Classification of Tobacco Leaf Quality Using Feature Extraction of Gray Level Co-occurrence Matrix (GLCM) and K-Nearest Neighbor (K-NN)

Aeri Rachmad[1(✉)], Rinci Kembang Hapsari[2], Wahyudi Setiawan[1], Tutuk Indriyani[2], Eka Mala Sari Rochman[1], and Budi Dwi Satoto[1]

[1] Faculty of Engineering, University of Trunojoyo Madura, Bangkalan-Madura, Indonesia
{aery_r,wsetiawan,budids}@trunojoyo.ac.id
[2] Faculty of Information Technology, Institute of Technology Adhi Tama, Surabaya, Indonesia
{rincikembang,tutuk}@itats.ac.id

Abstract. Tobacco is one of the largest agricultural products and is widely traded in the world market, including in Indonesia. In Indonesia, tobacco leaves are used as raw material for cigarettes which are mostly produced by cigarette companies. The quality of tobacco leaves greatly affects the quality of cigarettes, this is because the condition of tobacco leaves is influenced by several factors including pests, diseases, and climate. This study uses the Gray Level Co-Occurrence Matrix (GLCM) method for texture feature extraction, while for classification uses the K-Nearest Neighbor (KNN) method to classify the quality of tobacco leaves. The data used in this study is the image of tobacco leaves taken directly in TonDowulan Village, Plandaan District, Jombang Regency at the age of the leaves of approximately 2 months. Tobacco leaf images used were 300 images consisting of 3 classes, namely Normal, Perforated, and Withered based on the level of leaf damage. The GLCM features used are Contrast, Correlation, Energy, Homogeneity, and Entropy which will then be classified using the KNN method where before performing feature extraction the data must be processed first at the preprocessing stage. The result of the training using GLCM and K-NN feature extraction produces the highest accuracy value when the neighbor value 1, pixel distance 3, and k-fold 2 are 83.33%.

Keywords: Tobacco · GLCM · classification · K-NN

1 Introduction

Tobacco is one of the largest agricultural products that is widely traded throughout the world, including in Indonesia [1]. The part of tobacco that is widely used in tobacco leaves, these leaves are used as raw material for cigarettes [2]. Indonesia itself has several tobacco-producing regions including Palembang, Central Java, East Java, North Sumatra, and West Sumatra. Tobacco has an important role in the national economy,

I. H. Agustin (Ed.): ICONNSMAL 2022, AISR 177, pp. 30–38, 2023.
https://doi.org/10.2991/978-94-6463-174-6_4

this is because tobacco can increase the country's source of income through foreign exchange, excise, a source of income for farmers and can create jobs. The existence of tobacco needs to be maintained and further enhanced so that the state's source of income continues to increase and there are more jobs, this is from a commercial perspective [3].

Several factors affect the level of quality of tobacco plants, namely pests and diseases and the level of damage to tobacco leaves caused by different climates in each region. These pests and diseases are problems caused by pest attacks that attack tobacco leaves and are the most important problem for tobacco farmers, currently, pests and diseases that attack tobacco plants vary widely [1]. In addition, the level of damage to tobacco is also a very influential problem for farmers, this damage often occurs due to changes in climate, land, planting techniques, and care. Tobacco diseases usually have different symptoms, including hollow leaves caused by leaf caterpillars. As for the level of damage to the tobacco leaves themselves, it is indicated by the leaves looking wilted and then drying. This can be caused by the climatic conditions in each region being different [3].

The classification of tobacco leaves is carried out by a tobacco expert, namely a grader. The grader oversees measuring and analyzing the quality of the tobacco then the results will be grouped into certain grades. This classification can be done using two factors, namely internal factors, and external factors. Internal factors are more directed to the sense of smell (human sensory), testing with smoking tests, and chemical analysis. While external factors are more directed to the grader's sense of sight which is determined based on color, the level of maturity of the tobacco leaves, the texture of the surface, the size, and the shape of the tobacco leaves [2]. However, the sense of smell and the sense of sight cannot be used to fully classify tobacco. This is because human nature (such as emotions, fatigue, etc.) will affect the quality of the classification, which can result in the efficiency of the classification and the stability of the grade of tobacco not being achieved properly. Therefore, Digital Image Processing techniques are needed in the context of classifying the level of damage to tobacco leaves [1].

Previous research used the Gray Level Co-occurrence Matrix method for feature extraction and KNN for the identification of orchid species, this study used two types of orchids, namely the moon orchid and the larat orchid. This study produces an accuracy value of 80% with an average of 77%. The K value affects the success rate of identification, the greater the K value, the smaller the accuracy [4].

In this research we will use steps such as image acquisition, preprocessing, feature extraction, and object classification. The method we used in this study is the Gray Level Co-Occurrence Matrix (GLCM) which is used for texture feature extraction. After performing texture feature extraction then the classification process. This process uses the K-Nearest Neighbors (KNN) classification algorithm with measurement of accuracy using the Confusion Matrix.

2 Research Method

This section describes the stages of the research which consist of dataset input, split data training and testing, cross-validation, GLCM feature extraction, K-NN classification, and testing using the best model results from the training process. The complete process of the research stages is shown in Fig. 1.

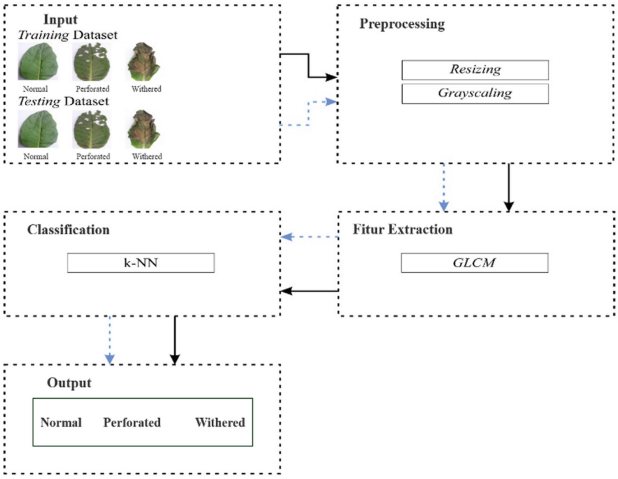

Fig. 1. The proposed system designs

Fig. 2. Three levels of damage to tobacco leaves (a) normal tobacco leaves (b) perforated tobacco leaves and (c) withered tobacco leaves

2.1 Dataset

The image data used is the image of a tobacco leaf measuring 3000x4000 pixels which was taken directly by the researcher. The dataset used is 300 images which will be divided into 3 classes based on the level of leaf damage, namely normal tobacco leaves, perforated tobacco leaves, and withered tobacco leaves as shown in Fig. 2. The data consisted of 100 data on normal tobacco leaves, 100 data on perforated tobacco, and 100 data on wilted leaves, where the data was taken in Tondowulan Village, Plandaan District, Jombang Regency. Tobacco leaf images were taken at approximately 2 months of age. The data collection time is in the morning at around 08.00 am. The shooting technique was taken using a personal cellphone by equalizing the shooting distance and lighting.

2.2 Preprocessing

The tobacco leaf data used is the tobacco leaf image which has been segmented and the background image is changed to 0. The tobacco leaf data is then changed to grayscale. Then the image will enter the resize stage to generalize the image size from 3000 x 4000 pixels to 256 x 256 pixels. Resizing is done because the images of tobacco leaves have

different sizes, in addition to shortening the computational time in image processing [5, 6].

2.3 Gray Level Co-occurrence Matrix (GLCM)

The Gray Level Co-occurrence Matrix (GLCM) is a co-occurrence matrix in which each element is the number of occurrences of pixels that have a certain gray value and each pair of pixels is at a certain distance and in a certain direction [7]. The distance (d) used is usually d = 1 and is expressed in pixels, while the angular direction is expressed in degrees (0º, 45º, 90º, and 135º) [8]. The four directions are shown in Fig. 3.

Some of the GLCM features used in this study are:

1. Contrast, which is a feature that represents the difference in the level of color/grayscale that appears in an image from a pair of adjacent pixels. Can be searched with the following formula:

$$CON = \sum_{i,j} (i-j)^2 p_{i,j} \tag{1}$$

2. Correlation, which is a feature that represents the linear relationship of the degree of gray image. The Correlation is between -1 to 1. It can be found using the following formula:

$$Corr = \sum_{i,j} \frac{(i-\mu i)(j-\mu j)p(i,j)}{\sigma i \sigma j} \tag{2}$$

3. Entropy is used to measure the randomness of the intensity distribution. Can be searched using the following formula:

$$En = -\sum_{i}^{m}\sum_{j}^{n} p(i,j)\log\{p(i,j)\} \tag{3}$$

4. Energy, namely the degree of similarity between a pixel in the image. The higher the value of energy in the image. The higher the similarity between pixels. The energy value can be found using the following formula:

$$Eng = \sum_{i}\sum_{j} p(i,j)^2 \tag{4}$$

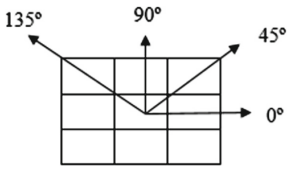

Fig. 3. Directional GLCM co-occurrence drawing

5. Homogeneity is used to measure the homogeneity of image intensity variations. The homogeneity value (Hom) can be calculated by the following formula:

$$Hom = \sum_{i} \sum_{j} \frac{p(i,j)}{1 + |i - j|}. \tag{5}$$

2.4 K-Nearest Neighbor

K-Nearest Neighbor (KNN) is an algorithm used to classify an object based on data that has the closest distance to the object. The way the KNN algorithm works is to find the shortest distance between the testing data and training data with a value of k equal to the number of nearest neighbors [9]. The distance between training data and testing data can be calculated by various methods, including using the Euclidean equation [10].

$$D(x, y) = \sqrt{\sum_{i=1}^{N} (xi - xy)^2} \tag{6}$$

2.5 Confusion Matrix

From testing the training data and testing data, it produces a confusion matrix as shown in Table 1. This confusion matrix is used to measure performance in machine learning classification problems where the output can be in the form of two or more classes [11, 12]. From the confusion matrix table, accuracy, recall, precision, and F1 score will be calculated.

a. Accuracy is a description of how accurate the model is used to classify the entire data; how close the predicted value is to the actual value. TP and TN are divided by the total number of testing data.

$$Accuracy = \frac{TP}{TP + FP + TN + FN} \tag{7}$$

Precision is the ratio of correct positive predictions (TP) compared to the overall positive predicted results (TP + FP).

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

Table 1. Confusion matrix

	True	**Negative**
True	True Positive (TP)	False Negative (FN)
Negative	False Positive (FP)	True Negative (TN)

b. A recall is a true positive prediction compared to all true positive data. Recall itself describes the success of the model in retrieving information.

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

c. F1 Score is a comparison of the average precision and recall which is weighted. F1 Score can be used as a reference if the number of False Negative and False Positive data is not close.

$$F1Score = \frac{2 * precision * recall}{precission + recall} \tag{10}$$

3 Result and Discussion

In general, the needs and types of software that researchers will use to help researchers complete this research are Windows 10 Home Single 64 Bit, the programming language uses Python, the processor is Intel Celeron, and the text editor uses Sublime Text.

The trials in this study used 300 tobacco leaf images measuring 256x256 pixels which were divided into 3 classes with a total of 100 data per class. Data sharing is 80% training data and 20% testing data. This study uses the K-Fold Cross Validation test with K = 5. The data set will be divided into 5 folds so that each partition will be 60 data. The parameters used in this study are the value of neighbors, pixel distance, and K-fold. The best model will be saved from the best results. In addition, the classification test process makes use of the best model in the testing data.

The experiment we did was to carry out 20 trials with neighboring values of 1, 3, 5, 7, and 9 and pixel distances of 1, 2, 3, and 4.

The results of image classification of tobacco leaves using KNN with GLCM feature extraction without combining pixel distance produce the best precision, recall, f1 score, and accuracy values when the neighbor value is 1, pixel distance 3, and k-fold 2, namely precision value 83.15%, recall 83.33%, f1 score 83.24%, and accuracy 83.33% as shown in Table 2 and Fig. 4.

The best classification results are by merging pixel distance when the neighbor value is 1. The 2nd fold with a precision value of 86.99%, a recall of 86.67, an f1 score of 86.83%, and an accuracy of 86.67% as shown in Table 3 and Fig. 5.

Table 2. GLCM Test Results Without Mixing Pixel Distance

No	Pixel Distance	K-Fold	Precision	Recall	F1 Score	Accuracy
1	3	1	67,21%	66,67%	66,94%	66,67%
2	**3**	**2**	**83,15%**	**83,33%**	**83,24%**	**83,33%**
3	3	3	65,42%	61,67%	63,49%	61,67%
4	3	4	65,93%	63,33%	64,60%	63,33%
5	3	5	69,75%	68,33%	69,04%	68,33%

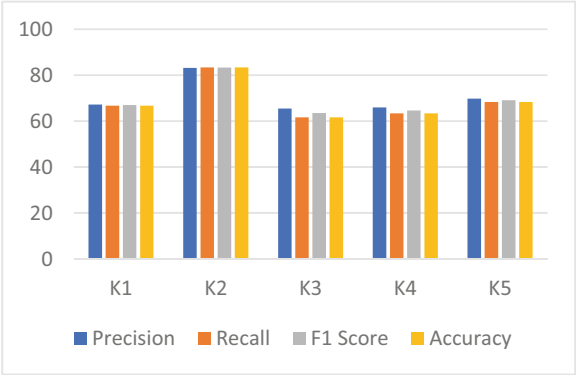

Fig. 4. Classification results using K-NN with GLCM without combining pixel distance

Table 3. GLCM Test Results with Mixing Pixel Distance

No	K-Fold	Precision	Recall	F1 Score	Accuracy
1	1	75,00%	75,00%	75,00%	75,00%
2	**2**	**86,99%**	**86,67%**	**86,83%**	**86,67%**
3	3	75,21%	75,00%	75,11%	75,00%
4	4	77,51%	71,67%	74,47%	71,67%
5	5	68,83%	66,73%	67,73%	66,67%

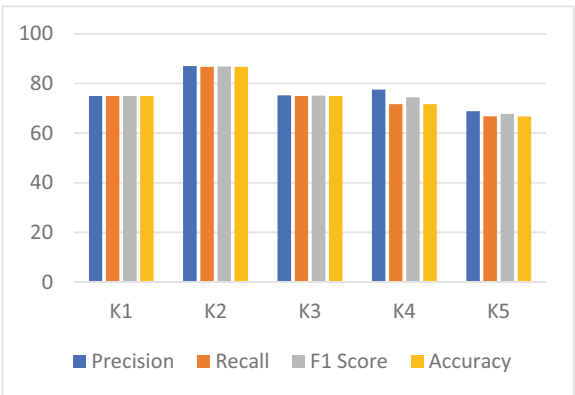

Fig. 5. Classification results using K-NN with GLCM with combining pixel distance

4 Conclusion

The experimental results above show that the results of image classification of tobacco leaves using GLCM with pixel distance incorporation have better accuracy results of 86.67%. While the accuracy for classification using GLCM feature extraction without merging pixel distance is 83.33%.

Acknowledgment. Many thanks to the Institute of Research and Community Service and the Faculty of Engineering at the University of Trunojoyo Madura for allowing the researchers to complete this research. We also thank the research team at the Multimedia and Networking Laboratory for supporting the success of this research.

References

1. F. Damayanti, A. Muntasa, S. Herawati, M. Yusuf, & A. Rachmad. Identification of Madura Tobacco Leaf Disease Using Gray-Level Co-Occurrence Matrix, Color Moments and Naïve Bayes. In Journal of Physics: Conference Series, Vol. 1477, No. 5, p. 052054, 2020.
2. H. Avila-George, T. Valdez-Morones, B. Acevedo-Ju, & W. Castro, Using artificial neural networks for detecting damage on tobacco leaves caused by blue mold. International Journal of Advanced Computer Science and Applications, Vol. 9, pp. 579–583, 2018.
3. D. Djajadi, Tobacco diversity in Indonesia. Berkala Penelitian Hayati, Vol. 20, pp. 27–32, 2015.
4. D. P. Pamungkas, Ekstraksi Citra menggunakan Metode GLCM dan KNN untuk Identifikasi Jenis Anggrek (Orchidaceae). Jurnal INNOVATICS: Innovation in Research of Informatics, Vol. 1, pp. 51–56, 2019.
5. A. Rachmad, N. Chamidah, & R. Rulaningtyas, Mycobacterium Tuberculosis Identification Based on Colour Feature Extraction using Expert System. *Ann. Biol*, Vol. *36*, pp. 196–202, 2020.
6. F. Damayanti, & A. Rachmad, Recognizing Gender Through Facial Image using Support Vector Machine. Journal of Theoretical & Applied Information Technology, Vol. 88, pp. 607–612, 2016.
7. O. R. Indriani, E. J. Kusuma, C. A. Sari, & E. H. Rachmawanto, Tomatoes Classification using K-NN based on GLCM and HSV Color Space. In *2017 international conference on innovative and creative information technology (ICITech),* . IEEE, pp. 1–6, 2017.
8. Ch. R. Babu, D. S. Rao, V. Sravan Kiran, N, Assessment of Plant Disease Identification using GLCM and KNN Algorithms Rajasekhar, International Journal of Recent Technology and Engineering (IJRTE), Vol. 8, pp. 4900–4904, 2020.
9. F. Damayanti, S. Herawati, I. Imamah, & A. Rachmad, Indonesian License Plate Recognition Based on Area Feature Extraction. Telkomnika (Telecommunication Computing Electronics and Control), Vol. 17, pp. 620–627, 2019.
10. A. Rachmad, & M. Fuad, Geometry Algorithm on Skeleton Image Based Semaphore Gesture Recognition. *Journal of Theoretical & Applied Information Technology*, Vol. *81*, pp. 102–107, 2015.
11. A. Rachmad, N. Chamidah, & R. Rulaningtyas, Mycobacterium tuberculosis images classification based on combining of convolutional neural network and support vector machine. Commun. Math. Biol. Neurosci., Vol. 2020, pp. 1–13, 2020.

12. A. Rachmad, N. Chamidah, & R. Rulaningtyas, Classification Of Mycobacterium Tuberculosis Based on Color Feature Extraction Using Adaptive Boosting Method. In *AIP Conference Proceedings* (Vol. 2329, p. 050005, 2021.