Identification of Madura Tobacco Leaf Disease Using Gray-Level Co-Occurrence Matrix, Color Moments and Naïve Bayes

Fitri Damayanti*, **Arif Muntasa, Sri Herawati, Muhammad Yusuf, and Aeri Rachmad**

Departemen of Informatics, Faculty of Engineering, University of Trunojoyo,
Madura, Bangkalan, Indonesia

*fitrid@trunojoyo.ac.id

Abstract. Indonesia is one of the world's biggest tobacco crop producers. By tobacco farmer, this plant is often even dubbed "green gold". Madura Island is one of the best tobacco-producing areas in Indonesia. Tobacco is a significant trading crop in the eastern part of Madura Island, specifically in Pamekasan and Sumenep. The decline in tobacco yields is usually caused by pests and diseases that attack tobacco plants. Experts can easily detect conditions in plants (including tobacco) with their eyes, but this is very suitable and requires expensive operational costs when the size of the planting area is vast, and the distance of the planting area is far from the location of the expert. So that digital image processing techniques need to be applied to detect tobacco plant diseases earlier. By using data in the form of photographs of tobacco plant leaves, the condition will be identified. The method used in this research is GLCM (Gray Level Co-Occurrence Matrix) texture feature extraction, while CM (Color Moment) colour feature extraction and Naïve Bayes method are used for classification. The results of testing tobacco identification obtained the best accuracy of 82.2% for Pamekasan tobacco and 84.4% for Sumenep tobacco. The best results are obtained by using the colour feature extraction.

1. Introduction

Tobacco is one of the commodities of plantations that has been known since Indonesia was colonized by the Dutch. Indonesia has many islands that produce tobacco commodities, one of which is Madura Island. Tobacco from Madura Island has also been marketed in many regions.

One of the problems faced by tobacco farmers is the emergence of tobacco disease, which causes crop yields to decline or even experience crop failure, this is caused by the delay of farmers in identifying the types of diseases from tobacco that arise, so the handling method is also late. The delay will cause the condition of tobacco plants to worsen, so that the worst possibility is the occurrence of death from these plants.

Tobacco disease can be caused by several factors. The first factor is caused by aphids, in this case the lice can attack during breeding. The second factor is caused by caterpillars, where most of the tobacco leaves are eaten by caterpillars, so the leaves become hollow and the numbers are very large. The third factor is beetles that eat tobacco leaves and their bite marks cause irregular damage. The fourth factor is an animal that damages the base of the stem of a tobacco plant, so the plant will die if part of the base is damaged. The fifth factor is caused by a virus which causes the tobacco leaves to contract to look like crackers.

By considering the many types of diseases on tobacco leaves, it is necessary to conduct a study to produce a model and software to help tobacco farmers in identifying the types of diseases that attack these plants. Several studies to build models that can be used to classify tobacco leaf quality have been carried out by many researchers [1], [2], [3], [4]. But unfortunately some of the research conducted has many weaknesses in determining features before measuring similarities.

Thirumalesh B. V., Suresha M, and Shreekanth K. N., who have conducted research on the identification of rice plant leaf diseases stated that the Naive Bayesian classification method is able to provide better accuracy than the KNN and SVM methods [5].

A research which was conducted by Jayamala K. Patil and Raj Kumar (2011) entitled "Color Feature Extraction of Tomato Leaf Diseases" explains that color moments with 3 moments are able to obtain color features from the leaves of tomato plants. This study also explained to add a grouping algorithm so that the system is able to group the leaves of tomato plants in each disease so that it can be calculated the accuracy of the system built [6].

Research conducted by Gautam Kaushal and Rajni Bala (2017) written in his paper is entitled "GLCM and KNN based Algorithm for Plant Disease Detection". The paper describes the detection of plant diseases using extraction of texture, segmentation and classification features. The Gray Level Co-occurrence Matrix (GLCM) algorithm is applied which extracts texture features from the image. The k-mean algorithm is used to segment input images. The Support Vector Machine (SVM) classification is applied in an existing algorithm that will classify input images into two classes. To improve the performance of existing algorithms, the SVM classifier is replaced by the K-Nearest Neighbor (KNN) classification. This leads to an increase in accuracy of disease detection, let alone classifying data into classes. The performance of the proposed algorithm is tested in terms of accuracy and false positive rate with an increase of up to 10 percent compared to existing techniques [7].

In research conducted by A. Harshavardhan, Dr. Suresh Babu and Dr. T. Venugopal (2017) entitled "Analysis of Feature Extraction Methods for the Classification of Brain Tumor Detection". In this study using the Histogram, GLRLM and GLCM methods. In the research results it was written that the Histogram method obtained an accuracy of 83%, the GLRLM method obtained an accuracy of 85%, and the highest was the GLCM method with an accuracy of 87.9% [8].

Based on research that has been done by other researchers before, in this study an application was made to recognize the types of diseases in tobacco plant leaves. The method used in the feature extraction process is the Gray Level Co-occurrence Matrix (GLCM) and Color Moments (CM) method, while for the classification of tomato leaf disease was using the Naive Bayesian method.

2. Methods

In this study, the authors built a model in identifying diseases in tobacco leaves. The model was created using the extraction of texture and color features. The texture feature extraction method used the GLCM (Gray Level Co-Occurrence Matrix) method and the color feature extraction method usesd the CM (Color Moment) method. The classification method used was Naïve Bayes. The general system process in this study can be seen in Figure 1.

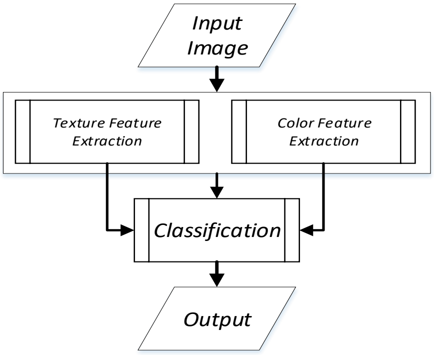

Figure 1. General System Process

2.1. *Gray Level Co-occurrence Matrix (GLCM)*

Gray Level Co-occurrence Matrix (GLCM) is one of the methods in digital image processing to extract texture features by looking for important texture features of the image. GLCM can also be used for statistical based texture analysis of objects. GLCM is formed by determining the location of adjacent pixels (*d*) and the angle of those adjacent pixels (Θ). To obtain GLCM values, angles use include lain 0^0, 45^0, 90^0, 135^0, 180^0, 225^0, 270^0, 315^0 and 360^0 [9].

GLCM uses angles $\Theta = 0^0$ for horizontal, angles $\Theta = 90^0$ for vertical, and angles $\Theta = 45^0$ and $\Theta = 135^0$ for the diagonal direction which can be seen in Figure 2. If the bottom-up and right-left directions are considered in the process of finding values GLCM, the angles used are 8 angles. However, if you get the GLCM value using 8 different angles, then not all angles must be calculated, but it is enough to count only 4 angles, namely: 0^0, 45^0, 90^0, and 135^0 because the angle 00 is symmetrical with 180^0, the angle is 45^0 symmetric with 225^0, and thereafter [9].

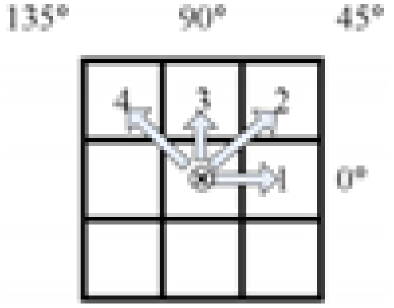

Figure 2. GLCM angles

GLCM has 9 features that can be used to get the texture characteristics of an image [9]. The 9 characteristics are:

1. *Angular Second Moment Feature* (ASM *Feature*).
   This feature serves to measure the homogeneity of pixels in an image [10]. ASM can be calculated using equation 1.

$$f_1 = \sum_{i=0}^{N_g-1} \sum_{j=0}^{N_g-1} P_{d,\theta}(i,j)^2 \qquad (1)$$

Where:

N_g = Number of rows / columns.
$P_{d,\theta}(i,j)$ = Pixel value in row *j*, column *i*.

2. *Contrast Feature*
   This feature is used to measure variations between the gray degrees of an area in the image [10]. This feature can be calculated using equation 2.

$$f_2 = \sum_{n=0}^{N_g-1} |i-j|^2 * \left\{ \sum_{i=0}^{N_g-1} \sum_{j=0}^{N_g-1} P_{d,\theta}(i,j) \right\} \qquad (2)$$

3. *Entropy Feature*
   This feature can measure the degree of irregularity in the shape of the texture in the image. The entropy feature value will be high if the image structure is relatively orderly, but if the entropy feature value is low if the image structure is irregular [10]. This feature can be calculated using equation 3.

$$f_3 = -\sum_{i=0}^{N_g-1} \sum_{j=0}^{N_g-1} P_{d,\theta}(i,j) * log\left(P_{d,\theta}(i,j)\right) \qquad (3)$$

4. *Variance Feature*
   This feature serves to distinguish each element in the GLCM matrix. Low intensity images have small variations [10]. To get the value of this feature, use equation 4

$$f_4 = \sum_{i=0}^{N_g-1} \sum_{j=0}^{N_g-1} (i-\mu)^2 * P_{d,\theta}(i,j) \qquad (4)$$

Where:
μ = Average on the matrix.

5. *Correlation Feature*
Correlation Feature is used to measure linear dependence of a gray level reference pixel matrix to its neighbors [9]. To get this feature, equation 5 can be used.

$$f_5 = \sum_{i=0}^{N_g-1}\sum_{j=0}^{N_g-1} P_{d,\theta}(i,j) * \frac{(i-\mu_x)(i-\mu_y)}{\sigma_x\sigma_y} \tag{5}$$

Where:
μ_x = Average lines in the matrix.
μ_y = The average column in a matrix.
σ_x = Standard deviation of row elements in the matrix.
σ_y = Standard deviation of column elements in the matrix

If the values in the GLCM are symmetrical, then $\mu_i = \mu_j$ and $\sigma_i = \sigma_j$. , While on the contrary, if the values of GLCM are not symmetrical, to get the values μ_i, μ_j, σ_i, and σ_j can use equations 6, 7, 8, and 9 [9].

$$\mu_x = \sum_{i=0}^{N_g-1}\sum_{j=0}^{N_g-1} i * P_{d,\theta}(i,j) \tag{6}$$

$$\mu_y = \sum_{i=0}^{N_g-1}\sum_{j=0}^{N_g-1} j * P_{d,\theta}(i,j) \tag{7}$$

$$\sigma_x = \sum_{i=0}^{N_g-1}\sum_{j=0}^{N_g-1} (i-\mu)^2 * P_{d,\theta}(i,j) \tag{8}$$

$$\sigma_y = \sum_{i=0}^{N_g-1}\sum_{j=0}^{N_g-1} (j-\mu)^2 * P_{d,\theta}(i,j) \tag{9}$$

6. *Inverse Different Moment Feature* (IDM *Feature*)
This feature shows the similarity of images with similar intensity values. An image that has a relatively high level of homogeneity will have a high IDM value to [10]. IDM features can be calculated using equation 10.

$$f_6 = \sum_{i=0}^{N_g-1}\sum_{j=0}^{N_g-1} \left(\frac{1}{1+(i-j)^2}\right) * P_{d,\theta}(i,j) \tag{10}$$

7. *Sum Average Feature*
This feature functions to obtain the texture characteristics of the image based on the average number that can be seen in equations 11 and 12 [9].

$$f_7 = \sum_{i=0}^{2(N_g-1)} i * P_{x+y}(i) \tag{11}$$

$$P_{x+y}(k) = \sum_{i=0}^{N_g-1}\sum_{j=0}^{N_g-1} P_{d,\theta}(i,j), \quad k = i+j = \{1,2,\dots 2(N_g-1)\} \tag{12}$$

8. *Sum Variance Feature*
This feature will calculate the amount of variance that results from calculations on the sum average feature as written in equation 13 [9].

$$f_8 = \sum_{i=0}^{2(N_g-1)} (i-f_7)^2 * P_{x+y}(i) \tag{13}$$

9. *Sum Entropy Feature*
This feature is the development of entropy features that can be calculated based on equation 14 [9].

$$f_9 = -\sum_{i=0}^{2(N_g-1)} P_{x+y}(i) * \log P_{x+y}(i) \tag{14}$$

2.2. Color Moment

Color moment is a measurement method used to distinguish images based on color features in the image. The basis of color moments is that the color distribution of an image can be interpreted as a distribution probability [6]. Color moments are the right choice for extracting color features of plant leaves because there are differences in color on normal leaves and affected leaves. Color moments are an effective and efficient feature extraction method for color based image analysis because they have low feature vector dimensions and low computational complexity when compared to other methods, such as color histograms, color correlograms and color structure descriptors [11].

The mean, variance and standard deviation of an image are known as color moments [12]. Generally, color moments use 3 features. The first color moment feature is also called the mean which is the average pixel value (Pij) in an image that can be calculated using equation 15 [13].

$$\mu_c = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N} P_{ij}^{c} \tag{15}$$

Where:
P_{ij}^c = pixel value (i, j) of the image in the color component c..
M = Image length.
N = Image width
c = Component color.
μ_c = Mean value of the color component c.

The second Color Moments feature is also called standard deviation which is the root of variance. Variance is the distribution area of distribution [6]. Standard deviation can be calculated using equation 16 [13].

$$\sigma_c = \left[\frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}\left(P_{ij}^{c} - \mu_c\right)^2\right]^{1/2} \tag{16}$$

Where:
σ_c = Image deviation standard on color components c.

The third Color Moments feature is also called skewness. Skewness is used to determine the size of the level of symmetry in the distribution. It is said to be symmetrical if it is balanced between left and right at the center point [8]. Skewness can be calculated using equation 17 [13].

$$\theta_c = \left[\frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}\left(P_{ij}^{c} - \mu_c\right)^3\right]^{1/3} \tag{17}$$

Where:
θ_c = Image skewness on color components c.

Color Moments with these 3 features indicate three results. Those 3 results can be obtained namely the results of the calculation of features 1, 2 and 3 so that the color features of the image can be known. From these results a further process can be carried out in the form of leaf classification to each disease [8].

This method uses a color image (RGB image) to be changed in the CIELAB image, unlike the

Gray Level Co-occurrence Matrix (GLCM) method that uses grayscale imagery. Therefore, this method only counts 3 features of the image. However, because the image used is a color image (RGB image) that has been changed in the CIELAB image. From each image 3 features for each layer could be obtained, so 9 features for each image input can be obtained too. However, the process to obtain 3 features from each layer has the same process at each layer.

2.2.1 Naïve Bayesian

Naive Bayesian is one of the classification methods using the similarity of the characteristics of an object. This method is classified as a fairly simple method, but is widely used in the fields of medicine, biometrics, text classification, and many more. Naive Bayesian uses the Gaussian distribution by considering 2 important parameters, namely the average μ and the variance σ [9]. In Naive Bayesian, Gaussian uses equation 18.

$$P(X_i = x_i \mid Y = y_i) = \frac{1}{\sqrt{2\pi}\sigma_{ij}} exp\left[-\frac{1}{2}\left(\frac{x_i - \mu_{ij}}{\sigma_{ij}}\right)^2\right] \tag{18}$$

Where:

P = Probability of attribute x_i.
x_i = Attribute sought.
i = Index for the value of the attribute
j = class index.
Y = Represent the class sought
μ = The average value represented.

For variance (σ), find the equation 19 [9].

$$\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu)^2 \tag{19}$$

To classify using the Naive Bayesian method, it is necessary to calculate the average and standard deviation of each class for each characteristic. For the final stage, the test data is entered into each class to determine the opportunities that exist in each class so that it can be determined in which class the image has the greatest opportunity [9].

3. Result

In this study the authors used two databases of tobacco leaves, namely tobacco leaves from Sumenep and Pamekasan Regencies. For each district, 150 samples were taken, consisting of 50 normal (healthy) leaves, 50 leaves eaten by caterpillars (holes) and 50 curly leaves (attacked by pests). Some samples of tobacco leaves can be seen in Figures 3, 4 and 5.

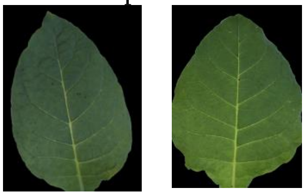

Figure 3.
Normal Tobacco Leaves

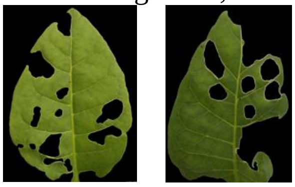

Figure 4.
Hollow Tobacco Leaves

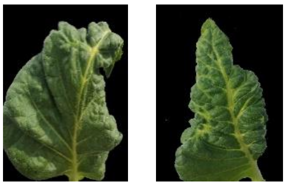

Figure 5.
Curly Tobacco Leaves

In this study there were 2 trial scenarios. Each scenario was calculated for its level of accuracy. Measurement of the level of accuracy was done by the K-Fold Cross Validation approach. The two test scenarios outline can be seen in table 1.

Table 1. Experiment Scenarios

Scenario	Feature / Attributes	Number of Features
Scenario 1	Texture, using angles 0^0	9 Features
Scenario 2	Color	9 Features

The first scenario used the texture feature by sharing the data using the k-fold cross validation approach with a 'k' of 5. The best results are at k = 3 with an accuracy value reaching 73.3% for sumenep tobacco and at k = 3 and 4 for Pamekasan tobacco with an accuracy of 71, 1% This is shown further in table 2.

Table 2. Accuracy Results Using the Texture Feature

K-Fold	**Accuracy Results**	
	Sumenep	**Pamekasan**
1	64,4%	64,4%
2	66,7%	62,2%
3	73,3%	71,1%
4	70,0%	71,1%
5	67,9%	68,9%
Average	68,5%	67,5%

The second scenario uses the color feature. By sharing the data using the k-fold cross validation approach with a 'k' of 5. The best results are at k = 1 with an accuracy value reaching 84.4% for sumenep tobacco and at k = 3 for pamekasan tobacco with an accuracy of 82.2% . This is shown in more detail in table 3.

Table 3. Accuracy Results Using the Color Feature

K-Fold	**Accuracy Results**	
	Sumenep	**Pamekasan**
1	84,4%	62,2%
2	68,9%	60,0%
3	68,9%	82,2%
4	64,4%	64,4%
5	77,8%	68,9%
Average	72,9%	67,5%

From the results of trial scenario 1 and scenario 2, it was found that the highest trial results in scenario 2 for Sumenep tobacco and Pamekasan tobacco. This is because the trial data have the characteristics of the leaf color. So the trial results are better with methods that use color features than using texture features.

The average results in scenario 1 shown in table 2 and the average results in scenario 2 shown in table 3, show that trials using Sumenep tobacco data are better than Pamekasan tobacco. This is because the trial data for Pamekasan tobacco was affected by the disease are not only holes in the middle of the leaves, but there are some leaves that are bitten by caterpillars on the edges. At the time of testing the leaves that were bitten by caterpillars on the edges were identified as normal (healthy) leaves.

4. Conclusion

From the research that has been done, several conclusions can be drawn including:

1. The use of color attributes with the CM method produces better accuracy than texture attributes with an accuracy rate of 84.4%.
2. The use of texture attributes in the classification of diseases in tobacco leaves using the GLCM method has results in lower accuracy than the CM method with the highest rate of only 73.3%.
3. From the experimental data it can be concluded that the use of Sumenep tobacco data has better accuracy than using Pamekasan tobacco data.

References

[1] M. Cormac,"image processing for tobacco grading in Zimbabwe", in *In Proceedings of IEEE International Symposium on Industrial Electronics*, pp. 327–331, 1993.

[2] D. S. Guru, P. B. Mallikarjuna, and S. M, "Segmentation and classification of tobacco seedling diseases", in *in Proceedings ofthe Fourth Annual ACM Bangalore Conference. ACM*, p. 32, 2011.

[3] Zhang, X. and Zhang, F, "Classification and Quality Evaluation of Tobacco Leaves Based on Image Processing and Fuzzy Comprehensive Evaluation", *Sensors*, 11(1), pp. 2369–2384. doi: 10.3390/s110302369, 2011.

[4] Avila-george, H., Valdez-morones, T. and Humberto, P, "Using Artificial Neural Networks for Detecting Damage on Tobacco Leaves Caused by Blue Mold", in *International Journal of Advanced Computer Science and Applications,* pp. 579–583, 2018.

[5] Suresha, M., Shreekanth, K. N., dan Thirumalesh, B. V., "Recognition of Diseases in Paddy Leaves Using kNN Classifier", *International Conference for Convergence in Technology*, pp. 663–666, 2017.

[6] Patil, J. K., dan Kumar, R. "Color Feature Extraction of Tomato Leaf Diseases", *International Journal of Engineering Trends and Technology*, vol. 2, no. 2, pp. 72–74, 2011.

[7] Kaushal, G., dan Bala, R., "GLCM and KNN based Algorithm for Plant Disease Detection", *International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering*, vol. 6, no. 7, pp. 5845–5852, 2017.

[8] Harshavardhan, A., Babu, S., dan Venugopal, T., "Analysis of Feature Extraction Methods for the Classification of Brain Tumor Detection", *International Journal of Pure and Applied Mathematics*, vol. 117, no. 7, pp. 147–155, 2017.

[9] Muntasa, A, "*Pengenalan Pola",* Yogyakarta: Graha Ilmu, 2015.

[10] Andono, P. N., Sutojo, T., dan Muljono, "*Pengolahan Citra Digital"*, 1st ed. Yogyakarta: ANDI(Anggota IKAPI), 2017.

[11] Herusutopo, A., Zuhrudin, R., Wijaya, W., dan Musiko, Y., "Recognition Design of License Plate and Car Type Using Tesseract Ocr and Emgucv", *International Journal of Communication and Information Technology*, vol. 6, no. 2, pp. 76–84, 2012.

[12] Yu, H., Li, M., Zhang, H. J., dan Feng, J., "Color Texture Moments For Content-Based Image Retrieval," *International Conference on Image Processing*, pp. 929–932, 2002.

[13] Sari, I. P., "Perancangan dan Simulasi Deteksi Penyakit Tanaman Jagung Berbasis Pengolahan Citra Digital Menggunakan Metode Color Moments dan GLCM", *Seminar Nasional Inovasi Dan Aplikasi Teknologi Di Indonesia*, pp. 215–220, 2016.