Journal of Computer Science, Information Technology and Telecommunication Engineering (JCoSITTE)
Vol. 6, No. 2, September 2025, pp. 1019~1027
ISSN: 2721-3838, DOI: 10.30596/jcositte.v6i2.26495

Herbal Plant Image Retrieval Using HSV Color Histogram and Random Forest Algorithm

Fadhillah Azmi[1], M Khalil Gibran[2], Amir Saleh[3]
[1]Department of Electrical Engineering, Universitas Medan Area, Medan, Indonesia
[2]Department of Computer Science, Universitas Islam Negeri Sumatera Utara, Medan, Indonesia
[3]Department of Computer Engineering and Informatics, Politeknik Negeri Medan, Medan, Indonesia

ABSTRACT

Herbal plants have significant importance in traditional medicine and are often useful in various natural health products. Visual identification of these plants is usually carried out based on the shape of the leaves and often encounters difficulties in distinguishing species due to similarities in shape and color. Therefore, a system capable of automatically and efficiently recognizing and searching for herbal plant images is needed. This study aims to implement an image search engine for herbal plants based on leaf color similarity. The method used includes color feature extraction using an HSV (Hue, Saturation, Value) histogram with an 8×8×8 bin configuration, resulting in a 512-dimensional feature vector. This histogram feature is then used as input for the Random Forest classification algorithm to group images based on the type of herbal plant. The dataset used consists of 450 herbal leaf images from 9 different classes, obtained through direct image capture using a digital camera. The test results indicates that the developed system is able to classify types of herbal plants with an accuracy of 95.56%. In addition, the computation time and system response during both training and testing processes are relatively fast and efficient. The advantage of this system lies in the simplicity of feature extraction while still being able to provide high classification performance. This system has great potential to be used as an educational tool as well as an initial component in the development of mobile applications for automatic herbal plant identification.

Keywords: Herbal Plants; **HSV Histogram**; **Random Forest; Image Retrieval; Image Classification**.

Corresponding Author:
Amir Saleh
Department of Computer Engineering and Informatics
Politeknik Negeri Medan
Jl. Almamater No. 1 Kampus USU Medan, 20155, Indonesia
Email: amirsaleh@polmed.ac.id

Article history:
Received Aug 30, 2025
Revised Sep 7, 2025
Accepted Sep 9, 2025

1. INTRODUCTION

Herbal plants have been widely used as an alternative to traditional medicine because they contain various bioactive compounds that are beneficial for health (Rahayu et al., 2020). In Indonesia, which is known for its rich biodiversity, there are thousands of herbal plant species utilized in traditional medicine (Iskandar et al., 2020; Jadid et al., 2020). However, the process of identifying herbal plant species is still mostly carried out manually by botanists or herbalists. This method heavily relies on experience and visual knowledge, making it inefficient and prone to identification errors (Falah & Hadiwibowo, 2017). In this context, technology-based systems that assist in the automatic identification of herbal plants have become increasingly important (Harjanti et al., 2020; Yunitarini & Widiaswanti, 2024).

One of the recent approaches in automatic plant identification systems is Content-Based Image Retrieval (CBIR), which enables image searching and classification based on visual feature similarities such as color, shape, and texture (Chugh et al., 2022). Among these visual features, color remains the most dominant aspect, particularly in leaf image recognition. A common color representation is the color histogram in the HSV (Hue, Saturation, Value) color space, as HSV can separate color information from light intensity and is more robust against illumination changes compared to the RGB color space. Recent studies have shown that integrating HSV histograms with machine learning algorithms such as Support Vector Machine (SVM) and Convolutional Neural Network (CNN) significantly improves the accuracy of classifying plant leaf images (Abinaya et al., 2021). Moreover, the use of HSV features combined with contour-based segmentation has also been proven to enhance system performance in automatic plant identification (Yunitarini & Widiaswanti, 2024).

In the development of plant image search and classification systems, the Random Forest algorithm is one of the widely used approaches due to its advantages in accuracy, efficiency in handling high-dimensional data, and robustness against overfitting. This algorithm combines the prediction results of multiple randomly constructed decision trees, resulting in more stable and reliable final classifications. Several recent studies have demonstrated the effectiveness of combining HSV color features with the Random Forest algorithm in detecting plant leaf diseases. For instance, Pandey & Vir, (2024) integrated HSV histograms with Random Forest and CNN for plant disease classification and achieved high accuracy. Meanwhile, Vishnoi et al. (2024) employed a stacking ensemble approach that included Random Forest for leaf disease classification and showed significant performance improvements. These findings indicate that the Random Forest algorithm is highly suitable for application in the development of herbal plant identification systems based on images.

Based on the aforementioned background, this study aims to develop an herbal plant image retrieval system capable of automatically identifying and searching for images based on the visual similarity of leaf characteristics. The proposed system employs color feature extraction through histogram representation in the HSV (Hue, Saturation, Value) color space, followed by classification using the Random Forest algorithm. The dataset used in this study was directly obtained through the image acquisition of leaves from nine types of herbal plants, for a total of 450 images. Each image was processed to extract its color features, which were then used in the training and testing stages of the classification model. It is expected that this system can make a major contribution to the field of digital image processing for plants, particularly in developing automatic identification systems that support education, the preservation of medicinal plants, and the development of mobile-based applications in the agricultural and traditional health sectors.

2. RESEARCH METHOD

This study adopts an experimental quantitative approach aimed at developing and testing an herbal plant image retrieval system based on color features. The system is designed to accept input in the form of herbal leaf images, extract color features from the images using HSV histograms, and classify them using the Random Forest algorithm. In addition to classification, the system also displays the most similar images from the database based on the degree of feature similarity.

A. System Architecture

The overall system architecture consists of five main stages: (1) input of herbal leaf images, (2) HSV color feature extraction, (3) labeling and storing features into the database, (4) training and classification using the Random Forest algorithm, and (5) retrieving the most similar images to the input image based on similarity distance. Figure 1 illustrates the proposed system architecture.

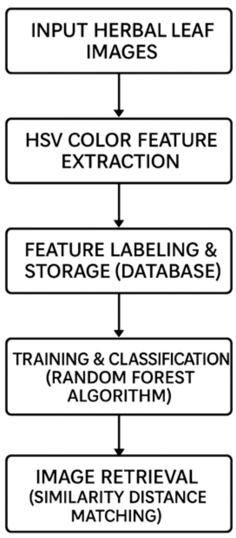


Fig 1. System Architecture

B. Herbal Plant Image Dataset

The dataset used in this study was obtained through direct image acquisition using a digital camera on commonly found plants. It consists of 9 classes of herbal plants, namely Acalypha australis L., Murraya paniculata, Murraya koenigii, Sauropus androgynus, Dimocarpus longan L., Vernonia amygdalina, Polyscias scutellaria, Artocarpus heterophyllus, and Syzygium polyanthum. Each image was converted into a fixed size of 512×512 pixels in .jpg format. Labels were manually assigned and grouped into folders to facilitate automatic labeling.

C. Image Preprocessing

The image preprocessing stage is a crucial step in the image classification system, aiming to ensure that all input data have a uniform format before feature extraction. The first step is resizing all images to a fixed resolution of 128 × 128 pixels to ensure dimensional consistency and computational efficiency (Yuan et al., 2024). Next, pixel value normalization was performed by reducing the intensity range from 0–255 to 0–1 using linear transformation according to Equation (1) below.

$$I_{norm}(x,y) = \frac{I(x,y)}{255} \tag{1}$$

In this context, *I(x,y)* represent the pixel intensity value at the coordinates *(x,y)*, while $I_{norm}(x,y)$ denotes the normalized result (Gonzalez & Woods, 2018).

The next step is the conversion of the color space from RGB (Red, Green, and Blue) to HSV (Hue, Saturation, and Value). The HSV color model was chosen because it is able to separate the color components (Hue and Saturation) from illumination (Value), making it more reliable in distinguishing color differences in biological objects such as plant leaves (Priya et al., 2025; Yuan et al., 2024). After this process, the converted and normalized images were stored in an array of sizes [128, 128, 3] with pixel values in the range [0, 1], making them ready for histogram feature extraction and the classification stage.

D. Feature Extraction with HSV Color Histogram

Image feature extraction was carried out using a three-dimensional (3D) color histogram based on HSV channels, where each channel (Hue, Saturation, Value) was divided into 8 bins. The combination of these three channels produced a total of 8×8×8 = 512 feature values for each image (Yuan et al., 2024). The histogram was computed by accumulating the pixel frequency in each HSV bin combination, followed by a normalization process to ensure that the distribution of feature values remained within a uniform scale. Normalization was performed using Equation (2) below.

$$H_{norm}(i,j,k) = \frac{H(i,j,k)}{\sum_{i=1}^{8}\sum_{j=1}^{8}\sum_{k=1}^{8} H(i,j,k)} \tag{2}$$

In this context, *H(i,j,k)* denotes the number of pixels in the *i, j, k-th* bin of the HSV histogram, while $H_{norm}(i,j,k)$ indicates the corresponding normalized histogram value (Gonzalez & Woods, 2018).

The normalized histogram is then flattened into a one-dimensional vector of 512 elements. This feature vector is used as input for the classification and image matching stages in the plant identification system. This approach is commonly applied in color-based image recognition because it can efficiently and representatively capture the distribution of color information (Ismail & Malik, 2022).

E. Classification Using Random Forest

The image classification stage was carried out using the Random Forest algorithm, an ensemble learning method that combines predictions from multiple decision trees. The algorithm builds a model based on a collection of decision trees trained independently using random subsets of both training data and features, with the final classification determined through a majority voting mechanism. Mathematically, the prediction result of Random Forest can be expressed using Equation (3) below.

$$\hat{y} = mode(T_1(x), T_2(x), \ldots, T_n(x)) \tag{3}$$

In this equation, $T_i(x)$ represents the classification result of the *i-th* decision tree for the feature vector *x*, while n indicates the total number of trees in the model (Izquierdo-Horna et al., 2025).

In this study, a total of 100 decision trees were used to construct the Random Forest model. The selection of this number of trees was based on considerations of prediction stability and computational efficiency. The Random Forest algorithm is known for its advantages in handling high-dimensional data and its robustness against overfitting.

F. System Performance Evaluation

The performance evaluation of the classification system was conducted using confusion matrix-based metrics, namely accuracy, precision, recall (sensitivity), and F1-score. These four metrics were selected to offer an in-depth assessment of the model's capability in automatically recognizing herbal leaf images, particularly under multi-class or imbalanced data conditions.

Accuracy measures the proportion of correct predictions to the total test data and is calculated using the following Equation (4):

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{4}$$

Precision indicates the degree of accuracy in positive classifications, namely the proportion of predicted positive instances that are truly positive, as shown in Equation (5):

$$Precision = \frac{TP}{TP+FP} \tag{5}$$

Recall (Sensitivity) measures the ability of the model to correctly identify all positive samples of a given class, as defined in Equation (6):

$$Recall = \frac{TP}{TP+FN} \tag{6}$$

The F1-score represents the harmonic mean of Precision and Recall, reflecting their balance as defined in Equation (7):

$$F1 - score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{7}$$

These four metrics are used complementarily to evaluate the generalization ability of the model and to avoid bias toward the majority class (Zhang et al., 2023).

In addition to numerical evaluation, the system was also tested in terms of retrieval efficiency (retrieval time), which refers to the time required from the input image until the search results are displayed. This aspect is crucial for the implementation of real-time application-based systems, such as mobile devices or field systems (Wu et al., 2024). The system was also qualitatively evaluated in terms of the visual relevance of the retrieval results, by comparing the visualization of the input image with the retrieved images based on similarities in color, shape, and texture patterns. The goal is to ensure that the results are not only statistically accurate but also perceptually meaningful according to human observation (Li et al., 2025).

3. RESULTS AND DISCUSSION

A. Image Preprocessing Results

After the dataset was collected, all herbal leaf images underwent preprocessing to standardize their dimensions and format. All images were resized to 128 × 128 pixels and converted from the RGB color space to HSV. The pixel values were normalized into the range [0, 1] to ensure stability during the feature extraction process. This conversion produced three color channels (Hue, Saturation, Value), which were used in the histogram process. An example of the image preprocessing results is shown in Figure 2 below.

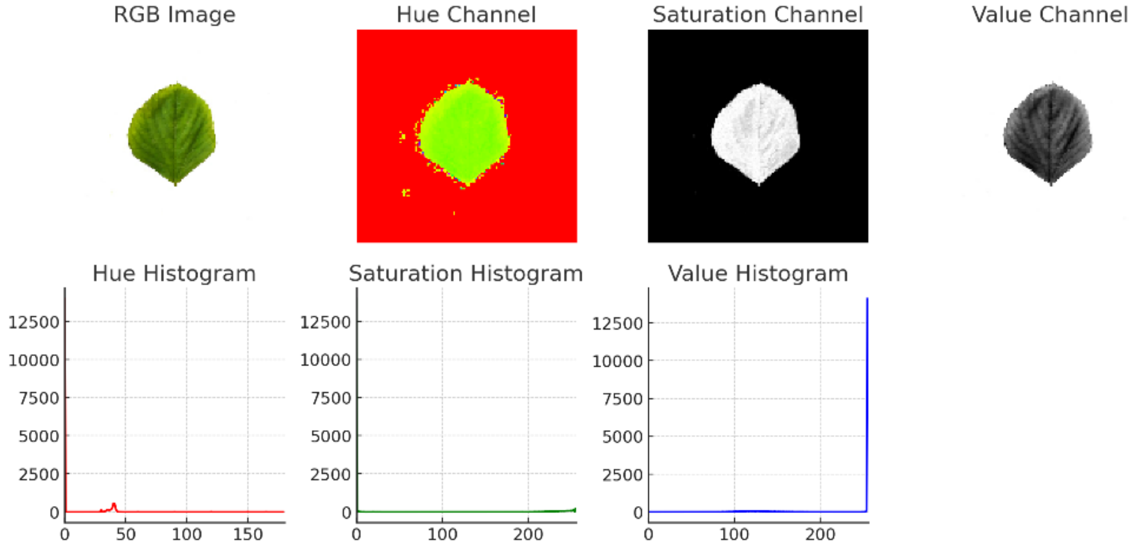


Fig 2. Image Preprocessing Results

B. HSV Histogram Feature Extraction Results

After the preprocessing stage, leaf images were converted into the HSV color space and a three-dimensional histogram was constructed by dividing each channel (Hue, Saturation, and Value) into eight bins. The result of this process is a color histogram representation with dimensions of 8 × 8 × 8, yielding a total of 512 elements.

Each histogram reflects the color distribution of an image in the form of frequency statistics for HSV value combinations. The obtained histograms were then normalized so that their values ranged between [0, 1], and subsequently flattened into a one-dimensional vector of 512 elements. This vector serves as the feature representation of the color characteristics of each leaf image. An example of a feature vector extracted from a single image is as follows:

[0.0000, 0.0027, 0.0068, 0.0019, ..., 0.0031]

Each image in the dataset produces a similar vector, which is then used in the classification and clustering processes to identify leaf species based on their color distribution.

C. Training and Classification Results of the Random Forest Model

The feature vectors extracted from HSV histograms were used as input for the model training process. The algorithm employed was the Random Forest Classifier, an ensemble learning method based on decision trees. In this study, the model was built with 100 trees (*n_estimators = 100*), chosen to achieve a balance between accuracy and computational efficiency.

The model was trained using leaf image data represented in 512-dimensional feature vectors. The training and testing processes were conducted using a cross-validation approach with a train-test split ratio of 80:20. The model's performance evaluation results are presented in Table 1 below.

Table 1. Random Forest Model Evaluation Results

Performance Evaluation	**Results**
Accuracy	95.56%
Precision	95.96%
Recall	95.56%
F1 – Score	95.54%

The accuracy value, as shown in Table 1, reached 95.56%, indicating that the model was able to correctly classify the majority of images. The balanced precision and recall values demonstrate that the model not only performs well in recognizing positive classes but also maintains a low rate of false negatives. The F1-score, which is close to 96%, further indicates that the model delivers stable and reliable overall performance.

In addition to quantitative evaluation metrics, an analysis of the confusion matrix was also conducted. The results showed that the model could recognize most plant classes well. Misclassifications generally occurred due to similarities in dominant leaf colors between classes, for example, in cases where two types of herbal plants had nearly identical leaf colors. This suggests that while color features are quite effective, they have limitations in distinguishing classes that are very visually similar.

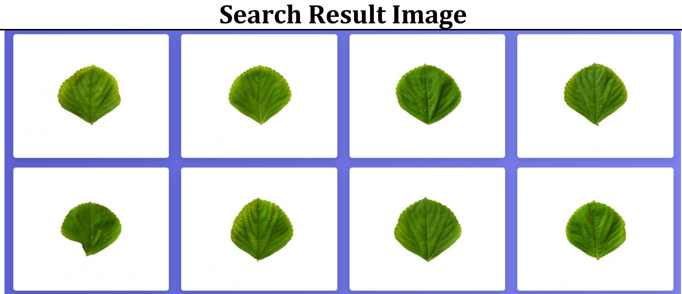
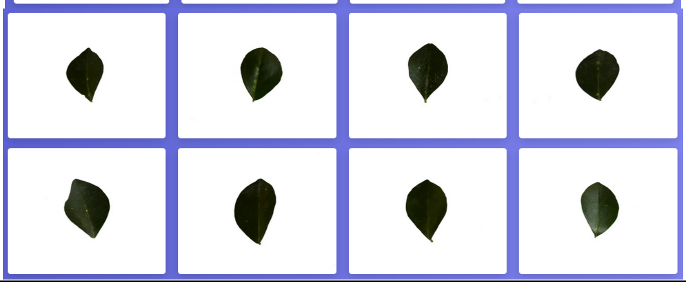
D. Image Retrieval System Implementation Results

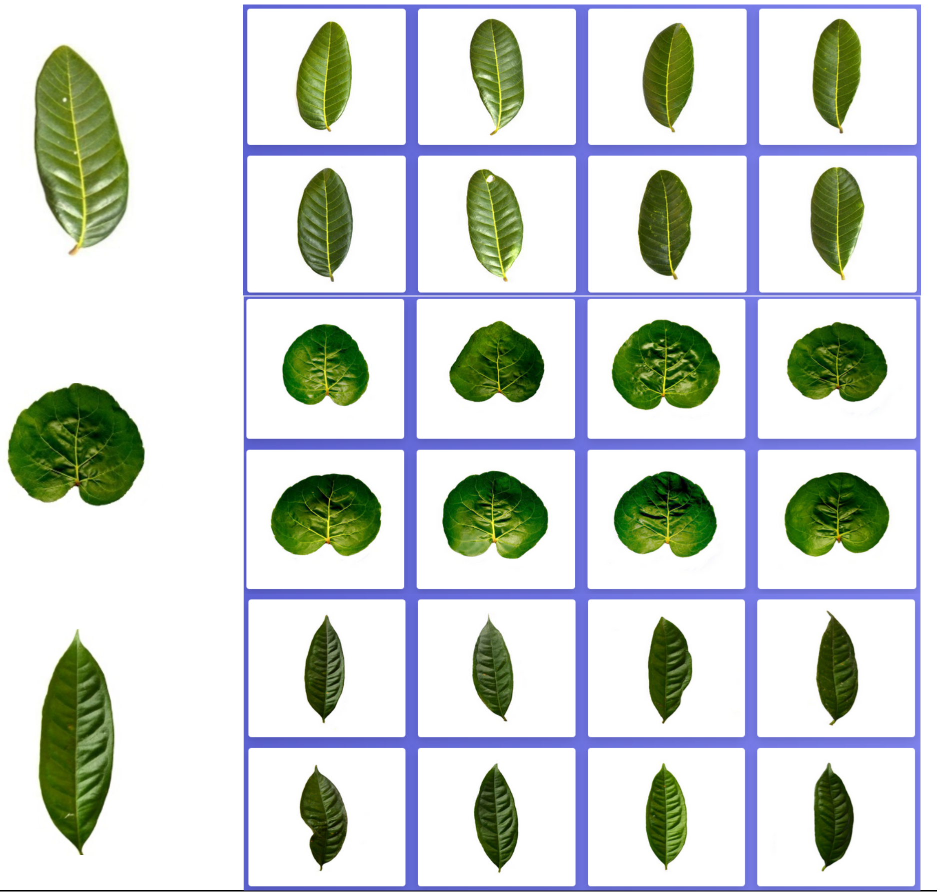
The developed system is capable of receiving an input in the form of a query image, then extracting its HSV color features in the form of a normalized three-dimensional histogram. The extracted features are subsequently fed into the trained Random Forest model to classify the image as one of the nine herbal plant classes. The input image process was carried out in the initial stage of the developed system by testing a single input image, as illustrated in Figure 3 below.

Fig 3. Form Input Image

After the classification process is completed, the system displays the predicted plant class label along with several other images from the same class as the most relevant or similar results. This approach enables the system not only to recognize the plant species from the input image but also to provide visual references from the dataset based on the classification results. The display of the test results or the system's response to the input image can be seen in Table 2 below.

Table 2. System Testing Results

Input Query	Search Result Image
	
	



From Table 2, the system successfully displayed images with leaf colors and structures very similar to the input image.

E. System Evaluation

The system evaluation was carried out from three main aspects: classification accuracy, average retrieval time, and retrieval relevance. Based on the experimental results, the system achieved a classification accuracy of 95.56%, demonstrating that HSV color histogram features are effective in distinguishing leaf colors across different herbal plant species. This proves that even when using only color information, the system is still able to achieve a high level of classification accuracy.

In terms of speed, the average time required by the system to perform retrieval—from input image to displaying the results—ranged from 0.8 to 1.3 seconds, depending on the size of the database and the hardware specifications used. This indicates that the system operates responsively and can be relied upon for practical applications.

Furthermore, the relevance of the retrieval results also showed satisfactory performance. Most of the retrieved images belonged to the same class or were visually very similar to the input image, in accordance with the classification results produced by the Random Forest algorithm. Thus, the system is not only fast and accurate but also capable of providing visually relevant and informative results for users.

F. Discussion

The implementation and evaluation results show that the method used in this study—feature extraction with HSV color histograms and classification using the Random Forest algorithm—provides satisfactory performance in identifying and retrieving herbal plant images. This is evident from the classification accuracy of 95.56% and the relatively fast system response time of less than 2 seconds per retrieval.

The use of the HSV color space contributes significantly to the quality of image feature representation. HSV is considered to better approximate how humans perceive colors compared to the RGB color space. The Hue channel captures information about color type, Saturation represents color intensity, and Value corresponds to brightness. By focusing on these three channels, the system is able to capture differences in leaf color even though shape and texture are not explicitly considered. This aligns with the system's objective, which relies primarily on color-based visual features.

The color histogram constructed with 8 bins per channel produced a feature vector of 512 values. Normalization of the histogram ensured that the color distribution was not affected by image size, making the extracted features relative and consistent across images. These histogram vectors served as the main input for the Random Forest algorithm in the training and classification process.

Random Forest was chosen because it is a proven ensemble learning method that reliably handles datasets with a large number of features and does not require extensive additional data normalization. This algorithm combines multiple decision trees and determines the classification result through majority voting. The advantage of Random Forest lies in its ability to avoid overfitting while delivering stable and consistent results.

In the context of image retrieval, the system does not explicitly calculate distances such as Euclidean distance. Instead, it presents images from the same class based on the classification output of Random Forest. This approach proved effective, as most of the retrieved results matched the correct class or were visually very similar to the input image. Thus, users not only obtain information about the classification of the herbal plant but can also view additional visual references of the same species.

Nevertheless, the system still has several limitations. One of them is its sensitivity to the quality of input images, such as those with excessive lighting, shadows, or blurriness. Since the system relies solely on color information, variations in texture and shape are not considered. This may lead to misclassification when plants have similar dominant colors. Moreover, the system currently lacks user-based validation features or interactive refinement mechanisms to improve the results.

From the results obtained, it can be concluded that the HSV color histogram and Random Forest-based approach are effective for a simple herbal plant image retrieval system. However, to further improve accuracy and robustness against environmental variations, the system can be enhanced by incorporating texture and shape features or even applying deep learning techniques such as Convolutional Neural Networks (CNN), which have been proven effective in complex image recognition tasks.

4. Conclusion

Based on the results and implementation of the system, it can be concluded that the application of HSV color histogram-based feature extraction and classification using the Random Forest algorithm is effective in developing an herbal plant image retrieval system. The conversion from RGB to HSV color space allows the system to capture leaf color characteristics more representatively, particularly through the Hue and Saturation channels, which play the most significant roles in visual identification. The three-dimensional color histogram with 512 bins was successfully reduced into a feature vector that can be efficiently used for classification and image retrieval. The Random Forest model trained with histogram feature vectors achieved excellent classification performance with an accuracy of 95.56%. This demonstrates that even with simple color-based features, the system is capable of reliably distinguishing different herbal plant species.

REFERENCES

Chugh, H., Gupta, S., Garg, M., Gupta, D., Mohamed, H. G., Noya, I. D., Singh, A., & Goyal, N. (2022). An Image Retrieval Framework Design Analysis Using Saliency Structure and Color Difference Histogram. *Sustainability (Switzerland)*, *14*(16). https://doi.org/10.3390/su141610357

Falah, F., & Hadiwibowo, N. (2017). Species Identification of Traditional Medicine Plants for Women's Health in East Kalimantan: Lesson Learned from Local Wisdom. *Indonesian Journal of Forestry Research*, *4*, 49–68. https://doi.org/10.20886/ijfr.2017.4.1.49-68

Gonzalez, R. C., & Woods, R. E. (2018b). *Digital Image Processing* (4th ed.). Pearson.

Harjanti, T. W., Madenda, S., Harlan, J., & Lussiana, E. T. P. (2020). Study and Research on the Identification of the Leaves of Indonesian Herbal Medicines Using Manhattan Distance and Neural Network Algorithms. *2020 Fifth International Conference on Informatics and Computing (ICIC)*, 1–6. https://doi.org/10.1109/ICIC50835.2020.9288564

Iskandar, B. S., Iskandar, J., Partasasmita, R., & Irawan, B. (2020). Various medicinal plants traded in the village market of karangwangi village, southern cianjur, west java, indonesia. *Biodiversitas*, *21*(9), 4440–4456. https://doi.org/10.13057/biodiv/d210963

Ismail, N., & Malik, O. A. (2022). Real-time visual inspection system for grading fruits using computer vision and deep learning techniques. *Information Processing in Agriculture*, *9*(1), 24–37. https://doi.org/https://doi.org/10.1016/j.inpa.2021.01.005

Izquierdo-Horna, L., Kahhat, R., & Vázquez-Rowe, I. (2025). Applying random forest to forecast municipal solid waste generation from household fuel consumption. *Resources, Conservation & Recycling Advances*, *27*, 200264. https://doi.org/https://doi.org/10.1016/j.rcradv.2025.200264

Jadid, N., Kurniawan, E., Himayani, C. E. S., Andriyani, Prasetyowati, I., Purwani, K. I., Muslihatin, W., Hidayati, D., & Tjahjaningrum, I. T. D. (2020). An ethnobotanical study of medicinal plants used by the Tengger tribe in Ngadisari village, Indonesia. *PLOS ONE*, *15*(7), e0235886-. https://doi.org/10.1371/journal.pone.0235886

Li, Z., Lu, S., Yuan, Z., Hou, B., & Bian, J. (2025). Interactive Instance Search: User-Centered Enhanced Image Retrieval with Learned Perceptual Image Patch Similarity. *Electronics (Switzerland)*, *14*(9). https://doi.org/10.3390/electronics14091766

N.S Abinaya, D., S., & S., R. K. (2021). Naive Bayesian fusion based deep learning networks for multisegmented classification of fishes in aquaculture industries. *Ecological Informatics*, *61*, 101248. https://doi.org/https://doi.org/10.1016/j.ecoinf.2021.101248

Pandey, A., & Vir, R. (2024). An Improved Approach to Classify Plant Disease using CNN and Random Forest. *2024 IEEE 3rd World Conference on Applied Intelligence and Computing (AIC)*, 795–805. https://doi.org/10.1109/AIC61668.2024.10730901

Priya, P., Kumar, D. A. N., Rithika, M., Sanjay, S., & Jeyamary, J. S. (2025). Enhancing Human Safety through Modesty Guard: A Comprehensive Review of OpenCV-Based Surveillance Systems. *2025 International Conference on Artificial Intelligence and Data Engineering (AIDE)*, 484–491. https://doi.org/10.1109/AIDE64228.2025.10987301

Rahayu, Y. Y. S., Araki, T., & Rosleine, D. (2020). Factors affecting the use of herbal medicines in the universal health coverage system in Indonesia. *Journal of Ethnopharmacology*, *260*. https://doi.org/10.1016/j.jep.2020.112974

Vishnoi, V. K., Kumar, K., Kumar, B., & Bhutiani, R. (2024). A stacking ensemble machine learning based approach for classification of plant diseases through leaf images. *Environment Conservation Journal*, *25*(3), 767–778. https://doi.org/10.36953/ECJ.28742840

Wu, G., Haider, A., Tian, X., Loweimi, E., Chan, C. H., Qian, M., Muhammad, A., Spence, I., Cooper, R., Ng, W. W. Y., Kittler, J., Gales, M., & Wang, H. (2024). Multi-modal video search by examples—A video quality impact analysis. *IET Computer Vision*, *18*(7), 1017–1033. https://doi.org/https://doi.org/10.1049/cvi2.12303

Yuan, H., Xiong, Z., & Chen, B. (2024a). Phenotypic Analysis and Variety Classification of Waxy Corn Seeds Based on Machine Vision. *2024 International Conference on Image Processing, Computer Vision and Machine Learning (ICICML)*, 912–915. https://doi.org/10.1109/ICICML63543.2024.10958153

Yuan, H., Xiong, Z., & Chen, B. (2024b). Phenotypic Analysis and Variety Classification of Waxy Corn Seeds Based on Machine Vision. *2024 International Conference on Image Processing, Computer Vision and Machine Learning (ICICML)*, 912–915. https://doi.org/10.1109/ICICML63543.2024.10958153

Yunitarini, R., & Widiaswanti, E. (2024a). Analysis and Design of Indonesian Traditional Medicine (Jamu) Information System by using Prototyping Model (Case Study: Madura Island). *E3S Web of Conferences*, *483*. https://doi.org/10.1051/e3sconf/202448303012

Yunitarini, R., & Widiaswanti, E. (2024b). Analysis and Design of Indonesian Traditional Medicine (Jamu) Information System by using Prototyping Model (Case Study: Madura Island). *E3S Web of Conferences*, *483*. https://doi.org/10.1051/e3sconf/202448303012

Zhang, J., Hou, Y., & Han, H. (2023). Hybrid evolutionary robust optimization-based optimal control for time-delay nonlinear systems. *Information Sciences*, *647*, 119395. https://doi.org/https://doi.org/10.1016/j.ins.2023.119395