Machine Learning Approach for Rhizomes Classification Based on Color

Bayu Agustian[1] Maimunah[2,*]

[1,2]*Department of Informatics Engineering, Universitas Muhammadiyah Magelang, Magelang, Indonesia.*
[*]*Corresponding author. Email: maimunah@unimma.ac.id*

ABSTRACT
Rhizome plants are often used as ingredients or traditional ingredients such as temuireng, temulawak and temumangga. Such rhizomes have the same color, shape, and smell characteristics. In this study, the types of rhizomes were classified based on their color into three classes, namely temuireng, temulawak, and temumangga. The color of the rhizome was captured using the camera. The image feature extraction process was carried out to obtain the color characteristics of the image by calculating the RGB value of the image. The obtained values were classified using several machine learning methods, namely Decision Tree, Naive Bayes, and KNN. This study used 120 data for all classes with the ratio of training data and test data was 70% and 30% respectively. In the initial stages of classification, data cleaning was carried out, before training and test data were used to create a classification model. From the classification results, the accuracy value of each method was obtained and then the method with the best accuracy can be selected. Result showed that the optimal method for classifying three types of temulawak (temuireng, temulawak and temumangga) based on RGB color features was using the KNN method with an accuracy of 87.5% that similar to previous researcher who used SVM method.

Keywords: *Rhizome, Machine Learning, Color.*

1. INTRODUCTION

Indonesia has the second largest biodiversity in the world after Brazil, with its Amazon Forest. This biodiversity including the presence of medicinal plants in high numbers. Most people in Indonesia, especially in rural areas, use traditional medicine known as jamu to treat several diseases. Jamu is a Javanese word which means traditional medicine [1]. Herbal medicine is made from natural ingredients taken from the roots, leaves, fruits and also parts of animals [2].

Research on traditional herbal medicine related to benefits, manufacturing technology, efforts to obtain quality herbal medicine, as well as components of compounds in medicinal plants. Five popular herbal medicines, i.e. jamupaitan, jamukunirasem, jamuberaskencur, jamutemulawak, and gulaasem have been studied [3]. The need for information related to herbal medicine has increased in recent years as herbal medicine has become a more economical option as alternative medicine. Previous study designed a system that has the ability to understand information related to herbal medicine and manage that information into knowledge. This system used a standard model that can represent all information about herbal medicine. The ontology is used to model the knowledge of herbal medicine information with the RDFS knowledge base [4].

Machine learning has been widely used for research on herbal medicine, e.g., rhizomes, leaves and fruits. Rhizome is an herbal ingredient found in the roots of herbal plants that have different properties. There are many types of rhizomes and most of them have the same characteristics. Because there are many types of rhizomes and have similarities, it is necessary to identify them. Ginger rhizome has been investigated to identify types of ginger which includes red ginger, emprit ginger and elephant ginger. It has also been studied using artificial neural networks with an accuracy of 60% [5]. Other types of rhizomes such as temuireng, temulawak and temumangga can also be classified based on RGB color features using the Support Vector Machine (SVM) with an accuracy of 87.5% [6]. Beside based on color features, another study used shape and size features to classify turmeric using an artificial neural network with an accuracy of 73.33% [7]. Previous study showed that the

feature of the smell or odor of the rhizomes can also be used to classify some herbal medicines, i.e. ginger, kencur, turmeric and temulawak rhizomes. The smell of rhizomes or empon-empon was obtained from the e-nose which was designed using the TGS2611, TGS813, and MQ136 sensors that were connected to the Arduino Uno. Based on the classification results, it was found that the deep neural network can classify types of empon-empon based on odors with an accuracy of 86% [8].

Research has also been carried out to analyze the influence of herbal formulation on disease using the Support Vector Machine One Versus All Recursive Feature Elimination (SVM OVA-RFE) which has succeeded in reducing the dimensions of the data from 3085 samples to 238 samples [9]. Machine learning was also used to identify the authenticity of compounds from herbal products using logistic regression. The results showed that the groups of compounds that affect the authenticity of herbal products were aliphatic CH (methyl, methylene, methylene groups), bending HCH (methyl, methylene), and unsubstituted aromatics with 100% accuracy [10]. To produce a good quality herbal medicine, the good quality herbal raw materials are also needed. Therefore, it is necessary to classify the right types of rhizomes to suit their respective properties since the rhizomes have similarities in shape, colour and smell. In this study, the classification of the rhizome types, i.e. temuireng, temulawak and temumangga was carried out using a machine learning approach.

2. METHOD

The stage of the current research includes rhizomes image acquisition, feature extraction, data training and testing process and machine learning classification as shown in Figure 1.

In this study, the image data used were temuireng,temulawak and temumangga classes which were captured using an 18MP Canon EOS 600D DLSR camera. The image retrieval process was carried out using a mini studio with a distance of 30 cm. The feature extraction process is carried out based on color features. From the resulting image data, the extraction stage is carried out to obtain the average value of r, g, b which is then normalized by transforming the values of r, g and b into values between 0 and 1. Prior to the training and testing stage, preprocessing is carried out in order to make a good quality data.

At the training and testing stage, the sample data used is 80% data as training data and 20% as testing data with a total of 120 data used for all classes. In the machine learning classification stage, several machine learning methods are applied to classify the types of rhizomes. The first time the classification is carried out using the Naive Bayes, Decision Tree and KNN methods and then an analysis will be carried out on the classification results obtained. At the result prediction stage, the final output data used is the classification result, before compared with the level of accuracy. Finally, a comparison of the accuracy results that have been obtained is carried out, hence it is known which method produces the best accuracy.

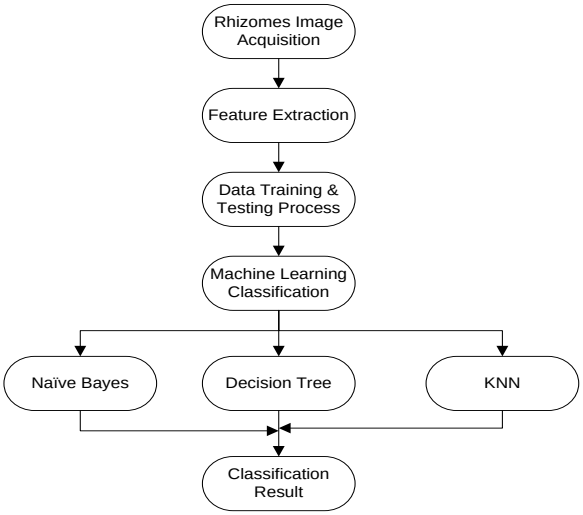

Figure 1 Detail of classification processs

3. RESULT AND DISCUSSION

There are 3 classes of image data in this study, i.e. temuireng, temulawak and temumangga. For each rhizome, 40 best images were collected. Rhizome was peeled and cut crosswise. Only the inner color of the rhizome flesh is taken for image data. The results of image data retrieval are stored in .jpeg extension format in size of 1296 x 864 pixels. Figure 2 shows the image data ready for classification.

The feature used in this study is rhizome's color. To get the color features, cropping is done from the input image. Only the flesh of the rhizome becomes the input data for classification. From this stage, RGB color features are obtained for 3 classes and 3 targets which are labeled 0, 1 and 2 where the data will be the input data for classifying. An example of the data obtained from the feature extraction stage is shown in Table 1. Label 0 in table 1 represents the temuireng, label 1 indicates the temulawak and label 2 represents temumangga. The distribution of data for each class feature r, g, b is shown in Figure 3. RGB data distribution shows the differences in the features of each class.

Figure 2 Temuireng, Temumangga, and Temulawak

Table 1. RGB Value

r	g	b	label
0.669368	0.670359	0.269494	0
0.59195	0.600923	0.247807	0
0.769081	0.388089	0.025866	1
0.815257	0.421135	0.020701	1
0.626964	0.515696	0.053322	2
0.744028	0.63336	0.049725	2

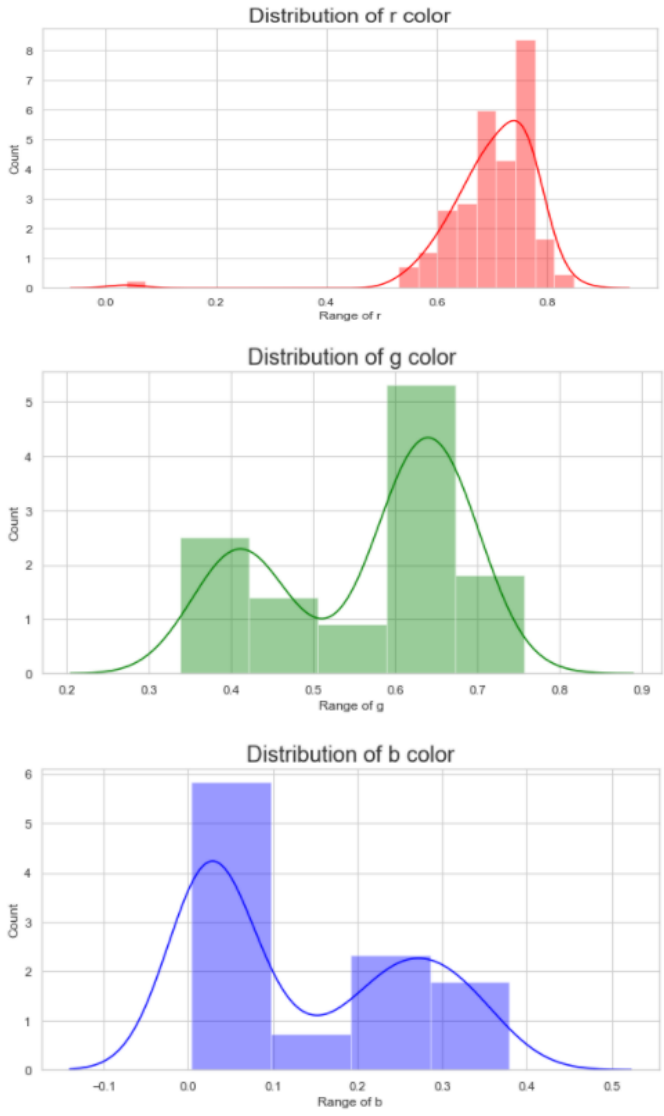

Figure 3 RGB data distribution

In the preprocessing stage, outlier data was checked using a boxplot for all classes. From Figure 4, it was found that there was only 1 outlier data, i.e. in class r, hence the outlier data was ignored.

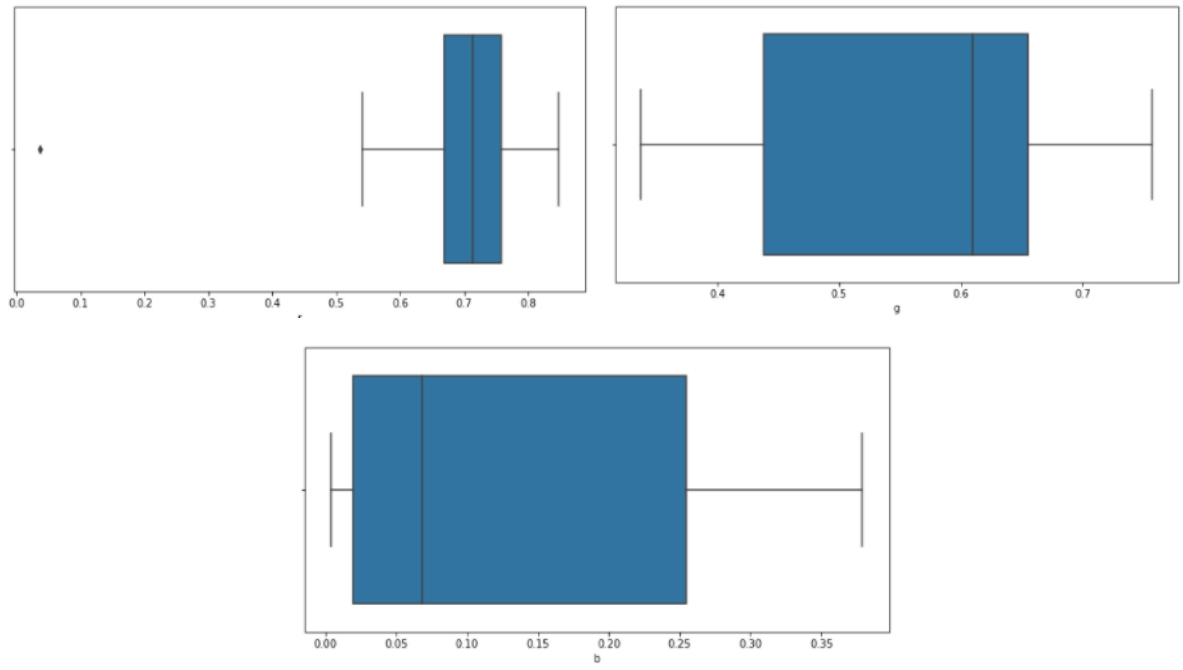

Figure 4 Boxplot

After preprocessing, the data is partitioned into training data and testing data with a composition of 80% for training and 20% for testing which is determined randomly. At the Machine Learning stage, the Decision Tree, Naive Bayes, and KNN classifications are carried out. The decision tree method is applied with the Gini and entropy parameters, a minimum sample split of 2,5,10,50 and a max depth of 3,4,5,6,7,8,9,10,11,12,13,14 and 15. Classification model of decision tree found the best parameters obtained includes the criteria of entropy, max depth and minimum sample split are 4 and 2 respectively. This decision tree model aims to determine the predictions. Table 2 shows the accuracy of predictions with confusion matrix.

Table 2 indicates that in predicting temuireng, there are 9 true results. No predictions of temulawak that actually temuireng but the predictions temumangga that actually temuireng is 2 results. There are no predictions of temuireng that actually temulawak, there are 3 accurate temulawak prediction and there are no temumangga prediction that actually temulawak. There are 2 predictions of temuireng that actually temumangga, there is no temulawak prediction result that actually temumangga and 8 true temumangga prediction. Therefore, the accuracy of predictions occurred in 20 samples, i.e. 9 true prediction of temuireng, 3 true prediction of temulawak and 8 true prediction of temumangga. Thus, the accuracy is 83%.

For classification using the Naïve Bayes method, the Gaussian function is used on the training data. The classification model created is used to find the predicted value so that the probability of the truth can be known. The accuracy of predictions is shown in the confusion matrix (Table 3). The classification results using the Naïve Bayes method is the same as the confusion matrix from the classification results using a decision tree. Thus, the accuracy obtained for the classification of rhizome types using Naïve Bayes is 83%. For classification using KNN method, it was tested with k number of 2, 3, 4, 5 and 6, using Euclidean and Manhattan distance functions through the GridSearchCV function. Table 4 shows the confusion matrix result of KNN.

Table 4 shows 11 true predictions of temuireng and there is no prediction of temulawak and temumangga that actually temuireng. There are 3 true predictions of temulawak and there is no prediction of temuireng and temumangga that actually temulawak. There are 7 true predictions of temumangga and there is no prediction of temulawak that actually temumangga but there are 3 predictions of temuireng that actually temumangga. Thus, the accuracy of using KNN obtained a value of 87.5%.

After classifying the types of rhizomes using 3 methods, i.e. Decision Tres, Naïve Bayes and KNN, it can be concluded that the best accuracy results are the classification using KNN. The results of this accuracy are also the same as those that have been carried out using the Support Vector Machine method which produces an accuracy of 87.5% [6]

Table 2. Confusion Matrix of Decision Tree

Prediction		Decision Tree		
		Temuireng	Temulawak	Temumangga
Actual	Temuireng	9	0	2
	Temulawak	0	3	0
	Temumangga	2	0	8

Table 3. Confusion Matrix of Naïve Bayes

Prediction		Naïve Bayes		
		Temuireng	Temulawak	Temumangga
Actual	Temuireng	9	0	2
	Temulawak	0	3	0
	Temumangga	2	0	8

Table 4. Confusion Matrix of KNN

Prediction		KNN		
		Temuireng	Temulawak	Temumangga
Actual	Temuireng	11	0	0
	Temulawak	0	3	0
	Temumangga	3	0	7

4. CONCLUSION

In this study, the classification of rhizomes was carried out through a machine learning approach in which three classes were used including temuireng, temulawak and temumangga. Three machine learning methods were used, i.e. decision tree, Naïve Bayes and KNN. Result showed that the optimal method for classifying three types of temulawak (temuireng, temulawak and temumangga) based on RGB color features was using the KNN method with an accuracy of 87.5% that similar to previous researcher who used SVM method. Further research is needed regarding the benefits of rhizomes as medicinal plants, for example the use of machine learning to identify all types of rhizomes in Indonesia based on odor and shape features.

REFERENCES

[1] Elfahmi, H. J. Woerdenbag, and O. Kayser, "Jamu: Indonesian traditional herbal medicine towards rational phytopharmacological use," *J. Herb. Med.*, vol. 4, no. 2, pp. 51–73, 2014.

[2] M. . Puspita, W. . Kusuma, A. Kustiyo, and R.Heryanto, "A Classification System for Jamu Efficacy Based on Formula Using Support Vector Machine and K-Means Algorithm as a Feature Selection," *ICACSIS*, pp. 215–220, 2015.

[3] W. Sumarni, S. Sudarmin, and S. S. Sumarti, "The scientification of jamu: A study of Indonesian's traditional medicine," *J. Phys. Conf. Ser.*, vol. 1321, no. 3, 2019.

[4] B. Susanto, H. Y. Situmorang, and G. Virginia, "RDFS-Based Information Representation of Indonesian Traditional Jamu," *Proc. Int. MultiConference Eng. Comput. Sci. ,IMECS 2019*, vol. 2239, pp. 352–357, 2019.

[5] Maimunah, "Identifikasi Jenis Jahe Berdasarkan Warna Menggunakan Jaringan Syaraf Tiruan," *Inf. Syst. Educ. Prof. J. Inf. Syst.*, vol. 2, no. 2, pp. 145–154, 2018.

[6] M. Maimunah and E. R. Arumi, "Use of Support Vector Machine to Classify Rhizomes Based on Color," *J. Phys. Conf. Ser.*, vol. 1381, no. 1, 2019.

[7] A. Kaur, N. Saini, R. Kaur, and A. Das, "Automatic Classification of Turmeric Rhizomes using the External Morphological Characteristics," *Int. Conf. Adv. Comput. Commun. Informatics, ICACCI 2016*, no. Sept, 21-24, 2016, pp. 1507–1510, 2016.

[8] Maimunah, M. Hanafi, and B. Agustian, "Deep neural network method to classify empon-empon herb based on e-nose," *2020 5th Int. Conf. Informatics Comput. ICIC 2020*, pp. 0–3, 2020.

[9] A. Fitriawan, I. Wasito, W. A. Kusuma, and R. Heryanto, "Support Vector Machine OVA-RFE Approach for Finding the Significant Plants of

Jamu," *2016 6th Int. Work. Comput. Sci. Eng. WCSE 2016*, no. June, pp. 650–654, 2016.

[10] B. P. R. Riau, F. M. Afendi, and R. Anisa, "Selection of Compound Group to Identify the Authenticity One of Jamu Product Using the Group Lasso for Logistic Regression," *J. Phys. Conf. Ser.*, vol. 1341, no. 9, 2019.