Journal of Physics: Conference Series

PAPER • **OPEN ACCESS**

Classification of Corn Diseases using Random Forest, Neural Network, and Naive Bayes Methods

To cite this article: A Ubaidillah *et al* 2022 *J. Phys.: Conf. Ser.* **2406** 012023

Classification of Corn Diseases using Random Forest, Neural Network, and Naïve Bayes Methods

A Ubaidillah[1]*, E M S Rochman[2], D A Fatah[3], A Rachmad[4]

[1-4]Departemen of Informatics, Faculty of Engineering, University of Trunojoyo Madura, Bangkalan, Indonesia

Corresponding author's e-mail: Ubaidillah.ms@trunojoyo.ac.id

Abstract. Corn is one of the staple foods consumed by many people after rice plants, especially in Indonesia. High consumer demand requires corn production in large quantities to meet these needs. However, corn production is not always in large quantities due to several factors, namely diseases in corn plants. Unhealthy corn plants can reduce the amount of production. Healthy and unhealthy corn plants can be identified manually, but this method is not efficient, so in this study, it is proposed to classify corn diseases using the Random Forest, Neural Network, and Nave Bayes methods. The dataset used is a collection of corn leaf images taken from farmers' fields in the Madura Region with four target classes, namely healthy, gray leaf spot, blight, and common rust. Based on the test results, the classification using the Neural Network method provides a better accuracy value than the other two methods in classifying corn leaf datasets, namely the AUC value reaches 90.09%, classification accuracy is 74.44%, f1 score is 72.01%, precision is 74.14% and recall by 74.43%.

1. INTRODUCTION

Corn is one of the staple foods consumed by many people after rice plants, especially in Indonesia [1]. High consumer demand requires corn production in large quantities to meet these needs. But on the other hand, several obstacles can trigger a decrease in corn production, one of which is the presence of diseases that attack the leaves of corn plants which result in crop failure if not handled quickly and appropriately [2]. Diseases that generally attack corn plants are gray leaf spot, which is a disease on corn leaves that causes brownish yellow spots from the fungus Helminthoporium maydis, leaf blight disease or known as corn blight [3], and common rust disease, namely rust disease. leaves [4].

Prevent crop failure, it can be done by monitoring corn plants for diseases that corn plants are susceptible to. Corn farmers generally only carry out manual monitoring so that because of the time required using this method it takes a long time to result in crop failure because it is not handled immediately. Based on these problems, this study utilizes the science of digital image classification which aims to identify diseases in corn plants by categorizing corn plant leaves into classes that have certain characteristics [5]. Digital image classification can be done using machine learning methods which are currently very popular.

There are many machine learning methods used to classify and analyze digital images, such as Random Forest, and Naïve Bayes. Not only that but classification can also be solved using deep learning methods, one of which is an Artificial Neural Network. So based on this explanation, this study applies several different Random Forest algorithms to classify leaves of corn plants and compare the performance results obtained from the algorithms proposed in this study. In 2016, research was conducted using the Neural Network method used to classify corn plant pests. The data used is questionnaire data with a sample of 100 corn farmers. The accuracy results obtained using the Neural Network method are 96.60% [6]. In 2018, a study was carried out that detected diseases on leaves of corn plants, for the use of the RGB extraction method and the random forest classification method, with an accuracy of about 87% [7] was carried out. In 2019, a study was conducted to detect disease in sunflower leaves using the HSV and GLCM extraction methods and then using the Random Forest classification method which obtained an accuracy of 95% [8]. In 2022, research was carried out on corn leaves through the Support Vector Machine (SVM) method approach by utilizing color features and texture features. The number of features used is 22 features consisting of 18 color features and 4 texture features. The accuracy results obtained are 99.5% [9].

Based on the description above, the author proposes a system to classify diseases on corn leaves based on corn leaf images using the Histogram of Oriented Gradients (HOG) feature extraction method and the Neural Networks classification method.

2. METHODOLOGY

2.1. Feature Extraction

This research involves a feature extraction process, feature extraction is a process that aims to remove noise in the image to maximize the accuracy obtained [10]. In this study, feature extraction was carried out using the Histogram of Oriented Gradients (HOG) method and feature extraction at the statistical level.

2.2. Histogram of Oriented Gradients (HOG)

Histogram of Oriented Gradients is a method used in image processing that can aim to detect objects in the image [11]. Each pixel in the cell has a different histogram value based on the value generated in the gradient calculation [12]. Cells have a size of 4x4 pixels and blocks have a size of 2x2 cells or the equivalent of 8x8 pixels in an image. The following is an illustration of the description shown in Figure 1.

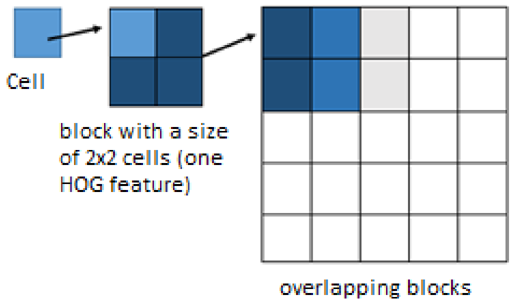


Figure 1. HOG block illustration

Based on Figure 1 above, the block is said to be a HOG feature. The steps of the HOG method are described below.

a. Specifies the block and cell sizes.
b. Calculates the gradient to obtain the outline of the object in the image. To calculate the gradient, the following equation can be used.

$$|G| = \sqrt{lx^2 ly^2} \quad (1)$$

$$\theta = \tan - 1(ly)(lx) \quad (2)$$

Information:

$|G|$: big gradient
θ : a measure of angle
ly : matrix row
lx : column matrix

c. Specifies the bin orientation to divide the image into several smaller areas called cells.
d. Normalize the block to avoid contrast and illumination variations in the image based on the cell gradient neighbor values. Histogram normalization can be done using the following equation.

$$Vn = \sqrt{\frac{v}{||v|| + \varepsilon}} \tag{3}$$

Information:

Vn : normalized vector value
v : unnormalized vector value
$|v|$: vector norm v
ε : constant with a small value to avoid division by zero.

e. Finally, combine the results of block normalization into a single vector to obtain a HOG feature vector.

2.3. Statistics Level Feature

Statistical feature extraction is one of the methods used in performing feature extraction on digital classification. This method uses a statistical calculation of the histogram distribution by measuring the contrast, granularity, and area roughness of the neighboring relationship between one pixel and another pixel in the image [13]. From the resulting histogram value, it can be calculated with several characteristic parameters, namely the mean, standard deviation, kurtosis, and skewness [14]. The following is a description of some of these characteristic parameters.

a. The mean is one of the characteristic parameters that shows the size of the central tendency or the value of the observation data center.

$$m = \sum_{i=0}^{L-1} p(i) \tag{4}$$

b. Standard deviation is a feature parameter that provides information about the size of the contrast level of an object.

$$\sigma = \sqrt{\sum_{i=0}^{L-1} (i-m)^2 \; p(i)} \tag{5}$$

c. Skewness is the degree of non-uniformity of distribution or it can also be said as a measure of the slope.

$$skewness = \sum_{i=0}^{L-1} (i-m)^3 \; p(i) \tag{6}$$

d. Kurtosis is the opposite of skewness which shows the uniformity of the distribution which is generally taken relative to a normal distribution.

$$m_k = \sum_{i=0}^{L-1} (i-\mu f)^k \; p(i) \tag{7}$$

2.4. Classification Process

In this study, digital image classification was carried out using several classification methods, namely Random Forest, Neural Network, and Naïve Bayes. The following is a description of each of these algorithms.

2.4.1. Random Forest

Random Forest is a supervised learning algorithm that is one of the algorithms using ensemble techniques (applying bagging and random feature selection methods). The ensemble learning used aims to overcome the problem of unstable classification by combining some basic learning to reduce errors during prediction. This method works by making modeling using several decision trees or in other words, namely a collection of several decision trees [15]. Where in each tree there is an estimated classification that we can consider (called a vote) to combine each possibility from each tree, then choose the classification that has the greatest number of classifications to produce optimal and stable predictions [16]. For more details, here are the stages of the Random Forest algorithm.

1. The first stage is to create a bootstrap sample randomly with a size return of n in the data cluster.
2. Determine the value of m randomly from p where the determination of the predictor variable m is chosen randomly, with m << p.
3. For a response to a prediction observation, it is done by combining the results of the prediction of as many as k trees based on the majority vote (majority vote). In the process of determining

the majority vote on a tree, then the first step is to calculate the entropy value using equations (2) and (3).

$$entropy(S) = \sum_{i=1}^{c} p_i log_2 p_i \quad (8)$$

Information:

S : a set of datasets
c : the number of classes
pi : the probability of class i frequency in the dataset

$$entropy(T, X) = \sum P(c)E(c) \quad (9)$$

Information:

(T, X) : features T and X
P(c) : the probability of feature class
E(c) : entropy result of the feature class

2.4.2. Neural Network

Artificial Neural Network (ANN) is a knowledge engineering concept in artificial intelligence developed by taking over the human nervous system, with the main human nervous system located in the brain [17] [18]. The smallest part of the human brain is the nerve cell, which is the basic processing unit. This unit is often referred to as a neuron. By using these neurons simultaneously, the human brain can process information in parallel and fast, even faster than today's fastest computers.

A biological neuron consists of a cell body (soma), a series of fibers that carry information to the neuron (dendrites), and a single fiber that exits the neuron (axon). It travels through synapses to neurons and then to somatic cells, then is released through action and re-examined by other neurons or outputs as the result of brain processes. By analogy with the working system of the human brain [19]. ANN consists of processing units, neurons (axons in the case of the human brain), auxiliary and activation functions, some weights (synapses in the human brain), and several input vectors (dendrites) [20]. The activation function helps and regulates the work of neurons, there are several kinds of activation functions, namely linear, step, binary sigmoid, and bipolar sigmoid.

There are two types of architecture in the Neural Network algorithm, namely single-layer perceptron and multilayer perceptron. The following are the steps of the Neural Network method with multilayer perceptron.

1. The first step is to initialize the weight and bias values with random values between 0 to 1.
2. The second step is to calculate the input value from the input layer to the hidden layer using the equation below:
$$input_j = \sum_{i=1}^{n} O_i W_{ij} + \theta_j \quad (10)$$
3. In the third step, the output from the hidden layer to the output layer is obtained using the specified activation function.
4. The next step is to calculate the prediction error (error) using the following equation.
$$Error_j = Error_j . (1 - Output_j) . (Target_j - Output_j) \quad (11)$$
5. The fifth step is to calculate the error value in the hidden layer using the equation below.
$$Error_j = Error_j . (1 - Output_j) \sum_{i=1}^{n} Error_k W_{jk} \quad (12)$$
6. The weight value will continue to be updated using the previous error value.
7. If the maximum limit of the specified loop has been reached or the error value is 0, then the loop will stop.

2.4.3. Naive Bayes

Naïve Bayes Classifier is a technique for classifying data derived from the Bayes theorem. This method can predict future data based on previous data or existing data by calculating opportunities or probabilities from test data with data stored in the training process [21]. The main feature of the Naïve Bayes method is a strong assumption of independence from each condition [22]. The following equation is used for the calculation of Naïve Bayes:

$$P(Y|Z) = \frac{P(Y)}{P(Z)} P(Z|Y) \tag{13}$$

Information:

Y : Data without unknown categories
Z : Hypothesis data
P(Y|Z) : Probability of hypothesis Y to condition Z
P(Z) : Probability Z
P(Z|Y) : Probability of Z based on condition on Y
P(Y) : Probability Y

2.5. Evaluation

The evaluation of the model in this study uses the confusion matrix method. The confusion matrix is a method that can be used to measure the performance of the classification model. The confusion matrix displays the performance results of the classification model in the form of a matrix [23]. Based on the confusion matrix, there are 4 possible results from the comparison between the prediction class and the actual class, namely true positive (TP), false positive (FP), false negative (FN), and true negative (TN). The matrix can be seen in the image below.

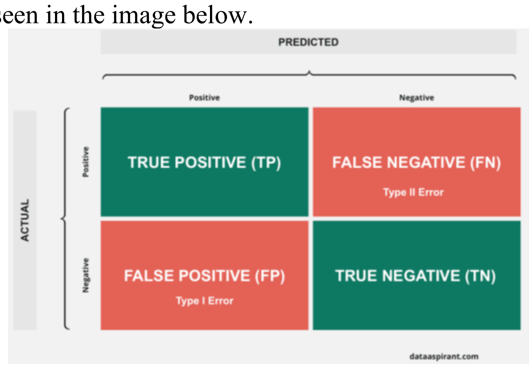


Figure 2. Confusion Matrix

With the confusion matrix as shown in the figure 2 above, accuracy, recall, precision, f1 scores, and AUC values can be generated. Accuracy is Recall i.e., F-measures [24]. In this study, the value of AUC (area under the curve) will also be calculated or it can be called probability which is a method for calculating under the ROC curve, where the higher the AUC value, the classification method used can be applied properly in a study [25].

1. Accuracy is the degree of closeness between the actual result and the predicted result
$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{14}$$
2. Precision is the level of conformity of the answers generated from the system with the information desired by the user.
$$Precision = \frac{TP}{TP + FP} x\ 100\% \tag{15}$$
3. A recall is the level of success of the system in finding back information.
$$Recall \quad = \frac{TN}{TN + FN} x\ 100\% \tag{16}$$
4. F1 score is the average result of the combination of recall and precision calculations.
$$F1\ score = 2\ x \frac{Recall\ x\ Precision}{Recall + Precision} x\ 100\% \tag{17}$$
5. AUC is a method for calculating under the ROC curve where the higher the AUC value, the classification method used provides better accuracy.

$$AUC = \frac{1}{2} x \left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right) \quad (18)$$

Information:

TP : the number of data correctly predicted positive,
TN : the number of data with the original class is positive but the prediction result is negative,
FN : the number of data correctly predicted negative,
FP : the number of data with the original class is negative but the prediction result is positive.

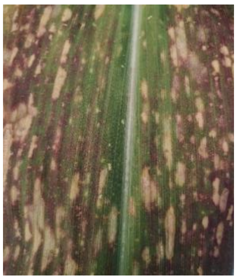
3. Result and Analysis

3.1. Data collection

This study used a disease dataset of 3500 maize leaves obtained from the Kaggle.com website https://www.kaggle.com/datasets/arhasnaazzahra/cornleavediseases. The dataset contains a collection of data in the form of images of corn plant leaves from farmers' fields in the Madura region uploaded by A.R Hasna Azzahra. The total target classes in this dataset are four, namely Healthy, Gray Leaf Spot, Blight, and Common Rush. The dataset used in this study can be seen in Table 1.

Table 1. Description of Target Class on Dataset

Dataset	Target	Description
 **Figure 3**. Healthy	Healthy	Figure 3 below is an example of a dataset of a healthy maize leaf class.
 **Figure 4.** Grey leaf Spot	Grey Leaf Spot	Gray leaf spot is a disease on corn leaves that comes from the fungus Helminthoporium maydis. As seen in Figure 4, this disease is characterized by brownish-yellow spots on the leaves.
 **Figure 5.** Blight	Blight	Figure 5 shows an example of a corn leaf infected with Leaf Blight, also known as corn blight.

 **Figure 6.** Common Rust	Common Rust	Figure 6 is an example of a dataset that is categorized into the common rust class, namely leaf rust.

3.2. Result

In this section, it will be explained how the research was carried out using the proposed method, as a solution to the problems raised. The stages of the process will be visualized in the Input-Process-Output diagram below.

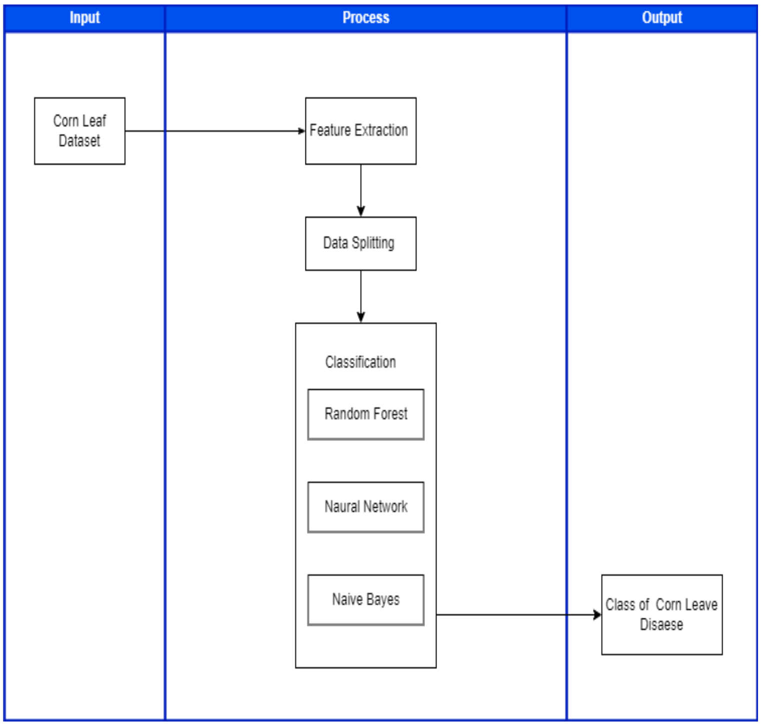


Figure 7. IPO Diagram

Based on Figure 7 above, the process of making a classification system is divided into several parts, namely, the input process, the core process, and the output process. The description of the three parts of the stages is as follows.

1. Data Input Process
   The input section is a process for entering data to be classified in this study, where the input data used is a dataset in the form of a corn plant leaf image of 3500 records and 4 classes or targets.
2. Feature Extraction
   In this study, feature extraction is carried out using two types of methods, namely feature extraction with Histogram of Oriented Gradients (HOG) and statistical feature extraction with the histogram results obtained will calculate several statistical feature parameters, namely the mean, standard deviation, kurtosis, and skewness.

3. Data Sharing Process
   At this stage, the training data and test data are divided using the data splitting method. Training data is used to build a model that is formed with a certain amount of data and for test data taken from the rest of the data that is not used in the training process which is used to test the performance of the model that has been trained. The comparison of training data applied in this study is 80% training data and 20% test data.
4. Classification Process
   At the classification stage, a learning process is carried out to build a system model using several algorithms, namely Random Forest, Neural Network, and Naïve Bayes. With the different algorithms used, it aims to find out which algorithm gives the best accurate results.
5. Output
   The output produced after the entire process is run is in the form of class predictions of diseases on corn leaves.

3.3. Analysis

In this study, from 3500 records of corn leaf imagery that have been carried out feature extraction using the HOG method and statistical feature extraction. The results of HOG feature extraction and Statistics produce feature extraction values such as Mean, Standard deviation, Kurtosis, and Skewness. The mean value of corn leaves has the lowest value of 0.06 and the highest value of 0.14. The Standard deviation has a value ranging from 0.08 to 0.15. while the lowest Kurtosis value is 0.36 and the highest is 13.39. while the lowest Skewness value is 0.82 and the highest value is 3.22.

Then the Data will be divided into training data and test data using the data splitting method with a ratio of 80:20, namely 80% as training data and 20% as test data to obtain the results of the evaluation of the classification model using 3 different algorithms. The results of the method performance evaluation can be seen in Table 2 below.

Table 2. Test Result

Model	AUC	Accuracy	F1 Score	Precision	Recall
Random Forest	88.14 %	69.76 %	69.28 %	69.00 %	69.76 %
Neural Network	**90.09 %**	**74.44 %**	**72.01 %**	**74.14 %**	**74.43 %**
Naïve Bayes	83.28 %	60.25 %	56.04 %	56.53 %	60.25 %

Table 2 shows the test results from the three methods, it can be seen that the use of the Neural Network method has the best performance with the AUC value reaching 90.09%, then the classification accuracy is 74.44%, f1 score is 72.01%, precision is 74.14% and recall is 74.43 %.

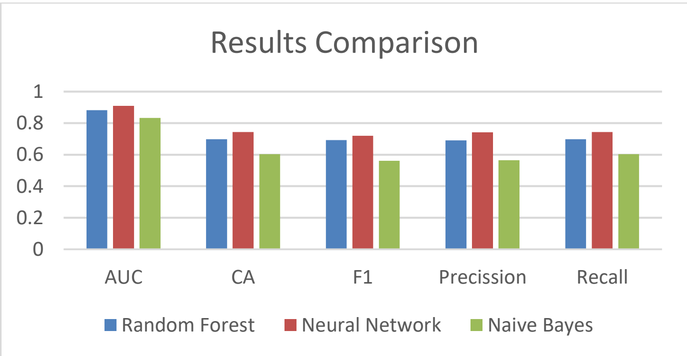


Figure 8. Evaluation Comparison Results

Based on Figure 8 above, it can be seen that the application of the Neural Network method is superior to the other two methods. This is because this method has more complex stages in classifying data objects using 3 processing layers, namely the input layer, hidden layer, and output layer.

4. Conclusion

The conclusion that can be drawn based on the research conducted is that the classification using the Neural Network method with feature extraction HOG provides a better accuracy value compared to the other two methods (Random Forest and Naïve Bayes) in classifying corn leaf datasets totaling 3500 records into 4 target classes, namely the AUC value reaches 90.09%, classification accuracy is 74.44%, f1 score of 72.01%, precision of 74.14%, and recall of 74.43%. Thus, the application of the Neural Network method is very capable of classifying digital images properly.

References

[1] K. Hidayat, M. K. Nasikin and Rakhmawati, "Product development of corn rice using value engineering method," IOP Publishing, In IOP Conference Series: Earth and Environmental Science, vol. 733, pp. 1-8, 2021.

[2] Y. Sulong, A. J. Zakaria , S. Mohamed, M. H. Sajiili and N. Ngah, "Survey on Pest and Disease of Corn (Zea Mays Linn) grown at BRIS Soil Area," JOURNAL OF AGROBIOTECHNOLOGY, vol. 10, no. 1S, pp. 75-87, 2019.

[3] L. O. Lopez-Zuniga, P. Wolters, S. Davis, T. Weldekidan, J. M. Kolkman, R. Nelson, K. S. Hooda, E. Rucker, W. Thomason, R. Wisser and P. Balint-Kurti, "Using Maize Chromosome Segment Substitution Line Populations for the Identification of Loci Associated with Multiple Disease Resistance," G3 Genes|Genomes|Genetics, vol. 9, no. 1, pp. 189-201, 2019.

[4] Sopialena, Suyadi and S. A. Noor, "Ecosystem Monitoring on Leaves of Leaf Rust Disease of Maize (Zea mays L.)," Journal of Sustainable Agriculture, vol. 37, no. 1, pp. 89-99, 2022.

[5] J. Al-doski, S. B. Mansor and H. Z. M. Shafri, "Image Classification in Remote Sensing," Journal of Environment and Earth Science, vol. 3, no. 10, pp. 141-147, 2013.

[6] Rais. "Klasifikasi Hama dan Penyakit Tanaman Jagung dengan Menggunakan Neural Network Berbasis Algoritma Genetika." *Prosiding Seminar Nasional IPTEK Terapan (SENIT) 2016 Pengembangan Sumber Daya Lokal Berbasis IPTEK*. Vol. 1. No. 1. 2016.

[7] M. Amini, Y. Resti, & D. A. Zayanti, Identifikasi Hama Dan Penyakit Pada Tanaman Jagung Menggunakan Metode Random Forest Dan Fuzzy Decision Tree (Doctoral Dissertation, Sriwijaya University). 2021.

[8] A. Salsabila, R. Yunita, and C. Rozikin. "Identifikasi Citra Jenis Bunga menggunakan Algoritma KNN dengan Ekstrasi Warna HSV dan Tekstur GLCM." *Technomedia Journal* 6.1 Agustus (2021): 124-137.

[9] R. Suhendra, "Identifikasi dan Klasifikasi Penyakit Daun Jagung Menggunakan Support Vector Machine." Jurnal Teknologi Informasi 1.1, 2022.

[10] T. G. Nisia and S. Rajesh, "Extraction of High-level and Low-level feature for classification of Image using Ridgelet and CNN based Image Classification," IOP Publishing, In Journal of Physics: Conference Series, vol. 1911, no. 1, pp. 1-6, 2021.

[11] D. Rahmadani, A. Virgono and C. Setianingsih, "Food Detection Using Histogram of Oriented Gradient (HOG) as Feature Extraction and K-Nearest Neighbors (K-NN) as Classifier," International Journal of Advanced Trends in Computer Science and Engineering, vol. 9, no. 5, pp. 219-225, 2020.

[12] S. V. Shidlovskiy, A. S. Bondarchuk, S. Poslavsky and M. V. Shikhman, "Reducing dimensions of the histogram of oriented gradients (HOG) feature vector," Journal of Physics: Conference Series, vol. 1611, no. 1, pp. 1-7, 2020.

[13] M. Wati , Haviluddin, N. Puspitasari, E. Budiman and R. Rahim, "First-order Feature Extraction Methods for Image Texture and Melanoma Skin Cancer Detection," IOP Conf. Series: Journal of Physics, vol. 1230, no. 1, pp. 1-9, 2018.

[14] V. Y. Kulkarni and D. P. K. Sinha, "Random Forest Classifier: A Survey and Future Research Directions," International Journal of Advanced Computing, vol. 36, pp. 1144-1153, 2013.

[15] A. Rachmad, N. Chamidah, & R. Rulaningtyas. (2020). Mycobacterium tuberculosis identification based on colour feature extraction using expert system. *Ann. Biol*, *36*, 196-202, 2020

[16] U. Kumari and G. Aggarwal, "Artificial Neural Networks," International Journal of Engineering and Applied Sciences (IJEAS), vol. 4, no. 5, pp. 64-68, 2017.

[17] A. Rachmad, E. M. S. Rochman, D. Kuswanto, I. Santosa, R. K. Hapsari, T. Indriyani, & E. Purwanti. Comparison of the traveling salesman problem analysis using neural network method. In *International Conference on Science and Technology (ICST 2018)*, pp. 1057-1061, 2018. Atlantis Press.

[18] W. J. Park and J. B. Park, "History and application of artificial neural networks in dentistry," European Journal of Dentistry, vol. 12, no. 04, pp. 594-601, 2018.

[19] P. L. Prasanna and D. R. Rao, "Text classification using artificial neural networks," International Journal of Engineering & Technology, vol. 7, no. 1, pp. 603-606, 2018.

[20] D. R. Anamisa, A. Rachmad, M. Yusuf, A. Jauhari, R.D.T. Erdiansa, & M. Y. Hariyawan. (2021).

Classification of diseases for rice plant based on Naive Bayes classifier with a combination of PROMETHEE. *Commun. Math. Biol. Neurosci., 2021*

[21] 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." *Journal of Physics: Conference Series*. Vol. 1477. No. 5. IOP Publishing, 2020

[22] D. Berrar, "Bayes' Theorem and Naive Bayes Classifier," Encyclopedia of Bioinformatics and Computational Biology, vol. 1, pp. 403-4012, 2019.

[23] A. Rachmad, N. Chamidah, & R. Rulaningtyas. Mycobacterium tuberculosis identification based on colour feature extraction using expert system. Ann. Biol, 36, pp.196-202, 2020.

[24] 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., 2020.

[25] M. H. Ferris, M. McLaughlin, S. Grieggs, S. Ezekiel, E. Blasch, M. Alford, M. Cornacchia and A. Bubalo, "Using ROC Curves and AUC to Evaluate Performance of No-Reference Image Fusion Metrics," in IEEE National Aerospace and Electronics Conference, Pennsylvania, 2015.

.