



Image Classification of Rice Leaf Diseases with KNN Based Model using Stratified-KCV

Tasya Anisah Rizqi¹, Anwar Fitrianto², Kusman Sadik³

^{1,2,3}Department of Statistics, School of Data Science, Mathematics, and Informatics, IPB University, Bogor, Indonesia

¹tasya.anisah.rizqi@apps.ipb.ac.id, ²anwarstat@gmail.com, ³kusmans@apps.ipb.ac.id

Abstract

Rice is a staple food for people in the world, especially Indonesia. The rice harvest decreased in 2023, reducing harvest productivity and causing losses for farmers. Rice cultivation is often affected by diseases that hinder rice harvests. SKCV is a resampling method that performs more accurately because it can ensure that class frequencies are maintained. RGB and VGG16 are image processing methods that extract images into numerics. RGB image extraction is done by taking the average value of the red, green, and blue layers while VGG16 image extraction is done by taking the value of visual pattern features such as edges, textures, and object shapes. In this study, rice leaf diseases were classified using KNN-based models, including KNN, WKNN, CDNN, and ECDNN. This classification was performed to determine which method had better performance using SKCV and comparing the results of RGB and VGG16 image extraction. This classification also produces a comparison of SKCV and KCV results to determine the best resampling performance. The results of the analysis that have been carried out show that the ECDNN method produces the highest accuracy of 81.20% in classifying rice leaf diseases using SKCV with VGG16 extraction followed by CDNN and WKNN each at 68.80%, and KNN at 56.20% while RGB extraction only produces an accuracy of 43.8% using ECDNN and CDNN, 56.20% using WKNN, and 50% using KNN. The results of this rice leaf diseases classification analysis are expected to help farmers in increasing rice production in Indonesia.

Keywords: image classification; rice leaf diseases; rgb; stratified k-fold cross validation; vgg16

How to Cite: T. A. Rizqi, Anwar Fitrianto, and Kusman Sadik, "Image Classification of Rice Leaf Diseases with KNN Based Model using Stratified-KCV", *J. RESTI (Rekayasa Sist. Teknol. Inf.)*, vol. 9, no. 5, pp. 1056 - 1065, Oct. 2025.]

Permalink/DOI: <https://doi.org/10.29207/resti.v9i5.6590>

Received: April 27, 2025

Accepted: September 20, 2025

Available Online: [October 12, 2025]

*This is an open-access article under the CC BY 4.0 License
Published by Ikatan Ahli Informatika Indonesia*

1. Introduction

Rice is a cultivated plant that is a source of carbohydrates with nutritional content that is greatly needed by some of the world's population, especially Indonesia. It is known that the population of Indonesia in 2023 increased by 2,922,400 people compared to 2022 [1], while rice production in 2023 reached 53.98 million tons of GKG (Dry Milled Grain) which decreased compared to 2022, which was 54.75 million tons of GKG [2]. This results in an imbalance between the population and the need for rice in Indonesia. It is also known that the incidence of rice plant pest attacks in 2023 was 106,092.8 hectares or 1.04% of the rice harvest area of 10.21 million hectares [3]. In this case, classification needs to be done to overcome disease constraints in rice plants so that the stability of the rice harvest can be maintained by farmers.

Classification of diseases in rice plants can be done using images of diseases in rice plants. Images cannot

be processed directly by computers so that analog images must first be converted into digital images [4]. Research on anemia detection using RGB images obtained optimal accuracy [5]. Research on the classification of rice plant diseases images produced the highest accuracy in identifying disease types using VGG16 feature extraction, which was 93.11% [6]. Therefore, diseases in rice plants can be classified by image extraction.

Classification is a part of statistical machine learning commonly used in data science by systematically grouping data into certain classes or groups based on similar characteristics [7]. K-Nearest Neighbor (KNN) is one of the classification methods that is widely used because of its simplicity and works quite easily with multiclass data sets. The KNN method improves accuracy in image classification [8]. Research that trains several KNN-based models, namely Centroid Displacement-based KNN (CDNN), Radius NN

(RNN), KNN, Weighted KNN (WKNN), Nearest Centroid (NC) Neighbors, and Ensemble CDNN (ECDNN) which is a method proposed by researchers produces 4 methods with the highest average F1-score, namely ECDNN, CDNN, WKNN, and KNN, which are 81.25%, 80.52%, 79.93%, and 79.16% respectively [9].

The classification process requires resampling to obtain an accurate estimate of model performance. Cross Validation (CV) is a resampling method used by dividing the training data into several parts to obtain a prediction of the error rate in each part [10]. Stratified K-Fold Cross Validation (SKCV) is a part of CV that can ensure that the relative class frequency is maintained effectively when using stratified sampling compared to random sampling [11]. Research on the detection of long-distance homologous proteins using the KNN method produced the most superior accuracy with an SKCV of 98.9% [12]. Therefore, SKCV can be a resampling method that will produce a model performance evaluation with more accurate accuracy compared to the K-Fold Cross Validation (KCV) method.

Research on the classification of various plant diseases using VGG16 succeeded in achieving the highest accuracy of 97.82% [13]. Research on VGG16 model classification of rice diseases with disease classes blast,

bacterial blight, brown spot, narrow brown spot, and bacterial leaf streak in Thailand successfully achieved an accuracy of 90.77% [14]. Research on KNN classification of banana leaf diseases with disease classes yellow sigatoka, black sigatoka, and banana mosaic in Malaysia successfully achieved an accuracy of 77.08% [15]. These studies are relevant to the use of machine learning methods in the agricultural sector in Southeast Asia.

Based on the background that has been explained, this study conducted a study by proposing the implementation of Stratified K-Fold Cross Validation on K-Nearest Neighbor-based models, namely ECDNN, CDNN, WKNN, and KNN for the classification of rice leaf disease images in VGG16 image processing which proved to be more accurate than the KNN, WKNN, and CDNN methods. This study also shows that the use of VGG16 image extraction significantly improves classification performance compared to RGB image extraction.

2. Methods

Classification of rice leaf diseases through several stages of analysis is presented in Figure 1.

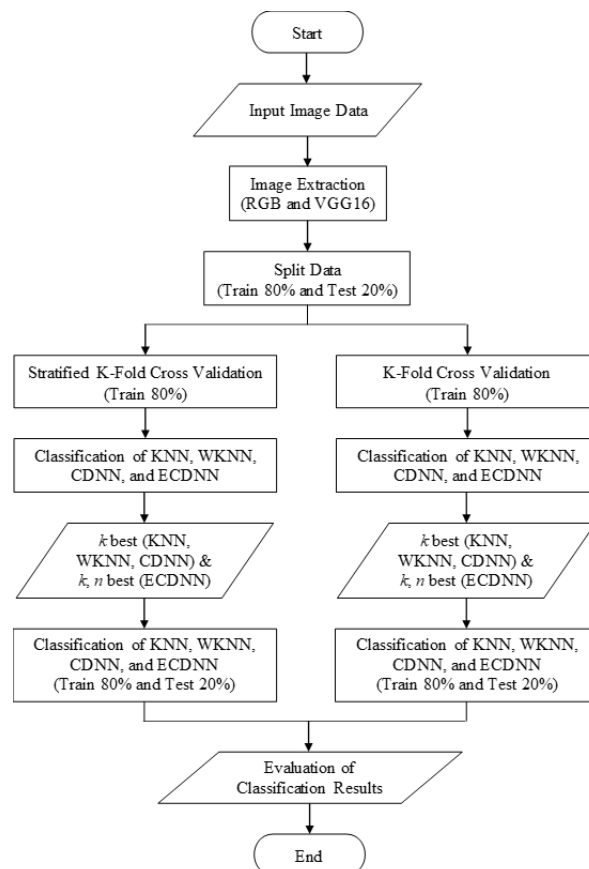


Figure 1. Flowchart of Analysis Stages

2.1 Digital Image Processing RGB

Manuscripts Digital images are numeric data with the smallest unit called pixels [16]. RGB (Red, Green, Blue) images are a representation of the color of an object that combines the three basic colors, namely red, green, and blue [17]. An illustration of the pixel intensity value in an RGB image is presented in Figure 2 [18].

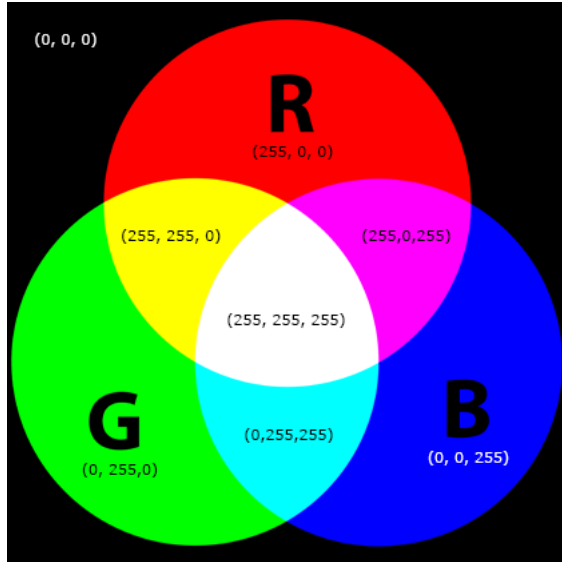


Figure 2. Illustration of the Pixel Intensity Value of an RGB Image

2.2 Digital Image Processing VGG16

Digital image processing is an image analysis process that aims to extract some useful information from it [19]. VGG16 is an architecture for feature extraction developed by [20]. VGG16 utilizes the depth of the system which is the main factor for better performance [21]. VGG16 uses a series of connecting layers (max pooling) and convolutional layers (convolution) that represent general concepts in image processing [22]. An illustration of the VGG16 architecture is presented in Figure 3 [23].

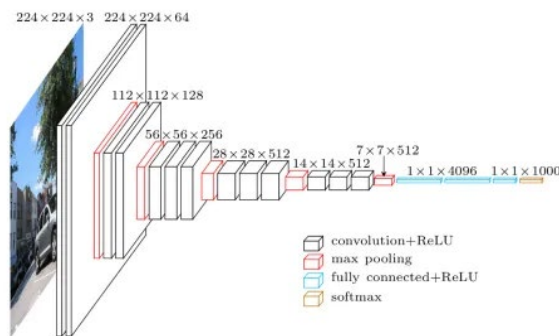


Figure 3. Illustration of VGG16 Architecture

2.3. Stratified K-Fold Cross Validation

Stratified K-Fold Cross Validation (SKCV) is a Cross Validation used with a stratified sampling method on train and test data that has almost the same proportion of features as the original data set [11]. Each subsample

taken will be used as data validation for testing a classification model with a process of k . The 5-fold cross validation procedure is presented in Figure 4.

Each fold in the train and test data is subjected to stratified sampling. An illustration of the use of random sampling and stratified sampling is presented in Figure 5.

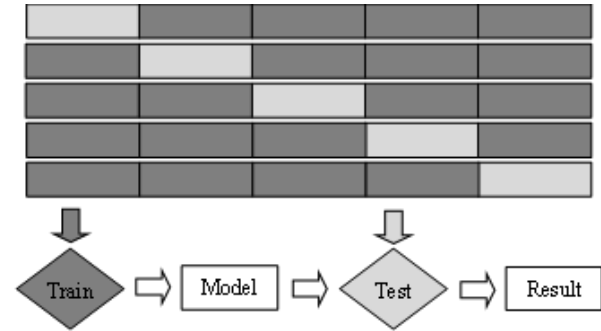


Figure 4. 5-Fold Cross Validation Procedure

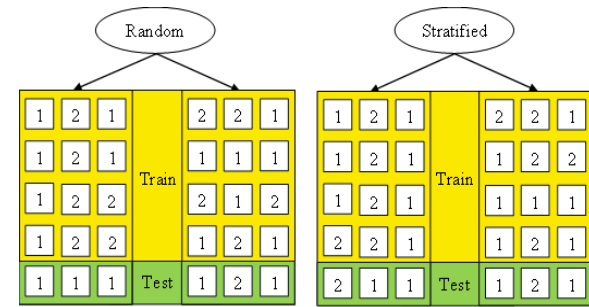


Figure 5. Illustration of Random Sampling and Stratified Sampling

2.4 K-Nearest Neighbor

K-Nearest Neighbor (KNN) was first introduced by [24] which is a classification method based on the consideration of the closest class of an object. The distance function is used to determine the distance between new data and existing data including Euclidean, Manhattan, and Minkowski [25]. KNN will determine the classification based on the majority class of the data that has the closest distance [26].

The first stage in the KNN classification process is to determine the parameter k , which is the number of nearest neighbors with odd k between 5 and 25. The second stage is to calculate the distance between new data and each data using the Euclidean distance using Formula 1.

$$d = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} \quad (1)$$

d is the Euclidean distance, x_i is the i th test data, and y_i is the i th train data.

The third stage is to sort the data based on the smallest Euclidean distance. The last stage is to determine the class of new data based on the majority class of the closest neighbors [27].

2.5 Weighted K-Nearest Neighbor

Weighted K-Nearest Neighbor (WKNN) is a development of the KNN classification method that aims to reduce the error rate in KNN. The distance in KNN causes problems in the contribution of selecting the k nearest neighbors. Therefore, the WKNN method emerged which can make the classification better. The determination of the weight in WKNN is taken from the inverse of the distance and squaring the inverse of the distance.

The first stage in the WKNN classification process is to determine the parameter k , which is the number of nearest neighbors with odd k between 5 and 25. The second stage is to calculate the distance between new data and each data using the Euclidean distance. The third stage is to sort the data based on the smallest Euclidean distance. The fourth stage is to select the top k data and then calculate the weight using Formula 2.

$$w = \frac{1}{d^2} \quad (2)$$

w is the weight.

The last stage is to add the same class weights then the new class is determined based on the total weight of the nearest neighboring class [28].

2.5 Centroid Displacement-based K-Nearest Neighbor

Centroid Displacement-based K-Nearest Neighbor (CDNN) is a classification method proposed by [29] that groups the nearest neighbors into separate sets of the same class label. The centroid of each set and the centroid displacement will be calculated if the testing data is included in the set. The testing data will be included in the set that has the minimum centroid displacement.

The first stage in the CDNN classification process is to determine the parameter k , which is the number of nearest neighbors with odd k between 5 and 25. The second stage is to calculate the distance between new data and each data using the Euclidean distance. The third stage is to sort the data based on the smallest Euclidean distance. The fourth stage is to select the top k data as the nearest neighbors. The fifth step is to divide the data into separate sets containing neighbors with the same class label. The sixth step is to calculate the centroid for each set using Formula 3.

$$C = \frac{1}{n} \sum_{i=1}^n x_i \quad (3)$$

C is centroid.

The seventh step is to calculate the new centroid for each set if new data x is entered into the set using Formula 4.

$$C_p = \frac{1}{n+1} ((\sum_{i=1}^n x_i) + x) \quad (4)$$

C_p is the new centroid and x is the new data and x is the new data.

The eighth step is to calculate the centroid displacement for each set using Formula 5.

$$D = \sqrt{\sum_{i=1}^n (C_i - C_{p_i})^2} \quad (5)$$

D is the displacement centroid, C_i is the i th centroid, and C_{p_i} is the i th new centroid.

The last step is to determine the class based on the set with the smallest centroid displacement [29].

2.6 Ensemble Centroid Displacement-based K-Nearest Neighbor

Ensemble Centroid Displacement-based K-Nearest Neighbor (ECDNN) is a classification method that combines KNN and CDNN based on the homogeneity of the nearest neighbors. The KNN method is basically used for high-confidence predictions, while the CDNN method is used for low-confidence predictions. Therefore, KNN or CDNN is applied to produce increased accuracy and be more efficient in terms of computation, where the use of ensembles aims to combine several models into a more robust and accurate model [30].

The first stage in the ECDNN classification process is to determine the parameter k , with odd k between 5 and 25. The second stage is to determine the parameter n , which is the number of nearest neighbors based on class homogeneity with $n < k$. The third stages is to calculate the distance between new data and each data using the Euclidean distance. The fourth stage is to determine the n nearest neighbors based on the Euclidean distance and group neighbors with the same class and check for homogeneity. The fifth step is to determine the k nearest neighbors based on the Euclidean distance if not homogeneous and group neighbors with the same class. The sixth stage is to calculate the centroid for each set of k nearest neighbors using formula 3. The seventh stage is to calculate the new centroid for each set if new data x is entered into the set using formula 4. The eighth stage is to calculate the centroid displacement using formula 5. The last stage is to determine the class based on the set with the smallest centroid displacement [9].

2.6 Multiclass Confusion Matrix

Multiclass confusion matrix is a visualization of the performance results of a method from a cross table that records the number of occurrences of n raters. The $n \times n$ confusion matrix with different numbers of classes shows the actual and predicted classifications [31].

The confusion matrix helps in determining the reliability of the model and the effectiveness of the model in making accurate predictions [32]. An illustration of the multiclass confusion matrix is presented in Table 1.

Table 1. Multiclass Confuion Matrix

		Predicted			
		Class 1	Class 2	...	Class n
Actual	Class 1	C_{11}	C_{12}	...	C_{1n}
	Class 2	C_{21}	C_{22}	...	C_{2n}
	...	...	...	...	...
	Class n	C_{n1}	C_{n2}	...	C_{nn}

The performance of a method is seen from the accuracy values of the multiclass confusion matrix which can be obtained using Formulas 6 - 9 [33].

$$Precision_i = \frac{C_{ii}}{C_{ii} + \sum_{j \neq i} C_{ji}} \quad (6)$$

$$Recall_i = \frac{C_{ii}}{C_{ii} + \sum_{j \neq i} C_{ij}} \quad (7)$$

$$F1 - score_i = 2 \times \frac{Precision_i \times Recall_i}{Precision_i + Recall_i} \quad (8)$$

$$Accuracy = \frac{\sum_{i=1}^n C_{ii}}{\text{jumlah data masukan}} \quad (9)$$

i has the value 1, 2, ..., n and j has the value 2, 3, ..., n . C_{ii} is the number of predictions of class i on actual class i , C_{ij} is the number of predictions of class i on actual class j , and C_{ji} is the number of predictions of class j on actual class i .

2.7 Data and Data Sources

The data used are primary data in the form of images of rice leaves taken directly in the rice fields of Rempoah Village, Central Java in 2024. The data was obtained by conducting field documentation using an OPPO A31 cellphone with a resolution of 12 MP (MegaPixel) on leaves affected by the disease with a white background stored in .png file format as many as 80 images.

The data collection criteria are based on rice leaves that show typical symptoms of three main types of diseases, namely Bacterial Leaf Blight, Bacterial Leaf Streak, and Brown Leaf Spot. In addition, only images with good quality are used such as sufficient lighting and clearly visible symptoms to obtain accurate results in the image extraction and classification process. A list of data descriptions used in this study is presented in Table 2.

Table 2. List of Data Descriptions

Class	Total
Bacterial Leaf Blight (BLB)	33
Bacterial Leaf Streak (BLS)	20
Brown Leaf Spot (BLT)	27

BLB is a disease that can reduce yields by up to 60% in severe conditions [34]. One of the image data of rice leaves attacked by BLB disease is presented in Figure 6.

BLB disease has symptoms which usually start from the tip of the leaf, then spread downwards forming yellowish to dark brown spots which eventually cause the leaves to dry out.



Figure 6. Image of Rice Leaves Infected with BLB Disease

Rice seeds can be a major source of Bacterial Leaf Streak (BLS) infection [35]. Therefore, control in the use of healthy seeds is highly recommended. One of the image data of rice leaves attacked by BLS disease is presented in Figure 7.



Figure 7. Image of Rice Leaves Infected with BLS Disease

The main symptom of BLS is the appearance of narrow yellow to brown lines which then enlarge and form transparent spots when humidity is high.

The fungus in Brown Leaf Spot (BLT) disease is a common pathogen in the dry season or in less fertile soil. This disease can be controlled with balanced fertilization [36]. One of the image data of rice leaves attacked by BLT disease is presented in Figure 8.

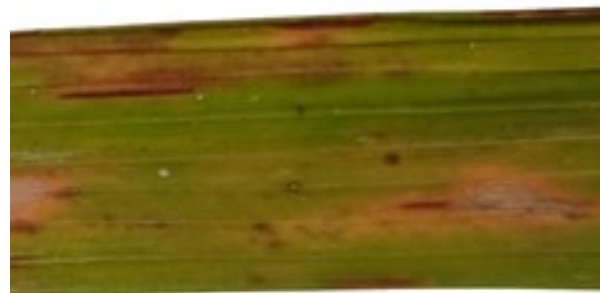


Figure 8. Image of Rice Leaves Infected with BLT Disease

BLT disease has symptoms characterized by the appearance of small brown or dark brown spots that can enlarge and form circles with dark edges and a grayish center.

Based on Table 2, the data variables used in this study using RGB image extraction are presented in Table 3.

Table 3. List of Data Variables with RGB Image Extraction

Variable	Description
Y	Class of rice leaf diseases
X_1	Average value of red layer RGB image extraction
X_2	Average value of green layer RGB image extraction
X_3	Average value of blue layer RGB image extraction

Data variables using VGG16 image extraction are presented in Table 4.

Table 4. List of Data Variables with VGG16 Image Extraction

Variable	Description
Y	Class of rice leaf diseases
X	Visual pattern features (edges, textures, and shapes of objects)

3. Results and Discussions

3.1 Data Exploration

Rice leaf diseases data in the form of images were subjected to RGB and VGG16 image extraction. Descriptive analysis of RGB image extraction of rice leaf diseases is presented in Table 5.

Table 5. Descriptive Analysis of RGB Image Extraction

Class	Red		Green		Blue	
	Avg	Std	Avg	Std	Avg	Std
BLB	0.791	0.049	0.759	0.034	0.668	0.049
BLS	0.782	0.065	0.749	0.043	0.653	0.073
BLT	0.739	0.081	0.755	0.058	0.647	0.082

Based on Table 5, the highest average value of the red and blue layers is obtained in BLB disease, while the lowest value is in BLT disease. This indicates that the red and blue layers tend to be stronger in BLB disease than in BLT disease. The highest average value of the green layer is in BLB disease, while the lowest value is in BLS disease. This indicates that the green layer tends to be stronger in BLB disease than in BLS disease. The largest data distribution in the red, green, and blue layers in BLT disease indicates that the color intensity is more varied than other diseases. Descriptive analysis of VGG16 image extraction of rice leaf diseases is presented in Table 6.

Table 6. Descriptive Analysis of VGG16 Image Extraction

Descriptive Statistics	BLB	BLS	BLT
Average	0.1034	0.0993	0.1183
Standard Deviation	0.1875	0.1934	0.2105

Based on Table 6, the highest average feature value is obtained in BLT disease, while the lowest value is in BLS disease. This indicates that the texture or visual pattern in BLT disease tends to be more contrasting than other diseases. The largest data distribution in the features in BLT disease indicates that the visual pattern in BLT disease is quite diverse, while the smallest data distribution in BLB disease indicates the consistency of the visual pattern in BLB disease.

3.2 Stratified K-Fold Cross Validation

Before classifying rice leaf diseases, it is necessary to divide the data first to be able to produce a classification accuracy value. Rice leaf diseases data was divided into

80% train data and 20% test data by stratification. The proportion of train data and test data by stratification is presented in Table 7.

Table 7. Proportion of Train and Test Data for Rice Leaf Diseases

Split Data	BLB	BLS	BLT
Train	26	16	22
Test	7	4	5

Based on Table 7, the proportion of data in the train data and test data is divided proportionally in each class of rice leaf diseases. After that, SKCV was carried out on the rice leaf disease train data as much as 5 folds with each fold divided into 80% train data and 20% test data. The proportion of data with SKCV on the 5-fold train data is presented in Table 8. Based on Table 8, the proportion of data in the train data and test data for each fold is divided proportionally in each class of rice leaf disease so that it can increase the accuracy of the performance of a method by ensuring that each disease is represented in the train data and test data.

Table 8. Proportion of Train Data with SKCV in Rice Leaf Diseases

Fold	Split Data	BLB	BLS	BLT
1	Train	21	13	17
	Test	5	3	5
2	Train	21	13	17
	Test	5	3	5
3	Train	21	12	18
	Test	5	4	4
4	Train	20	13	18
	Test	6	3	4
5	Train	21	13	18
	Test	5	3	4

3.4 Model Accuracy Using RGB and VGG16

RGB and VGG16 are digital image processing that converts image data into numeric data to perform classification with RGB being color-based extraction while VGG16 is feature-based extraction. At this extraction stage, a comparison of accuracy values with RGB and VGG16 image extraction is carried out. The comparison graph of accuracy values with RGB and VGG16 image extraction in rice leaf disease classification by applying SKCV to the train data is presented in Figure 9.

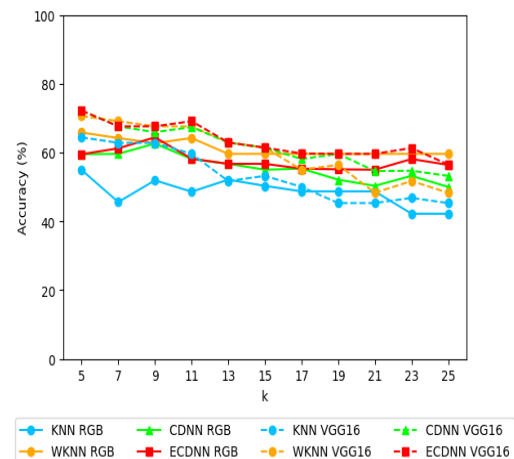


Figure 9. Accuracy Values of RGB and VGG16 on Train Data

Based on Figure 9, the highest average accuracy value of VGG16 image extraction was obtained by the ECDNN method, which was 63.41%, while the highest average accuracy value of RGB image extraction was obtained by the WKNN method, which was 61.28%. Therefore, it can be concluded that VGG16 image extraction using the ECDNN method is a good digital image processing for classifying rice leaf diseases.

3.5 Analysis Results Using KNN-Based Models

Based on Figure 9, the highest accuracy value is obtained in the ECDNN method with VGG16 image

extraction when $k = 5$ and $n = 2$, so that ECDNN classification is carried out with the best k and n on the train and test data. The highest accuracy value in the KNN method with RGB and VGG16 image extraction is obtained when $k = 5$, the WKNN method with RGB and VGG16 image extraction is obtained when $k = 5$. And the CDNN method with RGB image extraction is obtained when $k = 9$ and VGG16 when $k = 5$. The results of the analysis obtained from the multiclass confusion matrix on the KNN-based model with SKCV are presented in Table 9.

Table 9. Results of KNN-Based Model Analysis with SKCV

Methods	Evaluation	RGB			VGG16		
		BLB	BLS	BLT	BLB	BLS	BLT
KNN	Best k	5			5		
	Precision (%)	50	0	100	58.3	0	100
	Recall (%)	85.7	0	40	100	0	40
	F1-score (%)	63.2	0	57.1	73.7	0	57.1
	Accuracy (%)	50			56.2		
	Time (seconds)	0.003			0.002		
WKNN	Best k	5			5		
	Precision (%)	60	33.3	66.7	63.6	50	100
	Recall (%)	85.7	25	40	100	25	60
	F1-score (%)	70.6	28.6	50	77.8	33.3	75
	Accuracy (%)	56.2			68.8		
	Time (seconds)	0.011			0.002		
CDNN	Best k	9			5		
	Precision (%)	41.7	0	100	63.6	50	100
	Recall (%)	71.4	0	40	100	25	60
	F1-score (%)	52.6	0	57.1	77.8	33.3	75
	Accuracy (%)	43.8			68.8		
	Time (seconds)	0.001			0.001		
ECDNN	Best k	9			5		
	Best n	2			2		
	Precision (%)	45.5	0	100	70	100	100
	Recall (%)	71.4	0	40	100	50	80
	F1-score (%)	55.6	0	57.1	82.4	66.7	88.9
	Accuracy (%)	43.8			81.2		
	Time (seconds)	0.001			0.001		

Based on Table 9, the highest accuracy value was obtained in the ECDNN classification using VGG16 image extraction with the best k and n of 81.2% and also obtained the fastest and most efficient training time of 0.001 seconds. Therefore, the ECDNN method with VGG16 image extraction is a viable choice to be applied in real systems that require high performance and short response time.

The lowest accuracy value in the CDNN and ECDNN classification using RGB image extraction with the best k and n was 43.8%. It can be concluded that the ECDNN method produces the highest accuracy in VGG16 image extraction but produces the lowest accuracy in RGB image extraction and all methods produce better accuracy in VGG16 image extraction compared to RGB image extraction.

The best method in rice leaf diseases classification, namely ECDNN with VGG16 image extraction, was evaluated for method performance by analysis using the ROC-AUC metric presented in Figure 10.

Based on Figure 10, it can be interpreted that the ROC curve of the ECDNN method shows good and varied performance in distinguishing each class. The BLT class has the best performance with an AUC value of 0.982, which shows an almost perfect ability to distinguish one class from another. Meanwhile, the BLB class also shows very good performance with an AUC value of 0.929, indicating that the method is able to classify this class with high accuracy. The BLS class has a lower AUC value, which is 0.792, which is still quite good but shows that the model's ability to distinguish this class is not as strong as the other two classes.

The confusion matrix results obtained from the ECDNN classification analysis are presented in Figure 11.

Based on Figure 11, the actual BLB disease data was 100% successfully recognized correctly and the predicted data was only 70% predicted correctly. The actual BLS disease data was only 50% successfully recognized correctly and the predicted data was 100% predicted correctly. The actual BLT disease data was

only 80% successfully recognized correctly and the predicted data was 100% predicted correctly.

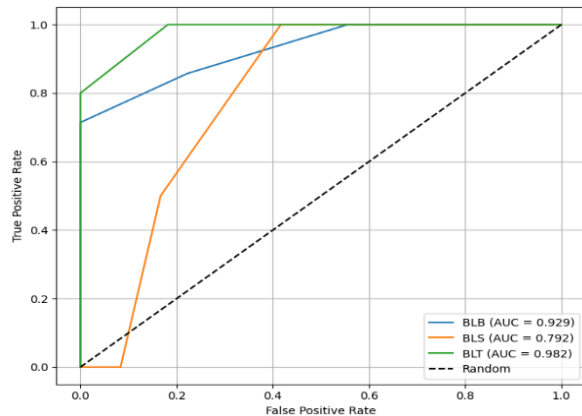


Figure 10. ROC-AUC Metric of ECDNN

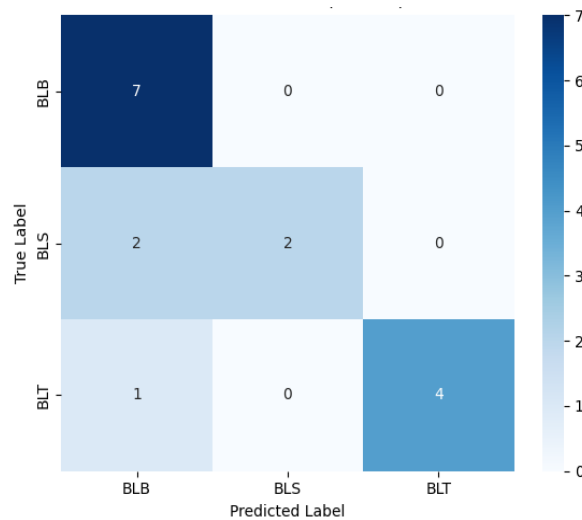


Figure 11. Confusion Matrix of ECDNN

3.6 Model Accuracy Using SKCV and KCV

Based on the results of the classification analysis with RGB and VGG16 image extraction, the highest accuracy was obtained in the ECDNN classification with VGG16 image extraction of 81.2%. The comparison graph of classification accuracy values with RGB and VGG16 image extraction using SKCV is presented in Figure 12.

The accuracy values of the KNN, WKNN, CDNN, and ECDNN classifications using SKCV need to be compared with KCV to obtain the accuracy of a method's performance. The comparison graph of accuracy values using SKCV and KCV with VGG16 extraction is presented in Figure 13.

Based on Figure 13, it can be interpreted that the ECDNN classification using SKCV with VGG16 image extraction produces the highest accuracy of 81.2%. Therefore, it can be concluded that SKCV is the best resampling compared to KCV with the best ECDNN method.

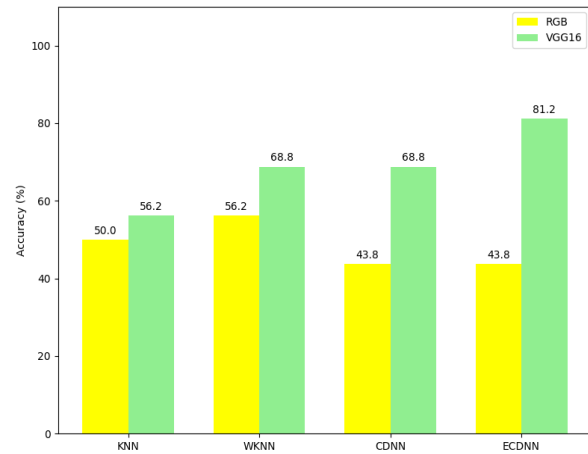


Figure 12. RGB and VGG16 Accuracy Value on SKCV

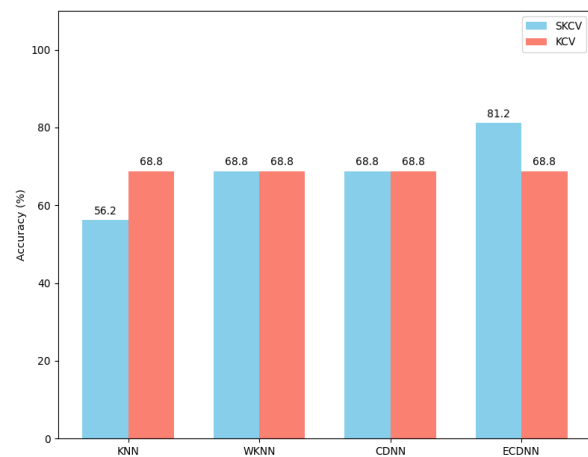


Figure 13. SKCV and KCV Accuracy Values with VGG16

3.7 Computational Cost

The analysis conducted in this study uses the Python programming language through the Google Colaboratory platform, which provides a cloud computing environment with general specifications in the form of an Intel(R) Xeon(R) CPU @ 2.20GHz processor, and around 12 GB of RAM. Computation is carried out without using a GPU, because the model used is lightweight and not based on deep learning.

The experiment was accessed using a device in the form of a Lenovo Ideapad D330 laptop with 4 GB of RAM and an Intel Celeron processor, so Google Colaboratory was chosen to avoid local limitations and ensure smooth execution.

4. Conclusions

Rice leaf diseases classification obtained the highest accuracy value in ECDNN classification using SKCV with VGG16 image extraction, which was 81.2% with an average precision of 90%, an average recall of 76.7%, and an average f1-score of 79.3%. It was obtained that from all the actual data obtained, 76.7% were successfully recognized correctly and the predicted data obtained were 90% predicted correctly with 79.3% indicating a fairly good method. It can be

concluded that SKCV is a very good resampling to obtain more accurate method performance than KCV and VGG16 image extraction is an image extraction that obtains better accuracy than RGB image extraction on rice leaf diseases images. The analysis results also obtained that the ECDNN method showed a very fast and efficient training time, making it a viable choice for application in a real rice leaf diseases classification system that requires high performance and short response time. Therefore, farmers can utilize the results of the rice leaf diseases classification analysis with an accuracy of 81.2% so that rice production in Indonesia can increase. This problem can be developed using Repeated Stratified K-Fold Cross Validation so that it can produce better accuracy. This classification analysis can also be done on other image data with more data using other convolution image extraction so that the accuracy value can be compared.

Acknowledgements

This research is supported by the Department of Statistics, School of Data Science, Mathematics, and Informatics, IPB University, which has provided research funding for the Masters of Statistics and Data Science program.

References

- [1] BPS, "Jumlah Penduduk Pertengahan Tahun." Accessed: May 19, 2024. [Online]. Available: <https://www.bps.go.id/id/statistics-table/2/MTk3NSMy/jumlah-penduduk-pertengahan-tahun.html>
- [2] BPS, *Ringkasan Eksekutif Luas Panen dan Produksi Padi di Indonesia 2023 (Angka Tetap)*. Badan Pusat Statistik, 2023.
- [3] BBPOPT, *Laporan Kinerja BBPOPT Tahun 2023*. Balai Besar Peramalan Organisme Pengganggu Tanaman Indonesia, 2023.
- [4] P. N. Andono, Muljono, and T. Sutojo, *Pengolahan Citra Digital*. Yogyakarta: ANDI, 2017.
- [5] Y. N. Fuadah, S. Sa'Idah, I. Wijayanto, R. Patmasari, and R. Magdalena, "Non Invasive Anemia Detection in Pregnant Women Based on Digital Image Processing and K-Nearest Neighbor," in *Proc. 37th International Conference on Biomedical Engineering (IBIOMED 2020)*, Institute of Electrical and Electronics Engineers Inc., Oct. 2020, pp. 60–64. doi: <https://doi.org/10.1109/IBIOMED50285.2020.9487605>.
- [6] V. K. Shrivastava, M. K. Pradhan, and M. P. Thakur, "Application of Pre-Trained Deep Convolutional Neural Networks for Rice Plant Disease Classification," in *2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)*, Mar. 2021, pp. 1023–1030. doi: <https://doi.org/10.1109/ICAIS50930.2021.9395813>.
- [7] P. W. Rahayu et al., *Buku Ajar Data Mining*. Jambi: PT. Sonpedia Publishing Indonesia, 2024.
- [8] P. Hasan Putra, M. Syahputra Novelan, and M. Rizki, "Analysis K-Nearest Neighbor Method in Classification of Vegetable Quality Based on Color," *Journal of Applied Engineering and Technological Science*, vol. 3, no. 2, pp. 126–132, 2022, doi: <https://doi.org/10.37385/jaets.v3i2.763>.
- [9] A. X. Wang, S. S. Chukova, and B. P. Nguyen, "Ensemble K-Nearest Neighbors Based on Centroid Displacement," *InfSci (N Y)*, vol. 629, pp. 313–323, Jun. 2023, doi: <https://doi.org/10.1016/j.ins.2023.02.004>.
- [10] A. Hermawan, I. W. Mangku, N. K. K. Ardana, and H. Sumarno, "Analisis Support Vector Regression dengan Algoritma Grid Search untuk Memprediksi Harga Saham," *MILANG Journal of Mathematics and Its Applications*, vol. 18, no. 1, pp. 41–60, 2022, doi: <https://doi.org/10.29244/milang.18.1.41-60>.
- [11] S. Prusty, S. Patnaik, and S. K. Dash, "SKCV: Stratified K-Fold Cross-Validation on ML Classifiers for Predicting Cervical Cancer," *Frontiers in Nanotechnology*, vol. 4, Aug. 2022, doi: <https://doi.org/10.3389/fnano.2022.972421>.
- [12] F. Gemci, T. Ibrikci, and U. Cevik, "Comparative of Success of KNN with New Propose K-Split Method and Stratified Cross Validation on Remote Homologue Protein Detection," *Eskişehir Technical University Journal of Science and Technology A - Applied Sciences and Engineering*, vol. 23, no. 1, pp. 87–108, Mar. 2022, doi: <https://doi.org/10.18038/estubtda.970169>.
- [13] F. Mohameth, C. Bingcai, and K. A. Sada, "Plant Disease Detection with Deep Learning and Feature Extraction Using Plant Village," *Journal of Computer and Communications*, vol. 08, no. 06, pp. 10–22, 2020, doi: <https://doi.org/10.4236/jcc.2020.86002>.
- [14] S. Mathulapransan, S. Patarapuwadol, K. Lanthong, D. Jetpittanapong, and S. Sateanpattanukul, "Rice Disease Recognition Using Effective Deep Neural Networks," *Journal of Web Engineering*, vol. 20, no. 4, pp. 853–878, May 2021, doi: <https://doi.org/10.13052/jwe1540-9589.20313>.
- [15] N. S. B. Mat Said, H. Madzin, S. K. Ali, and N. S. Beng, "Comparison of Color-based Feature Extraction Methods in Banana Leaf Diseases Classification Using SVM and K-NN," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 24, no. 3, pp. 1523–1533, Dec. 2021, doi: <https://doi.org/10.11591/ijeecs.v24.i3.pp1523-1533>.
- [16] H. Fitriyah and R. C. Wiandika, *Dasar-Dasar Pengolahan Citra Digital*. Malang: UB Press, 2021.
- [17] Zircon, "Pengembangan Metode Sortasi Buah Kopi Robusta Menggunakan Pengolahan Citra Digital," IPB University, Bogor, 2023.
- [18] J. T. Abel et al., "Display Characteristics and Their Impact on Digital Pathology: A Current Review of Pathologists' Future 'Microscope,'" *J Pathol Inform*, vol. 11, 2020, doi: https://doi.org/10.4103/jpi.jpi_38_20.
- [19] Armita, F. Marpaung, F. Aulia, N. Suryani, and R. C. Nabila, *Computer Vision dan Pengolahan Citra Digital*. Pustaka Aksara, 2022. [Online]. Available: www.pustakaaksara.co.id
- [20] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *International Conference on Learning Representations*, 2014, pp. 1–14. [Online]. Available: <http://arxiv.org/abs/1409.1556>
- [21] D. Singh et al., "Classification and Analysis of Pistachio Species with Pre-Trained Deep Learning Models," *Electronics (Basel)*, vol. 11, 2022, doi: <https://doi.org/10.3390/electronics11070981>.
- [22] Z. Omiotek and A. Kotyra, "Flame Image Processing and Classification Using a Pre-trained VGG16 Model in Combustion Diagnosis," *Sensors*, vol. 21, pp. 1–15, 2021, doi: <https://doi.org/10.3390/s21020500>.
- [23] Rohit Thakur, "Step by step VGG16 implementation in Keras for beginners," *medium.com*. Accessed: Mar. 03, 2025. [Online]. Available: <https://medium.com/towards-data-science/step-by-step-vgg16-implementation-in-keras-for-beginners-a833c686ae6c>
- [24] E. Fix and J. L. Hodges, *Discriminatory Analysis Nonparametric Discrimination: Consistency Properties*. Berkeley: University of California, 1951.
- [25] M. Çakir, M. Yilmaz, M. A. Oral, H. Ö. Kazancı, and O. Oral, "Accuracy Assessment of RFerns, NB, SVM, and KNN Machine Learning Classifiers in Aquaculture," *J King Saud Univ Sci*, vol. 35, no. 6, Aug. 2023, doi: <https://doi.org/10.1016/j.jksus.2023.102754>.
- [26] R. A. Rahman, K. Sadik, and A. Fitrianto, "Simulation for Time Series Classification Using Feature Covariance Matrices with K-Nearest Neighbor," in *AIP Conference Proceedings*, American Institute of Physics Inc., Dec. 2022. doi: <https://doi.org/10.1063/5.0108204>.
- [27] T. A. Rizqi, "Penerapan Stratified Sampling pada Klasifikasi Gambar Penyakit Daun Padi dengan Metode K-Nearest Neighbor," Jakarta: Seminar Nasional Official Statistics, 2024, pp. 481–488. doi: <https://doi.org/10.34123/semnasoffstat.v2024i1.2023>.
- [28] F. Tarakci and A. Ozkan, "Comparison of classification performance of kNN and WKNN algorithms," *Selcuk University Journal of Engineering Sciences*, vol. 20, no. 02, pp.

- 32–37, 2021, [Online]. Available: <https://sujes.selcuk.edu.tr/sujes/article/view/536>
- [29] B. P. Nguyen, W. L. Tay, and C. K. Chui, "Robust Biometric Recognition From Palm Depth Images for Gloved Hands," *IEEE Trans Hum Mach Syst*, vol. 45, no. 6, pp. 799–804, Dec. 2015, doi: <https://doi.org/10.1109/THMS.2015.2453203>.
- [30] H. Fransiska, A. M. Soleh, K. A. Notodiputro, and Erfiani, "Evaluation of Machine Learning Models Based on Household Food Insecurity Data in Indonesia," in *BIO Web of Conferences*, EDP Sciences, Apr. 2025. doi: <https://doi.org/10.1051/bioconf/202517102011>.
- [31] K. Ali, Z. A. Shaikh, A. A. Khan, and A. A. Laghari, "Multiclass Skin Cancer Classification Using EfficientNets – A First Step Towards Preventing Skin Cancer," *Neuroscience Informatics*, vol. 2, no. 4, pp. 1–10, Dec. 2022, doi: <https://doi.org/10.1016/j.neuri.2021.100034>.
- [32] C. K. K. Reddy et al., "A Fine-Tuned Vision Transformer Based Enhanced Multi-Class Brain Tumor Classification Using MRI Scan Imagery," *Front Oncol*, vol. 14, 2024, doi: <https://doi.org/10.3389/fonc.2024.1400341>.
- [33] I. Markoulidakis, I. Rallis, I. Georgoulas, G. Kopsiaftis, A. Doulamis, and N. Doulamis, "Multiclass Confusion Matrix Reduction Method and Its Application on Net Promoter Score Classification Problem," *Technologies (Basel)*, vol. 9, no. 4, Dec. 2021, doi: <https://doi.org/10.3390/technologies9040081>.
- [34] S. H. Ou, "Rice Diseases," 2nd ed., UK: Commonwealth Mycological Institute, 1985.
- [35] T. W. . Mew and P. . Gonzales, *A Handbook of Rice Seedborne Fungi*. IRRI ; Science Publishers, 2002.
- [36] G. N. Agrios, *Plant Pathology*, 5th ed. Burlington: Elsevier Academic Press, 2005.