medicinal plants based on leaf images using the gray level co-occurrence matrix and K-nearest neighbor algorithm. The object of this research is to classify herbal medicinal plants based on leaf images obtained using GLCM and KNN feature extraction. The results of identification using 9-fold cross-validation showed an accuracy value of 83.33% (Ni'mah et al., 2018b). Then in the same year, Rahmasari et al also identified plant species based on the Artificial Neural Network (leaf) image (Rahmadewi et al., 2018). In 2015 Liantoni and Nugroho conducted a classification of herbal leaves using the Naive Bayes and KNN methods, this study also used leaf images as the object of study (Liantoni & Nugroho, 2015). In 2018 Auliasri and Kertaningtyas conducted a study with the title of a comparative study of the classification of digital image texture patterns using the k-means and Naïve Bayes methods of lung images (Auliasari & Kertaningtyas, 2018). Based on some previous studies it can be concluded that the most frequently used method for identifying or classifying an image is KNN and Naive Bayes. Therefore, in this study, the researchers tried to apply the extraction of image feature values using GLCM and K-means clustering segmentation as a basis for the process of image identification.

MATERIALS AND METHODS

A. System Design

Several steps are needed to identify the image of herbal plants. To carry out image processing, the first step is image acquisition to produce training data and then enter the processing stage. The next step is to label the training data. Labeling is done following the types of herbal medicinal plants, amounting to 32 types. The processing performed using the GLCM algorithm will produce 4 features of the extraction results which are then stored in a database namely Microsoft Excel.

The system design for analyzing the texture of herbal plants based on leaf images is:



Source: (Anggraini et al., 2020)

Figure 1. System design training and testing process.

Based on Figure 1 it can be seen that in the testing process the image acquisition process is also carried out in the form of data testing and preprocessing. In the testing process, the testing data will be extracted using the GLCM feature and the image segmentation process is carried out using K-means clustering where the testing data will be processed and the extraction values closest to the data in the training data will be searched. This introduction phase will produce output in the form of image results.

B. Training Process

1. Digital Image Acquisition

Digital image acquisition aims to determine the data needed and choose digital image recording methods. In this stage, the researchers used a digital image recording method by searching for herbal plant image data through the Google search engine.

2. Praprocessing

The preprocessing phase aims to simplify the process of image identification. This stage consists of changing the pixel size of the original image to 688x800 pixels, changing the background color of the image in the segmentation process and changing the color of the RGB image to grayscale, LAB, and binary image to get the extraction of the shape value from the image.

3. Labeling

At this stage labeling of each image in the training, data is done. Labeling aims to separate data based on labels that will be used in segmentation and classification.

4. GLCM

At this stage, texture analysis is performed using the GLCM feature. This process is related to the quantization of image characteristics into a group of corresponding characteristic values. Texture analysis is generally used as an intermediary process for image classification and interpretation. The extracted features are contrast, correlation, energy, and homogeneity.

5. Training

At this stage, the training process is carried out using a set of training data that contains parameter features or features that are used to differentiate between one object and another object. The characteristics used are texture analysis with GLCM and leaf shape recognition. The training process maps training data towards the training target through an algorithm formulation used. Image identification using GLCM features and image classification using k-means clustering segmentation. The distance used is the euclidean distance (Sutojo et al., 2018).

6. Database data training

The training data database contains leaf image information that has passed the preprocessing, labeling, and training stages using texture analysis using the GLCM feature. The results of the training and labeling process are stored in Microsoft Excel and entered into the image processing application database. This aims to make the testing process easier and faster because the results of the extraction are already in the database without having to wait for the training process.

C. The Testing Process

The testing process is done after the training, with testing data entered into the application that has been designed and tested whether it is in accordance with the target or not. Testing data is used as an object to test methods in image processing. The testing data used were 32 types of herbs. The texture analysis stage is performed using the GLCM feature to quantify the image characteristics into a group of corresponding feature values. Characteristics of the features used in this study are contrast, correlation, energy, and homogeneity (Jundullah & Syahrul Mubarok, 2016). The classification process is done using the segmentation of the k-means clustering algorithm. Testing data that has been analyzed by GLCM and recognition of image shapes are classified based on the segmentation and cluster of each image. Testing data that has been analyzed by GLCM are classified based on the results of image segmentation and by calculating the closest distance from the training data and the image that has been trained using euclidean distance with equation 1.

$$D(a,b) = \sqrt{\sum_{i=1}^{n} b_i - a_i}^2 \quad \text{..... (1)}$$

Applications that have been made with MatLab will be tested by calculating the level of accuracy in recognizing or identifying herbal plants. Accuracy testing is done by comparing the real-time image of herbal plant leaves from the database of the testing or training stage and the results of leaf recognition with GLCM at the training stage. The process of testing or training on images is done using the k-means clustering segmentation method.

Source: (Goëau et al., 2013)
Figure 2. Example of Image Data

Figure 2 shows one of the image data to be tested to determine the species classification based on the characteristics or characteristics of the image.

B. Measurement Accuracy

The testing phase to validate the accuracy of the identification of the herbal leaf image is done by the k-means clustering segmentation method. The dataset was divided, which initially amounted to 352 data into two parts, namely training data and testing data. Training data are 320 data and testing data are 32 data. To find the accuracy value, use equation 2.

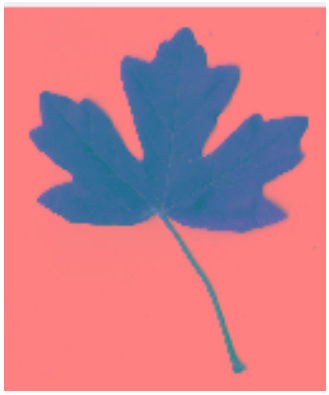
$$Si = \frac{\sum seluruh\ data\ benar}{\sum seluruh\ data\ salah} \times 100\% \quad \text{..... (2)}$$

RESULTS AND DISCUSSION

The data used in this research process is data taken from the website https://www.imageclef.org/2013/plant with the title ImageCLEF2013PlantTaskTestAndTaskPackage. This dataset contains data of herbal plants originating from France consisting of flowers, leaves, fruits, stems, and the overall image of the tree. The dataset contains 26077 images of 250 plants. Then filtering the data is carried out, because this research focuses on the processing of leaf images so the filtering process is carried out by searching for leaf images in the dataset. Leaves filtered data amounted to 352 data. From the data, it is divided into training and testing data with a total of 320 training data and 32 testing data.

A. Praprocessing

The original image in this study has a white background and different pixel values. At this stage the process of changing the pixel size on the entire image to 688x800 pixels. After changing the pixel, convert the image from RGB to LAB and grayscale and change the background color of the image to black.



Source: (Anggraini et al., 2020)
Figure 3. Image of LAB

Source: (Anggraini et al., 2020)
Figure 4. Grayscale image

Figure 3 shows the results of converting the original image to the LAB image. This conversion is done to simplify the process of image segmentation. Figure 4 shows the results of the conversion of the original image to grayscale. This conversion is done by taking all the color pixels in the image.

B. Labeling

The labeling function is to separate image data based on the label. Besides, this labeling also functions for grouping. In this study, labeling is done by classifying classes from training data or training images that have been separated through the previous process. After the data is obtained and has gone through the preprocessing stage, the data is labeled or class.

C. Extraction with GLCM and shape recognition

The GLCM process and the recognition of image shapes are carried out by looking at the feature extraction values in the image. Extraction

of dialed texture features by forming a co-occurrence matrix. This matrix is formed from an image by looking at the relationship between the two pixels at a certain distance and orientation. This matrix is used to extract texture features from an image. The distance used in this study is d = 1, while the angles using 00,450,900,1350 are averaged over each feature in each image. There are four texture features used in this study, namely contrast, correlation, energy, and homogeneity. As for the introduction of mind shape, two features are extracted, namely metric and eccentricity. This process is carried out in training data and testing data. The results of the training data feature extraction are stored in the training database and outlined in Table 1.

Table 1. Results of Data Training Feature Extraction

Metric	Eccentricity	Contrast	Correlation	Energy	Homogenity
0,22968	0,70204	0,10884	0,85798	0,58982	0,98517
0,3577	0,64679	0,11894	0,92679	0,54651	0,97646
0,4122	0,61491	0,082777	0,87229	0,67198	0,98522
...	...	...	...	...	...
0,32583	0,65561	0,087531	0,97629	0,6025	0,98176
0,12548	0,75306	0,14839	0,77095	0,65988	0,98399
0,18075	0,75976	0,087253	0,86181	0,62314	0,98916
...	...	...	...	...	...
0,50997	0,37769	0,1146	0,91842	0,43491	0,96446
0,52412	0,62132	0,12419	0,90134	0,49334	0,96643
0,60766	0,47859	0,087814	0,91391	0,52877	0,98729
...	...	...	...	...	...
0,48466	0,6206	0,042271	0,99194	0,65483	0,99369
0,17364	0,50058	0,1152	0,84397	0,58361	0,97958
0,17364	0,50058	0,16861	0,97817	0,56232	0,97774
...	...	...	...	...	...
0,72038	0,51016	0,068599	0,99155	0,57594	0,9918
0,36483	0,85876	0,042507	0,94278	0,73035	0,98751
0,67852	0,81573	0,087298	0,97495	0,74866	0,97756
...	...	...	...	...	...
0,12198	0,31909	0,11146	0,97452	0,65306	0,9633
0,67725	0,63192	0,12233	0,91801	0,42871	0,96589
0,52133	0,62435	0,077285	0,89048	0,70729	0,97138

Source: (Anggraini et al., 2020)

Table 1 shows the results of feature extraction from GLCM and shape recognition. The results are used as a data source for the image testing process through an application created using MatLab.

Source: (Anggraini et al., 2020)

Figure 5. Image Segmentation with K-means

Figure 5 is an image display generated after going through the segmentation stage using k-means clustering in the application made.

Source: (Anggraini et al., 2020)

Figure 6. Binary Image

Figure 6 is a display of binary images which is part of the process of recognizing image shapes. From the process of changing the RGB image into a

binary image, it can be seen the value of metric and eccentricity that can be taken to elude the shape of the leaf.

D. Accuracy Results

System accuracy measurement uses the image segmentation process and analysis of shapes and textures. The image testing process is carried out on 32 types of images obtained from the dataset.

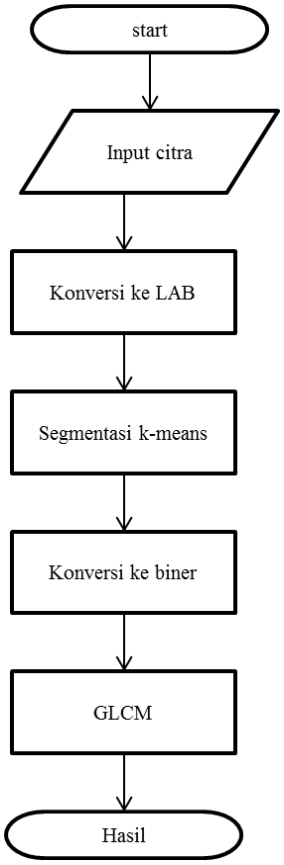
Table 2. Image Accuracy Values

LABEL	ACCURACY
Acer campestre	100
Acer monspessulanum	100
Alnus glutinosa	50
Buxus sempervirens	100
Carpinus betulus	100
Celtis australis	100
Cercis siliquastrum	100
Cornus mas	100
Cornus sanguinea	50
Corylus avellana	100
Eriobotrya japonica	100
Euphorbia characias	100
Hedera helix	100
Ilex aquifolium	50
Liquidambar styraciflua	100
Nerium oleander	50
Olea europaea	100
Paliurus spina-christi	100
Pittosporum tenuifolium	50
Pittosporum tobira	100
Platanus x hispanica	100
Populus nigra	100
Prunus dulcis	100
Punica granatum	50
Quercus petraea	50
Quercus pubescens	100
Quercus rubra	100
Rhamnus alaternus	50
Ruscus aculeatus	100
Ulmus minor	100
Viburnum opulus	100
Viburnum tinus	50
Average	85,94%

Source: (Anggraini et al., 2020)

Table 2 shows the results of the accuracy of each image type that has passed the testing phase using an application created with MatLab. The average accuracy is obtained by adding up the accuracy value of each image then dividing it by the number of image types, which is 32 image types.

The test scenario is done by inputting the image which will divert to the application then the process of converting the image into LAB, image segmentation, shape recognition, texture recognition then image type classification to connect the types of images of herbal planting. After the image is identified, the assessment process is carried out by calculating the results of calculations using calculation 2. The following is a flowchart of testing and accuracy testing performed:



Source: (Anggraini et al., 2020)

Figure 7. Testing Process Flowchart

Figure 7 shows the sequence of testing and identification of herbal plants in this study. The flowchart shows the process that must be carried out in running applications made with MatLab.

CONCLUSION

This research has succeeded in implementing the GLCM feature and k-means clustering algorithm to extract contrast, correlation, energy, and homogeneity features as well as shape recognition by extracting metric and eccentricity features on the image of herbal plant leaves and classifying them based on the closest distance between the training image and the testing image. The algorithm can be used to identify herbal plants with an average accuracy of 85.94%.

REFERENCE

Anggraini, R. A., Wati, F. F., Shidiq, M. J., Suryadi, A., Fatah, H., & Kholifah, D. N. (2020). *IDENTIFIKASI TUMBUHAN BERDASARKAN CITRA DAUN MENGGUNAKAN GLCM FEATURE DAN K-MEANS*.

Auliasari, K., & Kertaningtyas, M. (2018). Studi Komparasi Klasifikasi Pola Tekstur Citra Digital Menggunakan Metode K-Means Dan Naïve Bayes. *Jurnal Informatika*, *18*(2), 175–185.

Chaki, J., & Parekh, R. (2011). Plant leaf recognition using shape based features and neural network classifiers. *International Journal of Advanced Computer Science and Applications*, *2*(10), 41–47.

Ganis, K., Santoso, I., & Isnanto, R. R. (n.d.). *Klasifikasi Citra Dengan Matriks Ko-Okurensi Aras Keabuan (Gray Level Co-Occurrence Matrix-GLCM) Pada Lima Kelas Biji-Bijian*. Universitas Diponegoro.

Goëau, H., Joly, A., Bonnet, P., Bakic, V., Barthélémy, D., Boujemaa, N., & Molino, J. F. (2013). The ImageCLEF plant identification task 2013. *MAED 2013 - Proceedings of the 2nd ACM International Workshop on Multimedia Analysis for Ecological Data*, 23–28. https://doi.org/10.1145/2509896.2509902

Hidanti, M., Zahra, A. A., & Isnanto, R. R. (2016). SISTEM IDENTIFIKASI JENIS TANAMAN OBAT MENGGUNAKAN MATRIKS KOOKURENSI ARAS KEABUAN (GLCM) DAN JARAK CANBERRA. *FORTEI 2016*.

Jundullah, A., & Syahrul Mubarok, M. (2016). Analisis dan Implementasi Deteksi Citra Spam Menggunakan Gray Level Co-occurences Matrix dan Naive Bayes. *Indonesia Symposium on Computing (IndoSC) 2016*, 319–334. https://doi.org/10.21108/indosc.2016.164

Liantoni, F., & Nugroho, H. (2015). KLASIFIKASI DAUN HERBAL MENGGUNAKAN METODE NAÏVE BAYES CLASSIFIER DAN KNEAREST NEIGHBOR | Liantoni | Jurnal Simantec. *Jurnal Simantec*, *5*(1), 9–16.

Ni'mah, F. S., Sutojo, T., & Setiadi, D. R. I. M. (2018a). Identifikasi Tumbuhan Obat Herbal Berdasarkan Citra Daun Menggunakan Algoritma Gray Level Co-occurence Matrix dan K-Nearest Neighbor. *Jurnal Teknologi Dan Sistem Komputer*, *6*(2), 51. https://doi.org/10.14710/jtsiskom.6.2.2018.51-56

Ni'mah, F. S., Sutojo, T., & Setiadi, D. R. I. M. (2018b). Identifikasi Tumbuhan Obat Herbal Berdasarkan Citra Daun Menggunakan Algoritma Gray Level Co-occurence Matrix dan K-Nearest Neighbor. *Jurnal Teknologi Dan Sistem Komputer*, *6*(2), 51–56. https://doi.org/10.14710/jtsiskom.6.2.2018.51-56

Purnamasari, I., & Sutojo, T. (2017). PENGENALAN CIRI GARIS TELAPAK TANGAN MENGGUNAKAN EKSTRAKSI FITUR (GLCM) DAN METODE K-NN. *Jurnal VOI (Voice Of Informatics)*, *6*(1), 32–41.

Rahmadewi, R., Purwanti, E., & Efelina, V. (2018). Identifikasi Jenis Tumbuhan Menggunakan Citra Daun Berbasis Jaringan Saraf Tiruan Artificial Neural Networks. *Jurnal Media Elektro*, *VII*(2), 38–43. https://ejurnal.undana.ac.id/jme/article/view/427

Sutojo, T., Setiadi, D. R. I. M., Tirajani, P. S., Sari, C. A., & Rachmawanto, E. H. (2018). CBIR for classification of cow types using GLCM and color features extraction. *Proceedings - 2017 2nd International Conferences on Information Technology, Information Systems and Electrical Engineering, ICITISEE 2017*, *2018-January*, 182–187. https://doi.org/10.1109/ICITISEE.2017.8285491

Wu, S. G., Bao, F. S., Xu, E. Y., Wang, Y. X., Chang, Y. F., & Xiang, Q. L. (2007). A leaf recognition algorithm for plant classification using probabilistic neural network. *ISSPIT 2007 - 2007 IEEE International Symposium on Signal*

Processing and Information Technology, 11–16.
https://doi.org/10.1109/ISSPIT.2007.4458016