Rice Leaf Disease Image Classifications Using KNN Based On GLCM Feature Extraction

R A Saputra¹ *, Suharyanto², S Wasiyanti¹, D F Saefudin¹, A Supriyatna¹, A Wibowo²

¹Sistem Informasi Akuntansi, Universitas Bina Sarana Informatika, Indonesia
²Sistem Informasi, Universitas Bina Sarana Informatika, Indonesia
E-mail: rizal.rga@bsi.ac.id

Abstract. The disease that often occurs in rice leaves which causes crop failure. At present the attack of pests or leaf diseases in countries where the majority of rice as the main food of the community, especially in Indonesia is increasing. Early detection mechanism to reduce the risk of crop failure is very necessary. Data mining classification methods for the past few years have been very popular to be used in detecting the classification of rice leaf disease. Our paper uses 120 images of rice leaf disease from the UCI repository. The purpose of this study was to determine how to classify images of rice leaf disease consisting of three diseases namely Bacterial leaf blight, Brown spot, and Leaf smut. The study proposes the GLCM method as feature extraction for text analysis, with five feature values consisting of contrast, energy, entropy, homogeneity, and correlation. KNN (K-Nearest Neighbor) algorithm is used for the classification of rice leaf disease, by finding the maximum k value from the experiment k value 1 to 20. The results of our experiments show that the value of \( k = 11 \) has the highest accuracy value compared to other k values of 65.83% and kappa 0.485.

1. Introduction

The agricultural sector is important in a country’s growth, mostly for developing countries in the Southeast Asian region[1]. Indonesia is the largest country that consumes the highest rice in the world[2], rice is the main food for Indonesia peoples, rice production must increase because the population from year to year is also increasing[3]. Therefore rice must be the main concern of the central government in its production, to reduce rice exports[4]. Many factors will cause crop failure, one of which is disease[5], for this reason it is important to handle rice disease to reduce the risk of crop failure.

In the evaluation report on the main crop pest attack on rice plants in Indonesia in the planting season of 2018, it was stated that the most pests, namely rice stem borer, are predicted to attack 45,288.2 ha, brown stems, 22,747.5 ha, rats 46,944.5 ha, Tungro 3,212.8 ha, Leaf Blast 16,821.0 ha and Rotten leaf bacteria 22,747.5 ha. The maximum total estimated main pest attack in 2018 is 157,761.1 ha. The total incidence of major Padi attacks in the field in MT 2018 reached 145,636.5 ha or 92.3% of the forecast figure[4]. To prevent rice disease, it is not important in handling rice disease, so that the government and farmers can overcome the problem of rice crop failure.

Research on rice leaf disease has been done a lot before, such as research on the identification of symptoms of rice disease using image morphology to produce 12 leaf image data that can be
identified form of disease symptoms[6], previous studies using Bayes and SVM algorithms for detection of rice leaf disease resulted in an accuracy of 79.5% and 68.1%[7]. The classification of rice leaf disease research based on the results of the 4 angle interval GLCM feature extraction and backpropagation algorithm with 80% accuracy results but in this study, a testing process has not been carried out to improve accuracy and have not compared the results using k-fold[8]. The expert system research on rice plant identification using CBR, by entering symptoms into the system with the results of testing with a threshold similarity of 70% using the nearest neighbor method shows the system has a performance with 100% sensitivity and an accuracy rate of 82.69%[5]. Research on the identification of rice leaf disease with the Self-organizing map (SOM) technique, on the neural network algorithm and the zooming algorithm for image extraction, and resulted that the algorithm is good for the classification of rice leaf disease[9]. Research by extracting rice leaf mites based on images with BP Neural Network results that the algorithm is suitable for the classification of rice leaves[10].

The Gray level co-occurrence matrices (GLCM) method is one of the texture analysis methodologies that processes images used to describe spatial relationships and goes a step further to show the degree of application to gray level photomicrographs. Since then, GLCM has been widely used in applications[11]. The GLCM method proposed by Haralick in 1973, was used for classification of images using second-order statistical measurements[12], and it is proven that the GLCM method is a method that becomes an effective texture descriptor, and has better accuracy and computational time than other texture extraction methods[13].

Many data mining algorithms can be applied in the classification case[14], like K-Nearest Neighbor (KNN) algorithm, the KNN algorithm is included in the ten best classification algorithms[15] and it belongs to the instance-based group learning, but the KNN algorithm has a weakness that is in determining the variable at k value[16], the value of the variable must be solicited to produce maximum accuracy. This research will discuss a development in which 120 pictures of rice leaf disease data are extracted using the GLCM method with four angular intervals. After that, it is classified using the KNN algorithm, by finding the maximum k value in as many as twenty tests.

2. Methods
Classification by finding the maximum k value. At this stage describes the algorithm method used, then the proposed technique.

2.1. Algoritma KNN(K-Nearest Neighbor )
K-Nearest Neighbor (K-NN) algorithm is a classification method that is included in the supervised algorithm[17]. K-NN belongs to a group of simple algorithms, easily applied to machine learning algorithms that can be used to solve classification cases and regression cases. The application of the KNN algorithm in the case of classification is to find the value of group k on the object in the training data that is closest (similar) to the object in the testing data[15].

In general, the KNN algorithm is to find the distance between two objects x and y, and the Euclidean distance statistical formula is used in the equation[18][19][20]:

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$  \hspace{1cm} (1)

Note:
- d(x,y) : the distance between testing dan training data;
- x: data testing;
- y: data training;
- n: amount of training data.
2.2. Feature Extraction Algorithm Gray Level Co-occurrence Matrix (GLCM)
GLCM algorithm, one of the feature extraction methods used for texture analysis, the matrix in GLCM will calculate the probability value of the results of the relationship between two pixels with a certain intensity in the distance and orientation of a certain angle in the image [21].

The two-pixel coordinates have d distance and θ angle orientation. Distances are represented in pixels and angles are represented in degrees. Angular orientation will form into four angular directions, 0°, 45°, 90°, and 135°, and the distance between pixels is 1 pixel[22].

The steps taken in calculating the GLCM are as follows[21]:

1. Formation of the initial GLCM matrix from pairs of two pixels that line up in the direction of 0°, 45°, 90° or 135°;
2. Formation of a matrix by adding the initial matrix of GLCM;
3. Calculate the probability value of each element in the GLCM and divide it by the number of pixel pairs;
4. Calculate all feature extractions for each direction formed, namely:

   \[ \text{Contrast} = \sum_{i,j=0}^{N-1} P_{ij} (i - j)^2 \]  \hfill (2)

   \[ \text{Energy} = \sum_{i,j=0}^{n-1} (P_{ij})^2 \]  \hfill (3)

   \[ \text{Entropy} = \sum_{i,j=0}^{N-1} -\ln (P_{ij}) P_{ij} \]  \hfill (4)

   \[ \text{Homogeneity} = \sum_{i,j=0}^{N-1} \frac{P_{ij}}{1 + (j - j)^2} \]  \hfill (5)

   \[ \text{Correlation} = \sum_{i,j=0}^{N-1} \frac{P_{ij}}{1 + (j - j)^2} \]  \hfill (6)

2.3. Measurement of Accuracy of Algorithms
Measurement of the accuracy of an algorithm is the final stage to prove the level of performance measurement algorithm used in research. In this test, the confusion matrix is used as a measure of the performance of the KNN algorithm. Confusion Matrix is an evaluation of a data mining classification represented as a table [23]. Confusion matrix has information comparing the results of the classification carried out by the system with the results of the actual classification, in Table 1 below is a confusion equation matrix with two labels, True and False, with four different combinations of predictive values and actual values.

| Predicted Class | Predicted Class |
|-----------------|-----------------|
| True            | True (True Positif) | False (False Negatif) |
| False           | False (False Positif) | True (True Negatif) |
From table 1 the accuracy of an algorithm model can be calculated using the following equation:

\[
\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}
\]  

(7)

Note:
- TP: Positive data with positive classification results;
- FN: Represents positive data with negative classification results;
- FP: Is negative data with positive classification results;
- TN: Negative data with positive classification results.

2.4. Research Framework

Based on the description of the proposed methods in Figure 1, the stages in this study can be explained through the following steps:
- Starting from the extraction of rice disease features with the GLCM algorithm with 5 feature extractions with Matlab software;
- After feature extraction, the data set will be divided into 10 parts using ten-fold cross-validation, the data in the first part becomes the testing data and training data;
- The data is classified by the KNN algorithm with experimental testing looking for a maximum k value of 1 to 20.
- The final step is to measure the accuracy and kappa value of the KNN algorithm with Rapidminer Studio tools.

3. Result and Discussion
This stage explains the results of the experiment using rapid miner studio with experiments using k values ranging from 1 to 20. Figure 2 is a description of the dataset and algorithm performed in this study. From the calculation experiment of rice leaf classification using rapid miner studio, the accuracy and kappa KNN with the k value used can be seen in Table 2 below:
The results of testing the values of k 1 through 20 show that the value at k = 11 is the maximum k value with an accuracy level of 65.83% and a Kappa value of 0.485. Based on the Altman theory, that the kappa value of 0.485 includes the range 0.4-0.6, which is Strength Of Agreement Moderate [24].

4. Conclusions
The image classification of rice leaf disease with KNN and GLMC for feature extraction has medium accuracy and kappa values. This method can be developed with the optimization method of attribute selection or bagging techniques, so that it has high accuracy, and this method can be applied to applications so that it is easy for farmers to prevent crop failure.

Acknowledgments
The authors would like to thank the University of Bina Sarana Informatika for providing great support in conducting this research project. The authors would also like to thank our family for providing support throughout the research process.

References
[1] Presilla M and Rucianawati R, 2014, Pembangunan Sektor Pertanian di Asia Tenggara.
[2] Ishaq, Maulana., Rumiati, A T., Permatasari E O, 2017 Analisis Faktor-Faktor yang Mempengaruhi Produksi Kedelai di provinsi Jawa Yimur dengan metode regresi semiparametrik spline Sains dan Seni ITS 5, 2 p. 1–6.
[3] Lestari S and Bambang A N, 2017 Penerapan Minapadi dalam Rangka Mendukung Ketahanan Pangan dan Meningkatkan Kesejahteraan Masyarakat in Proceeding Biology Education Conference 14, 1 p. 70–74.
[4] Eka Putri A, 2017 Evaluasi Prakiraan Serangan OPT Utama Padi, Jagung, dan Kedelai di Indonesia MT2018, 53, 9, p. 1–62.
[5] Minarni and Warman I, 2017 Sistem Pakar Identifikasi Penyakit Tanaman Padi Menggunakan Case-Based Reasoning in Seminar Nasional Aplikasi Teknologi Informasi 5 Agustus 2017 ISSN: 1907 – 5022 p. 28–32.
[6] Zahrah S Saptono R and Suryani E, 2016 Identifikasi Gejala Penyakit Padi Menggunakan Operasi Morfologi Citra Suik October 2016 p. 100–106.
[7] Phadikar S, 2012 Classification of Rice Leaf Diseases Based on Morphological Changes Int. J. Inf. Electron. Eng. 2, 3 p. 460–463.
[8] Jani K and Noor A H, 2018 Klasifikasi Penyakit Daun Padi Berdasarkan Hasil Ekstraksi Fitur GLCM Interval 4 Sudut J. Teknol. Inf. dan Komun. STMIK ProVisi Semarang 03, 01 p. 1–6.
[9] Phadikar S and Sil J, 2008 Rice disease identification using pattern recognition techniques in Proceedings of 11th International Conference on Computer and Information Technology, ICCIT 2008 Iccit p. 420–423.
[10] Liu L and Zhou G, 2009 Extraction of the rice leaf disease image based on BP neural network in Proceedings - 2009 International Conference on Computational Intelligence and Software Engineering, CiSE 2009 p. 9–11.
[11] Jardine M A Miller J A and Becker M, 2018 Coupled X-ray computed tomography and grey level co-occurrence matrices as a method for quantification of mineralogy and texture in 3D Comput. Geosci. 111 p. 105–117.

[12] Pantic I Dimitrijevic D Nesic D and Petrovic D, 2016 Gray level co-occurrence matrix algorithm as pattern recognition biosensor for oxidopamine-induced changes in lymphocyte chromatin architecture J. Theor. Biol. 406 p. 124–128.

[13] Roberti de Siqueira F Robson Schwartz W and Pedrini H, 2013 Multi-scale gray level co-occurrence matrices for texture description Neurocomputing 120 p. 336–345.

[14] Han J and Kamber M, 2006 Data Mining Concepts and Techniques Second Edition San Francisco: Morgan Kaufmann.

[15] Wu X and Kumar V, 2009 The Top Ten Algorithms in Data Mining London.

[16] Karomi M A Al, 2015 Optimasi Parameter K Pada Algoritma Knn Untuk Klasifikasi Heregistrasi Mahasiswa IC-Tech 10, 1 p. 28–33.

[17] Han J and Kamber M, 2006 Mining Stream, Time-Series and Sequence Data 54.

[18] Leidiyana H, 2013 Penerapan Algoritma K-Nearest Neighbor Untuk Penentuan Resiko Kredit Kepemilikan Kendaraan Bermotor J. Penelit. Ilmu Komputer, Syst. Embed. Log. 1, 1 p. 65–76.

[19] Mustakim G O, 2016 Algoritma K-Nearest Neighbor Classification J. Sains, Teknol. dan Ind. 13, 2 p. 195–202.

[20] Yunita D, 2017 Perbandingan Algoritma K-Nearest Neighbor dan Decision Tree untuk Penentuan Risiko Kredit Kepemilikan Mobil J. Inform. dan Ilmu Komput. Progr. Inform. Teknol. 2, 11 p. 5769–5776.

[21] Widodo R Widodo A W and Supriyanto A, 2018 Pemanfaatan Ciri Gray Level Co-Occurrence Matrix (GLCM) Citra Buah Jeruk Keprok (Citrus reticulata Blanco) untuk Klasifikasi Mutu J. Pengemb. Teknol. Inf. dan Ilmu Komput. 2, 11 p. 5769–5776.

[22] Surya R A Fadil A Yudhan A Informatika M T Informatika P T dan Dahlan U A, 2017 Ekstraksi Ciri Metode Gray Level Co-Occurrence Matrix (GLCM) dan Filter Gabor Untuk Klasifikasi Citra Batik Pekalongan J. Inform. Pengemb. IT (JPIT, Vol. 02, No. 02, Juli 2017 02, 02 p. 23–26.

[23] Gorunescu F, 2011 Data Mining “Concepts, Models and Techniques” 12 Berlin, Heidelberg: Springer Berlin Heidelberg.

[24] Altman D G, 1991, Practical Statistics For Medical Research, (London: CRS Press), p. 277-300.