Journal of Artificial Intelligence and Engineering Applications

Website: https://ioinformatic.org/

15th February 2025. Vol. 4. No. 2; e-ISSN: 2808-4519

K-Means Algorithm to Improve Leaf Image Clustering Model for Rice Disease Early Detection

Gina Regiana Putri[1]*, Ade Irma Purnamasari[2], Agus Bahtiar[3], Edi Tohidi[4]

[1,2,3,4] *STMIK IKMI Cirebon*

ginaregianap8@gmail.com[1*], irma2974@yahoo.com[2], agusbahtiar038@gmail.com[3], editohidi00@gmail.com[4]

Abstract

This research aims to improve the accuracy of rice leaf image clustering in early disease detection using the K-Means algorithm. The approach used involves the Knowledge Discovery in Databases (KDD) method, which includes data selection, pre-processing, data transformation, data mining, evaluation, and presentation of results. The dataset used consists of images of healthy leaves and leaves infected with diseases such as Bacterial Leaf Blight, Brown Spot, and Leaf Smut. The images are processed through grayscale conversion, noise removal, size adjustment, and data augmentation. The K-Means algorithm is applied to cluster image features based on visual similarity. Evaluation results using Silhouette Score showed that the best clustering was obtained at K=2 with a score of 0.8340, resulting in two main clusters separating healthy and infected images. This study concludes that the K-Means algorithm is able to improve the efficiency and accuracy of rice disease detection, so that it can assist farmers in taking early preventive measures and increase agricultural productivity. This implementation shows significant potential in the development of smart agriculture technology.

Keywords: *K-Means; Early Detection; Rice disease; Image clustering; Silhouette score*

1. Introduction

Early detection of diseases in rice plants is essential to maintain agricultural productivity and food security, especially in agrarian countries like Indonesia [1]. Diseases such as Brown Spot, Leaf Smut, and Bacterial Leaf Blight often result in significant losses for farmers. Currently, disease detection methods still rely on visual observation, which is often less accurate and time-consuming. Therefore, a technology-based approach is needed that is able to provide faster and more accurate results. The K-Means algorithm offers a potential solution in clustering rice leaf images for automatic disease detection. Previous research has shown that methods such as Convolutional Neural Network (CNN) can provide high accuracy results in rice image classification. For example, CNN with MobileNet-V2 architecture achieved 97.56% accuracy [2], while Random Forest with Color Histogram extraction achieved 99.65% accuracy [3]. However, these methods have weaknesses, such as high computational requirements. This research aims to fill the gap by applying a lighter and more efficient K-Means algorithm. This study uses a dataset of rice leaf images consisting of two categories, namely healthy leaves and leaves infected with disease, collected from Jelegong Village. Using the Knowledge Discovery in Databases (KDD) approach, the leaf images were processed through the stages of data selection, pre-processing, data transformation, data mining, and result evaluation. Evaluation is done using the Silhouette Score metric to measure the quality of image clustering. This research is expected to provide a practical and efficient solution for detecting rice diseases, while making a significant contribution to the development of smart agriculture technology.

2. Research framework

2.1. Research framework

This research is designed to address the problem of early detection of disease in rice leaves using a technology-based approach. This research framework provides an overview of the systematic flow of each stage of the research conducted, starting from problem identification to achieving the final results. The following diagram explains the main steps in this research, including:

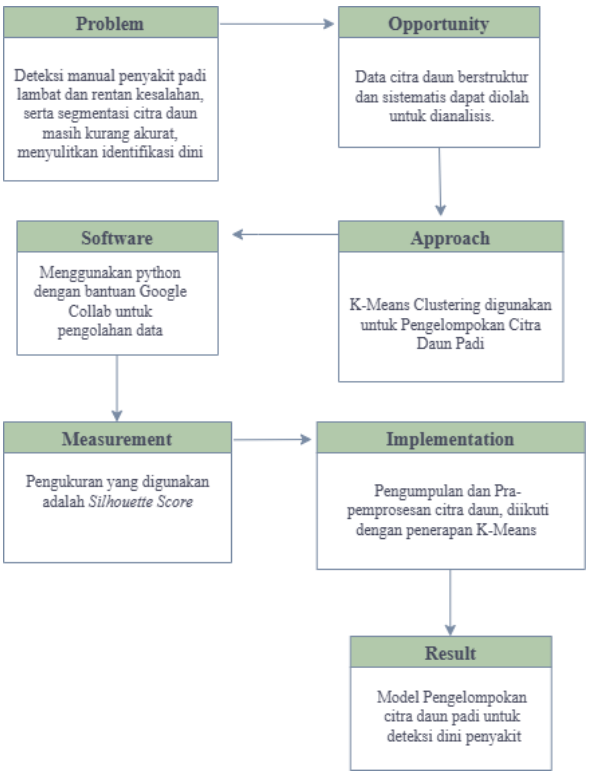

Fig. 1: Research framework

3. Literature review

3.1. Data mining

Data mining is a process in which processing data, the process of finding patterns or information based on the methods or techniques that will be used. The selection of the appropriate method used is highly dependent on the objectives and the overall KDD process [4].

3.2. Clustering

Clustering is a technique that searches for and groups data that have similar characteristics. The main focus of clustering is to group data or objects into clusters so that each cluster contains similar data [5].

3.3. K-means

K-Means is one of the techniques in data mining that performs the data analysis process without supervision, where the approach focuses on grouping based on similarities or existing patterns. Data that has similar characteristics will be grouped into one cluster, while data that has different characteristics will be placed in different clusters.[6]

4. Research methods

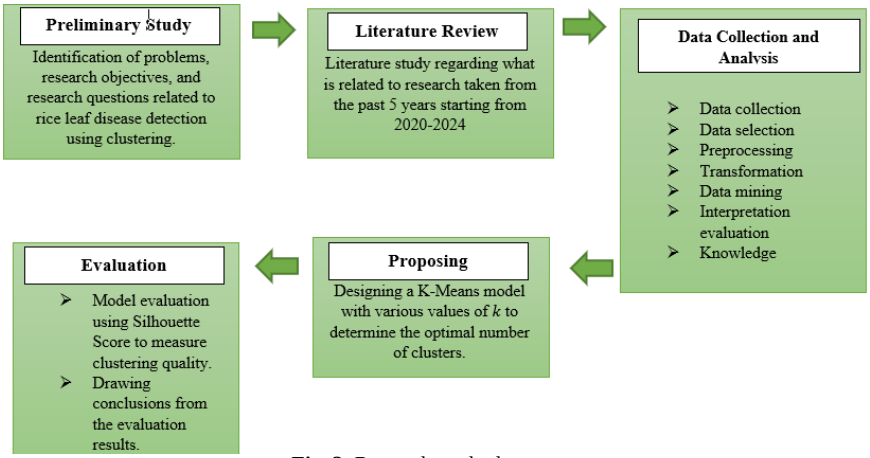

Fig. 2: Research methods

Table 1: Description of research method activies

Stages	activity	activity description
Preliminary Study	Identification of Problems	Identification of problems, research objectives, and research questions related to rice leaf disease detection using clustering.
	Research Objectives	Formulate the research objectives to be achieved.
	Research Questions	Asking specific research questions related to clustering methods
Literature Review	Literature Study	Conduct a literature study of related research from the last 5 years (2020-2024).
Data Collection and Analysis	Data Collection	Collecting rice leaf image data for research.
	Data Selection	Selecting relevant data for analysis.
	Preprocessing	Perform initial data processing, such as resizing or grayscale conversion and perform augmentation.
	Transformation	Transforming data into a format suitable for clustering analysis.
	Data Mining	Implementing the K-Means algorithm to cluster data.
	Interpretation and Evaluation	Analyze and evaluate clustering results.
	Knowledge	Drawing insights from the analysis results.
Proposing	Model Design	Designing K-Means algorithm model with various values of k to determine the optimal number of clusters.
Evaluation	Model Evaluation	Evaluating clustering quality using Silhouette Score.
	Drawing Conclusions	Summarize the results of the model evaluation to answer research questions.

5. Result and Discussion

5.1. Result

The results of the analysis and application of the K-Means algorithm performance in detecting disease types in rice can be seen as follows.

5.1.2. Data Selection

The dataset used consists of 160 images of rice leaves divided into four categories:

1. Healthy Leaves: Leaves without signs of disease.
2. Bacterial Leaf Blight: Leaves with watery spots caused by bacteria.
3. Brown Spot: Leaves with brown spots caused by fungi.
4. Leaf Smut: Leaves with small black spots caused by fungi.

The number of images in each category is 40, so the total dataset is 160 images.

5.1.3. Data Preprocessing

This stage aims to improve image quality and data consistency. The steps include:

Grayscale Conversion: Converting the image to grayscale to emphasize contrast.
Noise Reduction: Removing noise using Gaussian Blur.
Resizing: Resizing the image to standard resolution (128x128 pixels).
Data Augmentation: Additional variations on the dataset using rotation, shift, and zoom.

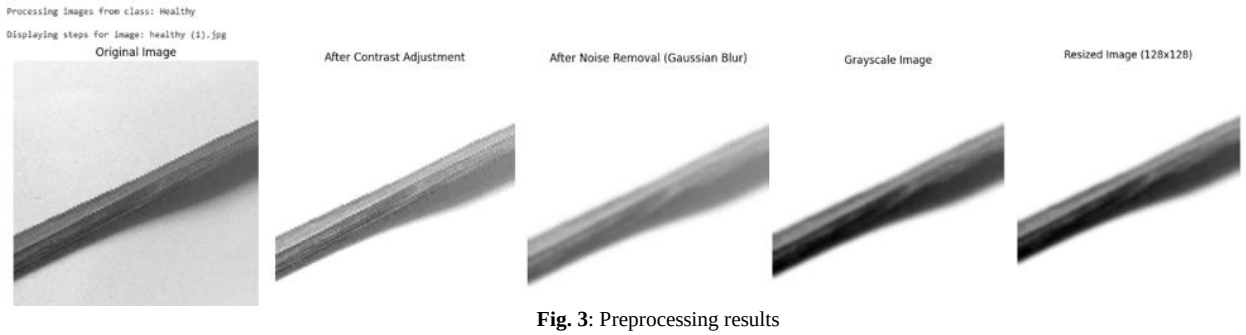

Fig. 3: Preprocessing results

5.1.4. Transformation data

Each image is converted into a feature histogram used for the clustering process. This histogram provides a numerical representation of the image suitable for the K-Means algorithm.

5.1.5. Data mining

At the data mining stage, researchers use the k-means algorithm which goes through the initialization stage and selecting the optimal number of clusters.

5.1.6 PCA Visualization

Dimension reduction using Principal Component Analysis (PCA) produces a two-dimensional graph that visualizes the image distribution based on clusters.

Fig. 4: PCA of k-means

5.1.7. Evaluation

In the evaluation stage, the cluster results will be tested using several K. The cluster with the highest Silhouette Score value or close to 1 is used as the best cluster. Researchers conducted experiments from cluster 2 to cluster 10 to determine the best cluster.

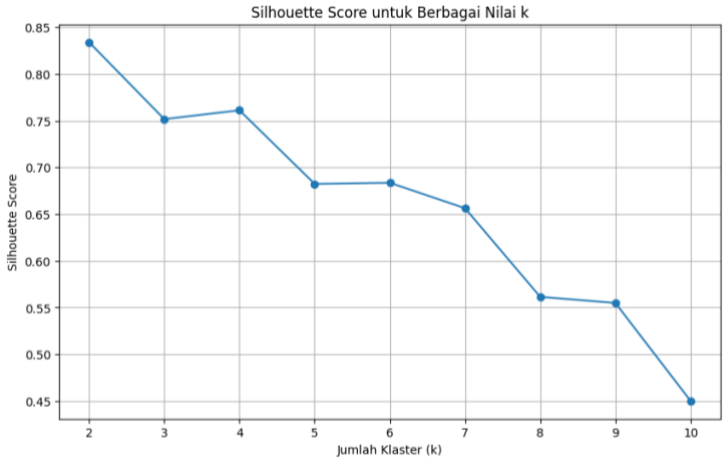

Fig. 5: Evaluation silhouette score

The K-Means algorithm was applied to group the images into clusters. The optimal number of clusters (**K = 2**) was determined using the **Silhouette Score**, which achieved a value of **0.8340**, indicating high-quality clustering.

- **Cluster 0**: Mixed healthy and infected images (e.g., 14 Bacterial Leaf Blight, 11 Brown Spot, 14 Healthy, 13 Leaf Smut).
- **Cluster 1**: Predominantly infected images (e.g., 26 Bacterial Leaf Blight, 29 Brown Spot, 26 Healthy, 27 Leaf Smut).

6. Conclusion and Suggestions

6.1. Conclusion

This study successfully applied the K-Means algorithm for clustering rice leaf images, achieving an optimal Silhouette Score of 0.8340 with K=2. The preprocessing steps enhanced dataset quality, enabling effective separation of healthy and diseased leaves. While efficient and lightweight, the model has limitations, including dataset diversity and lack of spatial feature consideration. Overall, K-Means demonstrates potential for early disease detection in smart agriculture systems.

6.2. Suggestions

The suggestions that can be conveyed from the results of this study are as follows:

1. Expand the dataset with samples from diverse regions for broader applicability.
2. Combine K-Means with advanced methods like DBSCAN or CNN for better accuracy.
3. Develop real-time mobile or web applications for farmers.
4. Extend the model to detect multiple rice diseases.
5. Focus on cost-effective solutions to support small-scale farmers.

References

[1] M. Deden Miftah Fauzi, T. Al Mudzakir, C. Emilia Sukmawati, and J. Indra, "Deteksi Jenis Penyakit Pada Tanaman Padi Menggunakan Yolo V5," *Media Online)*, vol. 5, no. 1, pp. 39–48, 2024, doi: 10.30865/klik.v5i1.2009.

[2] A. Julianto, A. Sunyoto, and W. W. Ferry, "OPTIMASI HYPERPARAMETER CONVOLUTIONAL NEURAL NETWORK UNTUK KLASIFIKASI PENYAKIT TANAMAN PADI," Dec. 2022. [Online]. Available: https://www.kaggle.com/tedisetiady/leaf-rice-dis-

[3] S. Agustiani, Y. Tajul Arifin, A. Junaidi, S. Khotimatul Wildah, and A. Mustopa, "Klasifikasi Penyakit Daun Padi menggunakan Random Forest dan Color Histogram," 2022. [Online]. Available: https://www.kaggle.com/vbookshelf/rice-leaf-

[4] A. A. Arrosyad, A. I. Purnamasari, and I. Ali, "IMPLEMENTASI ALGORITMA K-MEANS CLUSTERING UNTUK ANALISIS PERSEBARAN UMKM DI JAWA BARAT," Jun. 2024.

[5] A. Maulana, K. Nur Akbar, and Nurahman, "Penerapan Clustering Menggunakan Algoritma K-Means Sebagai Analisis Produksi Komoditas Perikanan Provinsi di Indonesia," Sep. 2021.

[6] I. R. Nr, E. Prasetyowati, and B. Said, "Penerapan Citra Berbasis K-Means Clustering untuk Mendeteksi Penyakit Bulai Pada Komoditas Jagung Madura," May 2023. [Online]. Available: https://jurnal.umj.ac.id/index.php/just-it/index