

Leaf Disease Detection Using Color Histogram and Random Forest on *Pongamia Pinnata* (L.) Pierre

Sarifah Agustiani, Agus Junaidi, Yoseph Tajul Arifin, Dwi Puji Hastuti, Sopiyan Dalis, Kartika Yuliantari and Fauzi Syarief

Universitas Bina Sarana Informatika, Jakarta, Indonesia

Keywords: Color Histogram, Random Forest, *Pongamia Pinnata* (L.) Pierre.

Abstract: Increasing energy consumption, which is disproportionate to the energy supply, has resulted in an urge to seek renewable alternative energy sources that are environmentally friendly to meet energy needs. One plant with great potential to be used as an alternative fuel/biodiesel which environmentally friendly is *Pongamia Pinnata* (L.) Pierre. Besides many benefits and advantages of growing fast in tropical and sub-tropical areas, maintaining growth, and meeting the supply of bioenergy, it is necessary to have an intelligent system that can detect diseases in these plants. The research aims to classify *Pongamia Pinnata* into healthy and diseased categories. Hopefully, this system will prevent plant disease transmission and less-than-optimal growth. The method uses a color histogram as a feature extraction to recognize the characteristics of each image. In contrast, it uses a random forest algorithm for the classification process, and the accuracy reaches 99.79%.

1 INTRODUCTION

Indonesia is currently facing energy problems. The energy situation in Indonesia is caused by an increase in energy consumption which is not proportional to the energy supply required. Due to the ever-increasing demand, this oil-exporting country has become an oil-importing country (Evans, 2021). According to the Central Statistics Agency (BPS), Indonesia's oil import volume from January to September 2022 was 30.06 million tonnes. This volume increased by 16.89% compared to the same period in the previous year. More precisely, imports of 11.23 million tonnes of oil increased by 7.7%, and implication of oil products amounted to 18.84 million tonnes, an increase of 23.15% from January to September 2022, which can be seen in Figure 1. (Kusnandar, 2022)

It has resulted in an urge to seek renewable alternative energy sources that are environmentally friendly to meet energy needs. One uses biofuels from biological sources, such as bioethanol, biodiesel, and biogas. Several plants that can produce biofuels are oil palm, coconut, jatropha, cotton, canola, rapeseed for biodiesel, cassava, sweet potato, sugar cane, sorghum, sago, palm sugar, nipa palm, and lontar for bioethanol (Maretta et al., 2016). One plant with great

potential to be used as an alternative fuel/biodiesel that is environmentally friendly is *Pongamia pinnata* (L.) Pierre plant (Tamin and Puri, 2020). This plant is one of the most potential biodiesel raw materials to be developed in Indonesia, spread naturally from Sumatra to Papua (A. Aminah and A. Suryani, 2017). Apart from the *Pongamia pinnata* plant, there are several other bioenergy crops such as oil palm, corn and soybeans, and others, but these plants' growth has been detention due to dependence on fertile soil (Evans, 2021).

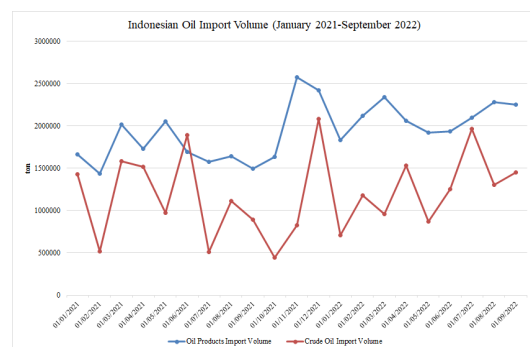


Figure 1: Oil Import Volume.

Meanwhile, the *Pongamia pinnata* plant grows fast in tropical and subtropical regions, even with

low rainfall (Dendang and Suhaendah, 2017). The *Pongamia pinnata* plant, apart from being a source of energy, is also versatile because it can be used as a greening plant, medicinal plant, windbreaker plant, animal feed, and vegetable pesticide (Suita and Syamsuwida, 2016). Several studies on *Pongamia pinnata* have been carried out, especially in the sustainability of renewable energy (Radhiana, 2023)(Khasanah, 2020)(Mitra, 2021) composition and excellence in agriculture (Usharani et al., 2019). But besides having advantages, of course, every plant also has threats; because there are many biotic and abiotic factors, the normal growth of these plants is affected in most cases, especially in the leaves, flowers, and fruit of these plants, often resulting in a decrease in the quality and quantity of the growth of these plants. So that growth is not optimal and even causes death for some time, which ultimately has an impact on reducing the environmental and economic value of plants (More et al., 2019). Clear and limited structural changes can characterize disease symptoms in these plants. It shows spots or spots on plant parts such as leaves, fruit, and roots (Sutarman, 2017). However, the most obvious symptoms can be from the leaves because changes in the color and shape of the spots on the leaves are more obvious, so the leaves can be used as a first step to detect disease in this plant.

The most modern approach is machine learning or deep learning using various algorithms to increase recognition and accuracy in detecting and diagnosing plant diseases (Maniyath, 2018), as previous research has been done in classifying leaf images on two different plants, namely *Jatropha Curcas L* and *Pongamia Pinnata L*, which uses machine learning with segmentation and classification techniques (Chouhan et al., 2021). This study also aims to classify the images of *Pongamia pinnata* leaves into two categories, healthy and diseased leaves, using the color histogram as a feature extraction. For the classification process, the random forest algorithm is used.

In some studies, the color histogram is used to track long-distance vehicles. The results show that the proposed method improves accuracy in real-time applications, as expected in the study (Yankun et al., 2021); other studies conducted in video data retrieval showed appropriate performance (Saravanan and Surendiran, 2019). Likewise, the use of the random forest algorithm has been carried out in several previous studies; it has an accuracy value of 99.65% in classifying rice leaf disease (Agustiani et al., 2022). Other research was predicting rainfall with an accuracy of 99.45% (Primajaya and Sari, 2018), the price prediction of mobile phones (Saadah and Salsabila, 2021), forecasting crypto assets in the futures market

(Orte et al., 3 01), Forest Mapping (Purwanto et al., 2023), Alzheimer's disease prediction (Singh et al., 2023) lupus disease prediction (Chen, 2022). Thus, in this research, it is proposed to apply the random forest method with color histogram feature extraction to classify *Pongamia* leaf images in healthy and diseased classes with the hope that this model can provide accurate model performance in the classification process.

2 METHODS

The stages started with collecting data in the form of secondary data, determining the location of the dataset, then making arrangements beginning from the number of images in each class, the size of the pixels, and the dimensions of the images used in the research, then implementing feature extraction to take one of the features of each image; thus the difference between one image and another is known, the application of this feature extraction will later produce feature values that can be used in the classification process. The feature extraction is color feature extraction, namely the color histogram, while the random forest is used for the classification process in this research.

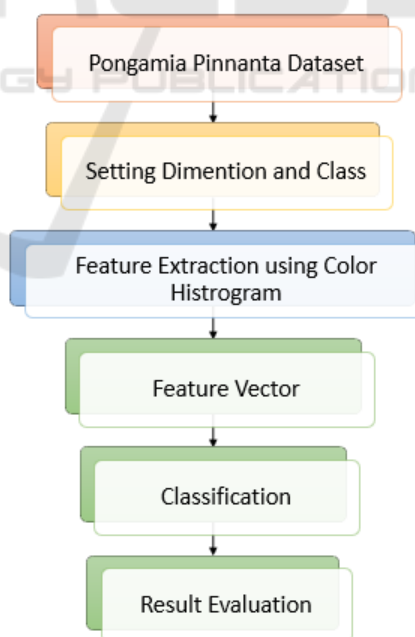




Figure 2: Research Method.

2.1 Dataset

The dataset used in this research is one of 12 types of plants found in the Collection of Different Categories of Leaf Images dataset. Each dataset is named from P0 to P11, and the Pongamia Pinnata (P7) dataset is used. All leaf images are collected from Shri Mata Vaishno Devi University, Katra. Shooting is done in a closed environment. This acquisition process is fully wi-fi enabled. All photos were taken using the built-in Nikon D5300 camera with performance times for JPEG shooting in single shot mode (seconds/frame, max resolution) = 0.58 and for RAW+JPEG = 0.63. Images were taken in jpg format with an 18-55mm lens with sRGB color representation, 24-bit depth, two resolution units, 1000-ISO, and no flash. Dataset details can be seen in Table 1:

Table 1: Pongamia Pinnata Dataset (Chouhan et al., 2019).

No	Citra Sample	Class	Total
1		Diseased	276
2		Healthy	322
Total Data			598

2.2 Feature Extraction

Feature Extraction or feature extraction is a feature taken by the score that will analyze for further processing. Feature extraction aims to find significant feature areas in images depending on their intrinsic characteristics and application. These regions can be defined in a global or local environment and distinguished by shape, texture, size, intensity, statistical properties, etc. (Verma and Salour, 2020)

2.3 Color Histogram

Color Histogram is a way of describing color content by counting each color that appears in an image. A color quantization process is needed for stains to develop faster because calculating the number of RGB color variations that occur takes a long time. Color quantization is the separation of color components into a certain range of distances (Beauty and Sari, 2019). A histogram can consist of 48 color ranges; each square defines a pixel. These areas represent the different intensity levels of each RGB component. Then the value of each bin is normalized by dividing

that value by the total number of pixels in the image (Simarmata et al., 2019). When specifying a color histogram, you can use the image's red, Green, and Blue (RGB) colors. This combination forms different colors, determining the value range (Gandhi and Lina, 2021).

The Color Histogram is one of the most effective and suitable color feature representations because the color space is clearly chosen and then displays the histogram concisely. In digital image processing, the RGB color space is generally used considering the advantages of concise representation, low computational complexity, and robustness to geometric changes (Zhou et al., 2018).

2.4 Random Forest

The RF comprises some decision trees, and each tree obtains its positional effect by utilizing a different classification. This method allows evaluation of the sampling allocation using the random sampling technique (Murugan et al., 2019) RF classification is an ensemble technique that continuously uses bootstrapping. Bootstrapping guarantees that whatever decision tree is in the random forest is different, lowering the RF variance. RF classification combines multiple decision trees for the final assessment; consequently, the RF classifier has strong generalizability. RF consistently outperforms all other classification techniques in terms of precision without difficulty with unbalanced and overfitting datasets (Balyan, 2022)

RF is based on constructing multiple decision trees, each of which functions as a classifier. In RF, each tree is sampled from the original data set to create sub-datasets. Sub-datasets are entered into each decision tree, and each decision tree outputs its results. The final decision result is determined by voting from all decision trees. Trees in RF don't use all the features to choose a form. Instead, some parts are randomly selected from all elements. Then the optimal amount is set from the randomly selected features. Because of this, the forest deviation usually increases slightly (relative to the variation of one non-random tree), but the variance also decreases, which can not only compensate for the increased deviation but also produce better results (Li et al., 2020)—the concept of tree decision calculated via the equation entropy value and information gain value. The following is the equation for the entropy value and the information gain value (Sandag, 2020).

$$Entropy(Y) = -\sum_i P(c|Y) \log_2 P(c|Y) \quad (1)$$

Note:

Y = Case Set

P(c—Y) = Proporsion Y'score to C class

$$Informationgain(Y,a) = Entropy(Y) - \sum_{vevalues(a)} \frac{Y_v}{Y_a} | Entropy(Y_v) \quad (2)$$

Information:

Values (a) = Possible values in a case set a

Y_v =Subclass of Y with class v, which is related to class a.

Y_a = All values corresponding to a

3 RESULT AND DISCUSSION

Based on research that has been done to classify the image of Pongamia pinnata leaves in the category of sick and healthy. The application of a color histogram can extract color features from a photo and represent the image as one-dimensional RGB, then form a histogram of an image. As in picture 3.

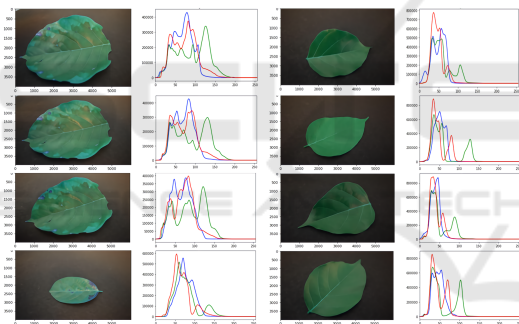


Figure 3: Extraction results of Pongamia Pinnata L.

Besides producing in the form of a histogram, this feature extraction also has different feature values in each image. The feature values resulting from feature extraction can be seen in table 2.

Table 2: Extraction results of Pongamia Pinnata L.

No	Leaf Sample	Feature Value
1	Leaf Healthy 1	774337
2	Leaf Healthy 2	891746
3	Leaf Healthy 3	891497
4	Leaf Healthy 4	856789
5	Leaf Diseased 1	375609
6	Leaf Diseased 2	344622
7	Leaf Diseased 3	398179
8	Leaf Diseased 4	593341

Table 2 shows the results of the extraction of sample

images in the form of feature values; from these results, the feature values produced in the sick class and the healthy class are different. The unhealthy type has the lowest score of 344622 and the highest score of 593341, while the therapeutic class has the lowest score of 774337 and the highest of 891746. Then classification is carried out from this feature value. The classification process is carried out using a random forest, with a data ratio of 8:2, meaning that 80% is used for training data, and 20% is used for test data. Apart from using RF, several other classification algorithms were tested as a comparison to see the performance of the different algorithms in classifying the Duan Pongamia image. The results of the test can be seen in Table 3.

Table 3: Machine Learning Algorithm Comparison.

Methods	Accuracy	Standard Deviation
RF	0,997917	0,006250
CART	0,991578	0,010316
LR	0,989583	0,016796
KNN	0,987367	0,016845
NB	0,985372	0,016289
LDA	0,910106	0,036019

Table 3 shows that RF produces the highest accuracy value reaching 0.997917, and the lowest accuracy is made from LDA, which equals 0.910106. The difference in accuracy results for the different methods can also be seen in Figure 4.

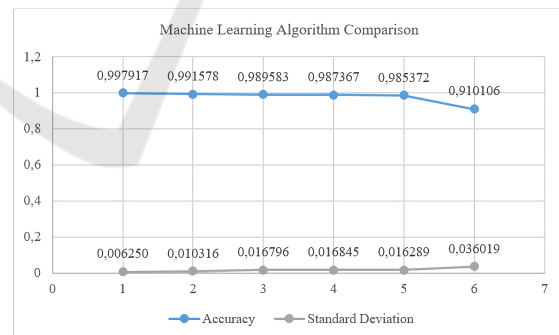


Figure 4: Algoritma Clarification Comparison.

3.1 Conclusion

Based on the results carried out on the Pongamia Pinnata dataset, which bring through the application of color histogram feature extraction and classification using a random forest which resulted in an accuracy performance of 99.79%. This study also compared several other classification algorithms such as Logistic Regression which produced an accuracy value of

98.95%, Linear Discriminant Analysis 91.01%, K-Nearest Neighbors 98.73%, Classification and Regression Tree 99.15% and Naïve Bayes 98.53%. The test results show that the Random Forest algorithm produces the highest accuracy value compared to several other methods.

As noted, the research does not aim to replace the plant disease diagnosis process carried out by experts but only seeks to complement it. The existence of machine learning methods can only predict with the level of accuracy by the accuracy obtained; still, testing through the laboratory is the most reliable way to diagnose plant diseases. However, the model implementation is used to help smallholders who need help to get a quick response from experts.

For further research, other feature extraction can be applied, which is not just one type; a feature extraction combination sees the difference in performance results in the extraction process. If there are larger data, it can be continued using deep learning and implemented in the system.

ACKNOWLEDGMENTS

The author would like to thank all parties who have supported this research process. The author also thanks Siddharth Singh Chouhan, Uday Pratap Singh, Ajay Kaul, and Sanjeev Jain for publishing the Pongamia Pinnata Leaf Image Data Repository

REFERENCES

- A. Aminah, Supriyanto, I. Z. and A. Suryani, P. J. S. S. B. B. B. C. M. P. L. (2017). Pierre from java island as biodiesel raw material. *J. Penelit. Has. Hutan*, 35(4):255-262.
- Agustiani, S., Arifin, Y., Junaidi, A., Wildah, S., and Mustopa, A. (2022). Klasifikasi penyakit daun padi menggunakan random forest dan color histogram.
- Balyan, A. (2022). A hybrid intrusion detection model using ega-pso and improved random forest method. *Sensors*, 22(16):1-20.
- Beauty, C. and Sari, Y. (2019). Temu kembali citra makanan menggunakan color histogram dan local binary pattern. *J. Pengemb. Teknol. Inf. dan Ilmu Komput*, 3(6):5514-5520.
- Chen, H. (2022). Establishment and analysis of a disease risk prediction model for the systemic lupus erythematosus with random forest. *Front. Immunol*, 13(November): 1-13.
- Chouhan, S., Singh, U., Kaul, A., and Jain, S. (2019). A data repository of leaf images: Practice towards plant conservation with plant pathology. *Conf. Inf. Syst. Comput. Networks, ISCON*, pages 700-707.
- Chouhan, S., Singh, U., Sharma, U., and Jain, S. (2021). Leaf disease segmentation and classification of jatropha curcas l. and pongamia pinnata l. biofuel plants using computer vision based approaches. *Meas. J. Int. Meas. Confed*, 171:108796.
- Dendang, B. and Suhaendah, E. (2017). Uji efektivitas insektisida terhadap hama maruca testulalis pada bibit malapari (pongamia pinnata (l.) pierre. *J. Pemuliaan Tanam. Hutan*, 11(2):123-130.
- Evans, M. (2021). Pongamia: Manfaat-manfaat potensial untuk restorasi dan bioenergi di indonesia. accessed Jan. 15, 2023).
- Gandhi, F. and Lina, L. (2021). Pengenalan jenis maker dengan metode color histogram dan euclidean distance. *J. Ilmu Komput. dan Sist. Inf*, 9(2):18.
- Khasanah, F. (2020). Pemanfaatan media sosial dan e-commerce sebagai media pemasaran dalam mendukung peluang usaha mandiri pada masa pandemi covid 19. *J. Sains Teknol. dalam Pemberdaya. Masy*, 1(1):51-62.
- Kusnandar, V. (2022). Indonesia impor minyak 30 juta ton pada januari-september 2022. accessed Jan. 15, 2023).
- Li, X., Chen, W., Zhang, Q., and Wu, L. (2020). Building auto-encoder intrusion detection system based on random forest feature selection. *Comput. Secur*, 95:101851.
- Maniyath, S. (2018). Plant disease detection using machine learning. *Conf. Des. Innov. 3Cs Comput. Commun. Control*, (April):41-45.
- Maretta, A. T. D., Devy, L., and Sulastrri (2016). Prosiding seminar nasional dan kongres perhimpunan agronomi Indonesia 2016. *Mult. Tunas Vitr. Satoimo (Colocasia esculenta Scott Var Antiq. pada Media MS dengan Penambahan 2iP, Glutamin, GA3, BAP, dan NAA*, (May):173-178.
- Mitra, S. (2021). A review on environmental and socio-economic perspectives of three promising biofuel plants jatropha curcas, pongamia pinnata and mesua ferrea. *Biomass and Bioenergy*, 151:106173.
- More, P., Agarwal, P., and Agarwal, P. (2019). Geminiviruses: Molecular biodiversity and global distribution in jatropha. *Physiol. Mol. Plant Pathol*, 108(May):101439.
- Murugan, A., Nair, S., and Kumar, K. (2019). Detection of skin cancer using svm, random forest and knn classifiers. *J. Med. Syst*, 43(8).
- Orte, F., Mira, J., Sánchez, M., and Solana, P. (2023-01). A random forest-based model for crypto asset forecasts in futures markets with out-of-sample prediction. *Res. Int. Bus. Financ*, 64:101829.
- Primajaya, A. and Sari, B. (2018). Random forest algorithm for prediction of precipitation. *Indones. J. Artif. Intell. Data Min*, 1(1):27-31.
- Purwanto, A., Wikantika, K., Deliar, A., and Darmawan, S. (2023). Decision tree and random forest classification algorithms for mangrove forest mapping in sembilang national park, Indonesia. *Remote Sens*, 15(1).
- Radhiana (2023). Strategi keberlanjutan pembangunan energi terbarukan jangka panjang indonesia: Kasus

- biomassa energi terbarukan di sektor pertanian, perkebunan dan kehutanan indonesia. *J. Serambi Eng*, VIII(1):4978-4990.
- Saadah, S. and Salsabila, H. (2021). Prediksi harga bitcoin menggunakan metode random forest. *J. Komput. Terap*, 7(1):24-32.
- Sandag, G. (2020). Prediksi rating aplikasi app store menggunakan algoritma random forest. *CogITo Smart J*, 6(2):167-178,.
- Saravanan, D. and Surendiran, J. (2019). Video data retrieval using image color histogram technique. *Int. J. Eng. Adv. Technol*, 8(6):2056-2060. Special Issue.
- Simarmata, S., Sari, Y., and Adinugroho, S. (2019). Klasifikasi citra makanan menggunakan algoritme learning vector quantization berdasarkan ekstraksi fitur color histogram dan gray level co-occurrence matrix. *J. Pengemb. Teknol. Inf. dan Ilmu Komput*, 3(3):2369-2378.
- Singh, A., Kumar, R., and Tiwari, A. (2023). Prediction of alzheimer's using random forest with radiomic features. *Comput. Syst. Sci. Eng*, 45(1):513-530,.
- Suita and Syamsuwida (2016). Pengaruh pengeringan terhadap viabilitas benih malapari. *J. Perbenihan Tanam. Hutan*, 4(1):9-16.
- Sutarman (2017). Dasar-dasar ilmu penyakit tanaman.
- Tamin, R. and Puri, S. (2020). Efektifitas fungi mikoriza arbuskula (fma) dan pupuk npk terhadap pertumbuhan bibit malapari (*Pongamia pinnata* (l.) pierre) pada tanah ultisol. *J. Ilm. Ilmu Terap. Univ. Jambi*, 4(1):50-58.
- Usharani, K., Naik, D., and Manjunatha, R. (2019). *Pongamia pinnata* (l.): Composition and advantages in agriculture: A review kv usharani, dhananjay naik and rl manjunatha. *J. Pharmacogn. Phytochem*, 8(3):2181-2187.
- Verma, N. and Salour, A. (2020). Feature extraction. *Stud. Syst. Decis. Control*, 256:121-173.
- Yankun, Y., Xiaoping, D., Wenbo, C., and Qiqige, W. (2021). A color histogram based large motion trend fusion algorithm for vehicle tracking. *IEEE Access*, 9:83394-83401.
- Zhou, J., Liu, X., Xu, T., Gan, J., and Liu, W. (2018). A new fusion approach for content based image retrieval with color histogram and local directional pattern. *Int. J. Mach. Learn. Cybern*, 9(4):677-689.