# Classification of Orchid Types using Random Forest Method with HOG Features

Oki Arifin[1,a)], Dewi Kania Widyawati[2], Zuriati[3], Rima Maulini[4)],
Dwirgo Sahlinal[5)], Sylvia[6)]

[1,6] Department of Software Engineering Technology, Politeknik Negeri Lampung, Bandar Lampung, Indonesia
[2,4,5]Department of Informatics Management, Politeknik Negeri Lampung, Bandar Lampung, Indonesia
[3]Department of Internet Engineering Technology, Politeknik Negeri Lampung, Bandar Lampung, Indonesia

[a)]Corresponding author: okiarifin@polinela.ac.id

**Abstract.** Orchids are one of the Indonesian people's most widely cultivated ornamental plants. Orchids are a family of plants in the Orchidaceae family that includes more than 700 genera and around 28,000 individual species. In terms of plant morphology, orchids can be distinguished based on the morphology of flowers, leaves, fruits, stems, and roots. Orchid leaves have their characteristics for each type of orchid, such as long, round, or lanceolate. All orchids have leaf veins that are parallel to the leaves. This makes it difficult to identify the type of orchid flower, especially for laypeople who are new to orchid cultivation and do not yet know the characteristics of various kinds of orchids. The individual shapes of orchid leaves can be classified using Random Forest and Histogram of Oriented Gradients (HOG). In this study, three types of orchids that are currently popular with orchid lovers were used, namely Cattleya, Phalaenopsis, and Vanda orchids taken from public data. The accuracy of this method in classifying orchid types based on leaf morphology can be measured using a confusion matrix that measures accuracy, precision, recall, and F1-score. The test results show that this method successfully achieved an accuracy of 98%, with an average precision, recall, and F1-score of 0.98 each. These findings indicate that the model built can classify orchid species with a high level of accuracy based on leaf morphology.

**Keywords:** Orchid Classification, Leaf Morphology, Random Forest, HOG

## INTRODUCTION

Orchids are one of the most widely cultivated and popular ornamental plants, especially among Indonesians. In Indonesia, there are an estimated 5000-6000 species of orchids spread throughout regions such as Kalimantan, Java, and Papua [1] [2]. Of the types of orchids spread throughout Indonesia, 29 species are included in the rare category. Therefore, orchids are included in the protected plants in Government Regulation Number 7 of 1999 concerning the Preservation of Plants and Animals.

Orchid plants have high economic value because they have special characteristics with unique lip shapes and colors or labellums that distinguish them from other plants and have aesthetic value. Orchids have the Latin name orchidaceae as one type of ornamental plant with all its stunning uniqueness that has attracted the attention of ornamental plant enthusiasts both from within and outside the country [3]. According to Pratyaswara, et al. (2017) one type of orchid flower can have characteristics in the form of different shapes and colors so that it can make it difficult for people to identify the type of orchid flower [4].

Orchid flowers have a basic structure of three sepals (petals) and three petals (flower crowns). One way to distinguish one type of orchid from another is by looking at the color, texture, and petals of the orchid flower [5]. By knowing these differences, someone can identify the type of orchid flower. However, in general, the types of orchid flowers have similarities in color, texture and petals. This is what causes someone to have difficulty in identifying the type of orchid flower, especially lay people or people who are new to the orchid cultivation business who do not yet know the characteristics of several types of orchid flowers. Therefore, effective technology is needed to classify orchid types with high accuracy and optimal time efficiency.

One of the technologies that can be utilized in orchid image classification is digital image processing that relies on the Histogram of Oriented Gradients (HOG) feature. HOG is a feature extraction technique in image processing that groups pixel gradient values according to the orientation of the direction in each local part of the image [6]. The HOG feature has been shown to be able to capture important texture and shape information in images, making it one of the popular methods in visual feature extraction for object classification [7]. In the context of plant classification, the HOG feature is able to represent variations in shape and texture well, which is essential for the identification of diverse orchid species.

The random forest method is an ensemble machine learning algorithm that has been successfully used in various classification applications [8]. According to Mekha and Teeyasuksaet (2021)  Random forest or collection of decision trees is a combined regression learning method for classification and other methods [9]. The random forest algorithm is created using many decision trees for training and displays the classification prediction results (regression) of each tree. This algorithm utilizes a collection of decision trees to produce accurate and stable predictions [10]. This method also has the advantage of reducing the overfitting problem that is common in single decision trees, and shows strong performance in classifying data with complex visual features, such as in plant images. Therefore, the combination of HOG features and the random forest algorithm is expected to improve performance in classifying orchid types accurately.

Previous research related to plant classification through leaf images has been carried out including the following: Research conducted by Meiriyama and Sudiadi (2022) on the classification of herbal leaf types using random forest based on the HOG feature. The images that have been separated between training data and test data are converted to grayscale and resized to 816x612 pixels, then the image is extracted using the HOG feature to produce a vector with a length of 1x3168. The random forest algorithm used for herbal leaf classification has an overall accuracy of 85.33% [11]. In the study Ibrahim et al. (2018) conducted a study on the classification of 10 types of herbal plant leaves using the HOG method, Local Binary Pattern (LBP), and Speeded Up Robust Features (SURF) with the classification algorithm used is Support Vector Machine (SVM) [12].

This study uses leaf data from three popular orchid species, namely Cattleya, Phalaenopsis, and Vanda, which are classified using the HOG and random forest methods. Model accuracy is evaluated using a confusion matrix that measures accuracy, precision, recall, and F1-score. This study aims to develop a classification model that combines HOG features and the random forest method in the process of identifying orchid species and is expected to provide high accuracy in identifying orchid species based on leaf morphology.

## METHODS

The following are several stages carried out to classify orchid types using random forest and HOG which can be seen in **FIGURE 1**.

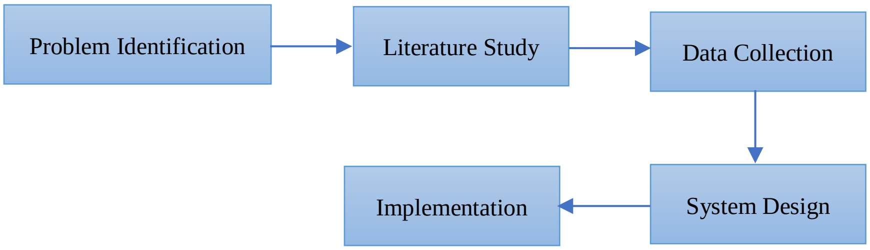

**FIGURE 1**. Research Stages

1. Problem Identification
   Defining the problem in automatic classification of orchid leaf types. The main challenge in this study is to recognize the specific characteristics of three different orchid species such as Cattleya, Phalaenopsis, and Vanda based on the visual characteristics of their leaves. At this stage, the study defines in detail the visual

characteristics of the leaves and establishes the classification boundaries needed to distinguish each type of orchid leaf.

2. Literature Study
   Literature study conducted to gather in-depth information about the methods and theories to be used. This study utilizes the random forest method as an ensemble-based machine learning algorithm that combines several decision trees to improve classification accuracy. In addition, HOG is used as a feature extraction method that captures texture and shape patterns from leaf images that are useful for distinguishing the visual characteristics of orchid leaves. The literature study also includes a literature review of the specific characteristics of orchid leaves that help understand the differences in shape, color, and texture patterns as distinguishing features in classification.

3. Data Collection
   At this stage, the data collection is taken from the public dataset on the Kaggle website (https://www.kaggle.com/datasets/raihanallaam/orchid-genus). The dataset consists of 2100 images. These images are divided into three types of orchids, namely: Cattleya, Phalaenopsis, and Vanda. Each type of orchid has 700 images, with a size of 222 x 224 pixels.

4. System Design
   The preprocessing stage is carried out on image data, namely feature extraction using HOG and random forest for classification. There are two stages, namely the training stage and the data testing stage. The training stage is carried out to determine the type of orchid leaf to be used. Each orchid leaf image will be converted using HOG feature extraction and the results will be used for the random forest method training process. After this process is complete, the model data can be used for testing. Furthermore, the type of orchid leaf will be converted into HOG feature extraction and the resulting image will be used in the checking process. After the checking process is complete, data is obtained that is close to or the same as the model.

5. Implementation
   Implementing the design into the interface. At this stage, the design that has been created is implemented into the program using the Python programming language. The existing dataset will be extracted with the HOG feature. At the evaluation stage, the Confusion Matrix is used to assess the performance of the model. The Confusion Matrix compares the model's prediction results against the test data with the actual labels, thus providing information on the number of correct and incorrect predictions for each type of orchid. By using the Confusion Matrix, we can calculate evaluation metrics such as accuracy, precision, recall, and F1-score, which provide an overview of the overall performance of the classification model.

## RESULTS AND DISCUSSION

In this section, we discuss the results of experiments using several machine learning models to identify orchid species. These experiments involved a comprehensive series of steps, including:

**Data Acquisition**

The data used in this study comes from a public dataset containing 2100 orchid leaf images from three types: Cattleya, Phalaenopsis, and Vanda. Each type of orchid has 700 images with a resolution of 222x224 pixels. This data is then split into training data and test data to build and test the model.

**Data Pre-processing**

At this stage, each image goes through a feature extraction process using the Histogram of Oriented Gradients (HOG) to obtain visual features that can represent the morphological characteristics of orchid leaves. The results of this HOG extraction are then used as input for the machine learning model.

**HOG Feature Extraction**

Feature extraction is done to obtain the characteristics of each image and these characteristics are used to recognize the image later. Some libraries used in this study include numpy which is used for numeric operations and multidimensional arrays. Then matplotlib which is used to create graphs and data visualization. While the skimage library is used to read and write images, and the sklearn library is used for the random forest algorithm.

The purpose of HOG extraction includes resizing the data to 128 x 64 pixels so that the image has a consistent size, most importantly for further processing and feature extraction. Then change the image shape into grayscale format because HOG features are usually extracted from grayscale images where color information is not needed to detect texture features. Next, determine the cell size to calculate the gradient histogram (in pixels) which is 8, 8, the block size containing several cells for normalization is 2, 2. This function will return two results, namely features in the form of a numeric array containing HOG features extracted from the image and used for model training and hog_image_rescaled which is a visualization image of the HOG feature for monitoring and analysis purposes. The following is an example of a Phalaenopsis orchid flower image used for the HOG process presented in **FIGURE 2**.

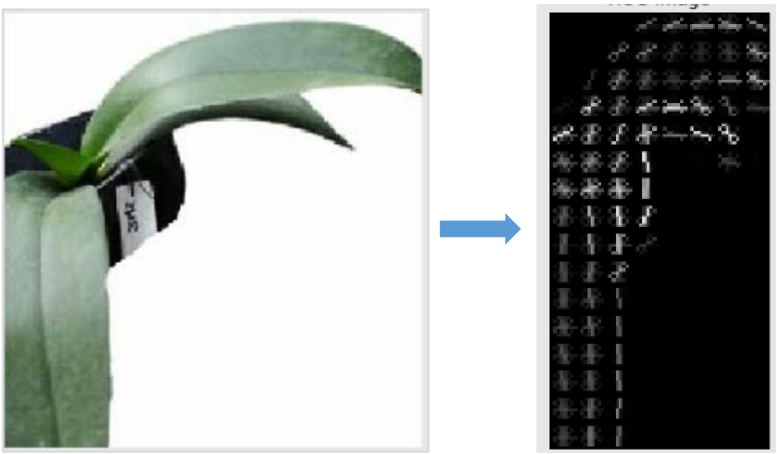

**FIGURE 2**. HOG Extraction of Phalaenopsis Orchid Type

The histogram form is the result of feature extraction that has been done using HOG using this feature and then classification can be done. The following are the results of HOG in the form of a histogram presented in **FIGURE 3**.

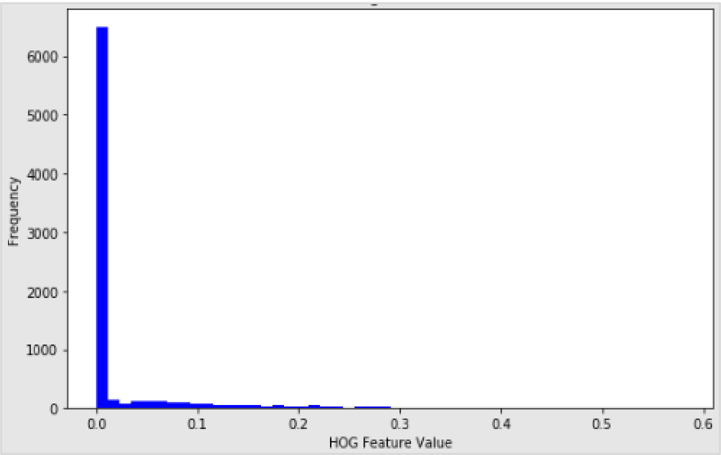

**FIGURE 3**. HOG Results in Histogram Form

## Classification Model

After successfully obtaining the characteristics of all types of orchids, the next process is to create a classification model. The classification model that will be built uses a random forest where this algorithm is able to classify well. Random forest works by building many decision trees and combining the results to improve prediction accuracy. The training process uses a random forest by initiating the value of the estimator to 100, this value is used to determine the number of decision trees that will be built in the forest. In this case, the model will build 100 decision trees. More trees usually improve model performance but also increase computing time. Furthermore, random_state = 42 is a parameter used to ensure consistent results every time the code is run. This is a random value used to control the decision tree creation process. By setting random_state to ensure that data distribution and decision tree selection are the same every time the model is run.

### Model Evaluation

The model training results are evaluated using a confusion matrix. Confusion matrix is an important evaluation tool in machine learning and statistics used to measure the performance of a classification model [13]. This visualization makes it easier to assess the classification error and accuracy of the built model which can be seen in **FIGURE 4**.

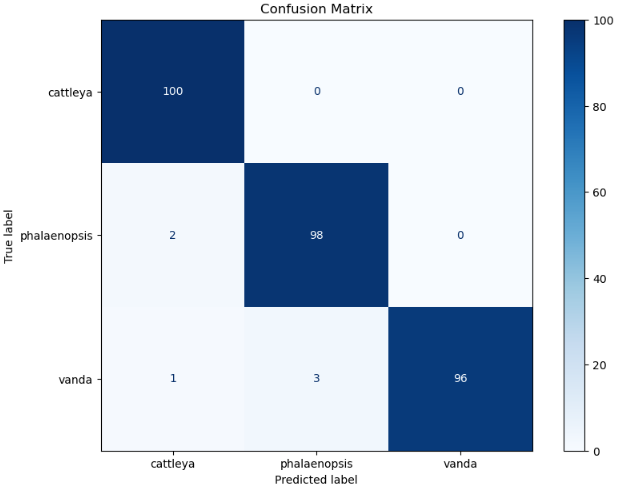

**FIGURE 4.** Confusion Matrix Results

The confusion matrix results of the built model can be seen in **FIGURE 4**, there are 100 images of the cattleya type with true positive values and 0 for the false positive, true negative, and false negative parameters, and so on for images of the phalaenopsis and vanda types. For the overall metrics, they can be calculated using the accuracy, precision, recall, and f1-score parameters. The results of the overall model evaluation metrics are as seen in **TABLE 1**.

**TABLE 1**. Overall Model Evaluation Metrics

| Label | Presisi | Recall | F1-Score |
|---|---|---|---|
| Cattleya | 0.97 | 1.00 | 0.99 |
| Phalaenopsis | 0.97 | 0.98 | 0.98 |
| Vanda | 1.00 | 0.96 | 0.98 |

In **TABLE 1**, the average Precision (Weighted) value is 0.98, the average Recall (Weighted) is 0.98, and the average F1-Score (Weighted) is 0.98. From the average results obtained, it can be concluded that the model built is good.

## CONCLUSIONS

Based on the results of this study, it can be concluded that the Random Forest method with the HOG (Histogram of Oriented Gradients) feature successfully classified three types of Cattleya, Phalaenopsis, and Vanda orchids with a very high level of accuracy, which is 98%. The evaluation results using the confusion matrix showed an average value of precision, recall, and F1-score of 0.98 each, indicating the reliability of the model in detecting and distinguishing the three types of orchids based on leaf morphology. This success shows that the developed model can be used as a reference in the process of identifying orchid types accurately and efficiently, which is useful for both beginners and practitioners in orchid cultivation.

## ACKNOWLEDGMENTS

We would like to thank the Center for Research and Community Service of the Politeknik Negeri Lampung for providing research funds through DIPA Polinela 2024. We would like to express our gratitude for the financial support provided, so that this research can be carried out.

## REFERENCES

[1] R. P. Putra, "Identifikasi Jenis Tanaman Anggrek Melalui Tekstur Bunga dengan Tapis Gabor dan M-SVM," *JOINTECS (Journal Inf. Technol. Comput. Sci.*, vol. 6, no. 1, p. 29, 2021, doi: 10.31328/jointecs.v6i1.1746.

[2] S. Andayani and L. Kusneti, "Using the Support Vector Machine Method with the HOG Feature for Classification of Orchid Types," *J. Teknol. Inf.*, vol. 21, no. 1, pp. 82–95, 2024, doi: https://doi.org/10.24246/aiti.v21i1.82-95.

[3] H. A. Shidiqy, B. F. Wahidah, and N. Hayati, "Karakterisasi Morfologi Anggrek (Orchidaceae) di Hutan Kecamatan Ngaliyan Semarang," *Al-Hayat J. Biol. Appl. Biol.*, vol. 1, no. 2, p. 94, 2019, doi: 10.21580/ah.v1i2.3761.

[4] E. C. Pratyaswara, N. K. Ayu Wirdiani, and G. M. Arya Sasmita, "Analisis Perbandingan Metode Canny, Sobel dan HSV dalam Proses Identifikasi Bunga Anggrek Hibrida," *J. Ilm. Merpati (Menara Penelit. Akad. Teknol. Informasi)*, vol. 5, no. 3, p. 11, 2017, doi: 10.24843/jim.2017.v05.i03.p03.

[5] D. P. Pamungkas, "Ekstraksi Citra menggunakan Metode GLCM dan KNN untuk Identifikasi Jenis Anggrek (Orchidaceae)," *Innov. Res. Informatics*, vol. 1, no. 2, pp. 51–56, 2019, doi: 10.37058/innovatics.v1i2.872.

[6] S. Akbar, R. A. Fahreza, F. Ferdianto, D. F. Wulandari, and E. W. Astuti, "Klasifikasi Pes Planus Menggunakan Ekstraksi Fitur HOG dan BoF dengan Random Forest," *J. Electron. Instrum.*, vol. 1, no. 3, pp. 103–113, 2024, doi: https://doi.org/10.19184/jei.v1i3.980.

[7] Q. Xia, H.-D. Zhu, Y. Gan, and L. Shang, "Plant Leaf Recognition Using Histograms of Oriented Gradients," in *Intelligent Computing Methodologies*, D.-S. Huang, K.-H. Jo, and L. Wang, Eds., Cham: Springer International Publishing, 2014, pp. 369–374. doi: 10.1007/978-3-319-09339-0_38.

[8] D. P. Mohandoss, Y. Shi, and K. Suo, "Outlier Prediction Using Random Forest Classifier," in *2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC)*, 2021, pp. 27–33. doi: 10.1109/CCWC51732.2021.9376077.

[9] P. Mekha and N. Teeyasuksaet, "Image Classification of Rice Leaf Diseases Using Random Forest Algorithm," in *2021 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunication Engineering*, 2021, pp. 165–169. doi: 10.1109/ECTIDAMTNCON51128.2021.9425696.

[10] J. Bal, M. K. Rath, and P. K. Swain, "Plant Leaf Identification Using HOG and Random Forest Regressor," in *Smart Computing Techniques and Applications*, S. C. Satapathy, V. Bhateja, M. N. Favorskaya, and T. Adilakshmi, Eds., Singapore: Springer Singapore, 2021, pp. 515–525.

[11] Meiriyama and Sudiadi, "Penerapan Algoritma Random Forest Untuk Klasifikasi Jenis Daun Herbal," *J. Teknol. Sist. Inf.*, vol. 3, no. 1, pp. 131–138, 2022, doi: https://doi.org/10.35957/jtsi.v3i1.3176.

[12] Z. Ibrahim, N. Sabri, N. Nabilah, and A. Mangshor, "Leaf Recognition using Texture Features for Herbal Plant Identification," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 9, no. 1, pp. 152–156, 2018, doi: 10.11591/ijeecs.v9.i1.pp152-156.

[13] B. A. Riyadi, S. Adi, A. D. Laksito, M. Hayaty, O. Arifin, and A. Fatkhurohman, "The Effect of Augmentation on Classification Algorithm to Determine Photo Angle," *Int. Conf. Electr. Eng. Comput. Sci. Informatics*, no. September, pp. 254–260, 2023, doi: 10.1109/EECSI59885.2023.10295875.