IAIC International Conference Series
Vol. 4, No. 1, 2023, pp. 165 ~ 171
E-ISSN: 2774-5899 | P-ISSN: 2774-5880
DOI: https://doi.org/10.34306/conferenceseries.v4i1.649

Comparative Analysis of the Performance of the Decision Tree and K-Nearest Neighbors Methods in Classifying Coffee Leaf Diseases

[1*]Suryadi, [2]Murhaban, [3]Rivansyah Suhendra
[1,2,3]Department of Information Technology, Teuku Umar University, Indonesia
E-mail: [1*]suryadi@utu.ac.id, [2] murhaban@utu.ac.id, [3] rivansyahsuhendra@utu.ac.id
***Corresponding author**

(Received November 8, 2023 Revised November 17, 2023 Accepted November 25, 2023, Available online December 19, 2023)

Abstract

This study aimed to develop and compare classification models utilizing Decision Tree and K-Nearest Neighbors (KNN) in the detection of diseases in coffee leaf images. The dataset comprises coffee leaf images categorized into four different disease types, namely Nodisease, Miner, Phoma, and Rust. To facilitate model training and testing, the dataset was divided into training and validation data using a cross-validation approach. Both the Decision Tree and KNN models underwent meticulous parameter tuning. The experimental results reveal that the Decision Tree model achieved an accuracy rate of 98.20% on the validation data, while the KNN model achieved an accuracy rate of 75.01%. Furthermore, the Decision Tree model exhibited an AUC of 0.9879, recall of 0.9820, precision of 0.9835, and an F1-score of 0.9819 on the validation data. Conversely, the KNN model achieved an AUC of 0.9465, recall of 0.7501, precision of 0.7569, and an F1-score of 0.7485. These findings suggest that the Decision Tree model surpasses the KNN model in accurately detecting coffee leaf diseases, as demonstrated by higher accuracy and other evaluation metrics. However, the relevance of the KNN model remains contingent on application requirements and modeling preferences. These outcomes may contribute to the development of automated systems for disease detection in coffee plants, ultimately promoting more sustainable agricultural practices.

***Keywords**: Coffee Leaf, Data Mining, Classification, Decision Tree, K-Nearest Neighbor*

1. Introduction

Coffee is one of the most important agricultural commodities in Aceh, and is also one of the main sources of income for farmers in the region. However, coffee production in Aceh often faces serious challenges in efforts to maintain high productivity and quality of coffee beans. One of the main problems affecting coffee plants is disease attacks on coffee leaves [1].

Coffee leaf diseases are a serious threat to coffee plants, resulting in reduced yields and quality of coffee beans. Manual identification of coffee leaf diseases by farmers is often difficult and time consuming, which can result in delays in appropriate treatment measures [2].

In an effort to overcome this problem, it is necessary to detect and classify coffee leaf diseases efficiently. Classification of coffee leaf diseases is a key step in effective plant disease monitoring and management. One approach that can be used to classify coffee leaf diseases is using digital image processing techniques and machine learning algorithms [3].

The Decision Tree and K-Nearest Neighbor (K-NN) methods are two methods commonly used in image classification and pattern recognition. Decision Tree is a machine learning method that can produce decision tree-based models for classification. Meanwhile K-NN is a distance-based classification method that uses proximity to nearest neighbors to classify objects [4].

Several previous studies have been carried out regarding the identification of leaf diseases in plants, as follows. Purnamawati et al in 2020 conducted research to detect leaf diseases in rice plants using the decision tree algorithm with 100% accuracy results and using the KNN algorithm with 84% accuracy results [5]. Kusuma et al in 2022 also conducted research on the classification of leaf diseases in corn plants using several algorithms, one of the algorithms used was k-nearest neighbor (KNN) which obtained accuracy results of 93.8%, precision of 93.9%, and recall 93.8% [6]. Wahyuningtyas et al in 2022 conducted research on disease identification on coffee leaves using the Local Binary Pattern and Random Forest methods. The results of this research using the Random Forest algorithm obtained accuracy results of 95.83% [7]. Wildah et al in 2023 conducted research related to the classification of coffee leaf diseases using random forest, obtaining accuracy results of 98% [8]. Based on previous research, the Decision Tree and KNN (K-Nearest Neighbor) algorithms have been successfully used to classify leaf diseases in plants by obtaining optimal curation results [9].

Based on related research, this research will compare the performance of the Decision Tree algorithm with the K-Nearest Neighbor algorithm for classifying images of leaf diseases on coffee plants [11]. This research was conducted to overcome the problem of diseases in coffee leaves which can result in a decrease in crop yields and coffee bean quality, as well as financial losses for farmers [12]. By developing an image-based classification model, this research has the potential to help farmers detect diseases in coffee plants early, allowing them to take timely action. The impact on the development of the coffee industry has been very positive, with increased productivity, coffee bean quality and a more sustainable approach to farming as a result [13].

2. Research Method

In the methodology section of this study, we will present the approach we used in detail. The research flow that we implemented can be seen in Figure 1, which will provide a comprehensive overview of the steps we carried out in this research.

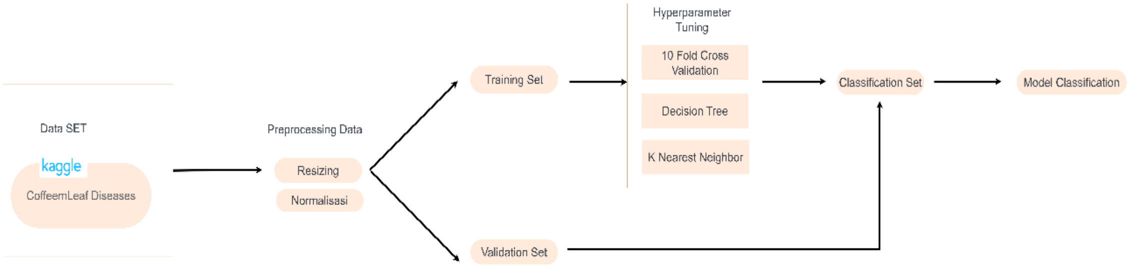
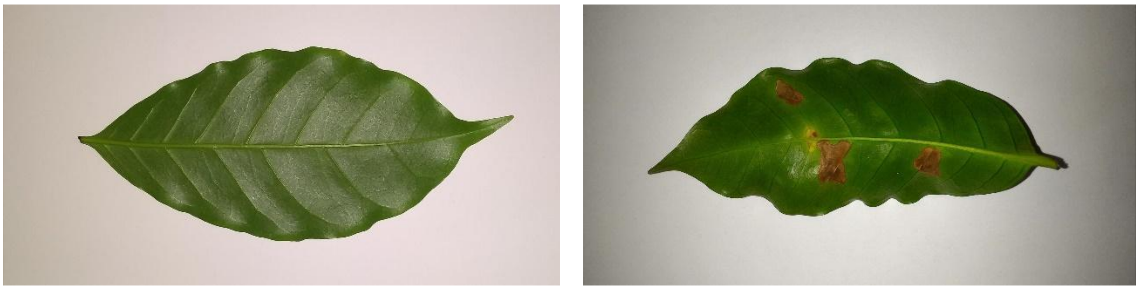


Figure 1. Flow of Research Methodology

2.1. Dataset

In this study, we used the coffee leaf diseases dataset, which was specially curated for the detection of coffee leaf diseases [14]. This dataset consists of 3 channel images with a resolution of 2048 × 1024 pixels. An overview of the dataset can be seen in Figure 2. This dataset focuses on the problem of detecting healthy and diseased coffee leaves [15]. This dataset is divided into 80% and 20% for training and validation purposes. The training set was further divided into 80% for training and 20% for validation. The distribution of images in each set can be seen in Table 1.



a b

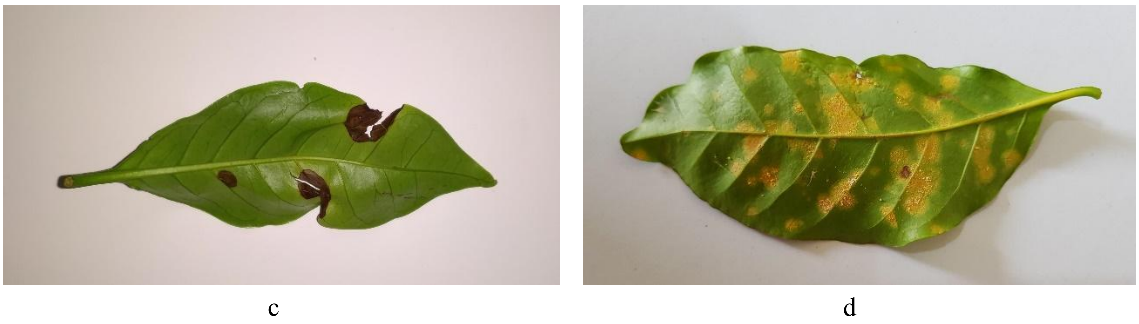


Figure 2. Dataset Used: (a) Class Nodisease, (b) Class Miner, (c) Class Phoma, (d) Class Rust

Table 1. Distribution of Data for Each Class

	Nodisease	Miner	Phoma	Rust	Total
Training	227	266	310	208	1011
Validation	57	66	78	52	253
Total	284	332	388	260	1264

2.2. Preprocessing Data

The data preprocessing stage is carried out by changing the image size (resizing) and normalizing the data. To prepare the dataset for training, all images were resized to a uniform size of 410 × 205 pixels (20% of the initial size). By changing the image size, the training process can be carried out more efficiently and quickly. Data normalization is carried out by changing each RGB value with a range of 0-255 into a range of 0-1.

2.3. Data Feature Extraction

In this research, feature extraction was carried out on each image. A total of 12 color features have been successfully extracted from the RGB and CMY color spaces using two statistical components, namely the mean and standard deviation [16]. The color feature is used because color is an important aspect of an image. In the context of detecting coffee leaf diseases, information about color changes in coffee leaves can be an important clue for identifying certain diseases [17]. For example, changing leaf color may be an early indication of an infection or health problem with the plant.

2.4. Image Classification Methods

In this research, we built a Decision Tree and K-Nearest Neighbors (KNN) model to classify coffee leaf images based on disease categories [18]. The data we use is divided into training data and validation data, with training data broken down into training data and validation data using a cross-validation approach [19].

We performed parameter tuning (hyperparameter tuning) on both models using a grid search algorithm to find the best parameters. For Decision Trees, the adjusted parameters include the split selection criteria (criterion), the maximum depth of the tree (max_depth), the minimum number of samples required to split a node (min_samples_split), and the minimum number of samples in each leaf of the tree (min_samples_leaf), in among other parameters. Meanwhile, for the KNN model, we adjust parameters such as the number of nearest neighbors (n_neighbors), distance metrics (metrics), and others [20].

After parameter tuning, we train both models using the prepared training data. Our Decision Tree and KNN models are trained to understand patterns in the training data and classify coffee leaf images based on adjusted parameters. We then measured the performance of these two models using validation data, by calculating evaluation metrics such as accuracy, AUC, recall, precision, and F1-score.

The results from testing the Decision Tree and KNN models are used to understand the extent to which these two models are effective in detecting coffee leaf diseases and to assist interpretation and taking further action in an agricultural context.

2.4. Evaluation of Classification Models

In this study, we use several evaluation metrics to assess the model performance. These metrics include accuracy, precision, recall, and F1-score. The equations for accuracy, precision, recall, and F1-score can be seen in equations 1, 2, 3, and 4, respectively [14].

$$Accuracy = \frac{TP + FN}{FP + FN + TP + TN} \quad (1)$$

$$Recall = \frac{TP}{FN + TP} \quad (3)$$

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$F1 - Score = 2\frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

3. Results and Analysis

3.1. Decision Tree and KNN models

In an effort to build an optimal Decision Tree model, we apply the cross-validation method. We used scikit-learn's DecisionTreeClassifier and performed careful parameter tuning. The result is a Decision Tree model with the most optimal parameters that we found, namely:

1. Criterion: 'gini'
2. Max Depth: There is no maximum restriction on tree depth (max_depth=None).
3. Max Features: There is no maximum limit on the number of features used in each split node (max_features=None).
4. Min Samples Leaf: The minimum number of samples required in each leaf of the tree is 1 (min_samples_leaf=1).
5. Min Samples Split: The minimum number of samples required to split a node is 2 (min_samples_split=2).
6. Random State: We use a seed (random_state=123) to ensure reproducible results.
7. Splitter: We select 'best' to select the best splitter on each node.

We also apply a cross-validation approach to build an optimal K-Nearest Neighbors (KNN) model. We used KNeighborsClassifier from scikit-learn and tried various parameters. The result is a KNN model with the most optimal parameters that we found, namely:

1. Algorithm: 'auto', which allows the KNN algorithm to select the appropriate method based on the given data.
2. Leaf Size: 30, is the leaf size used in the KNN tree structure.
3. Metric: 'minkowski', which indicates the use of the Minkowski metric to measure the distance between data points.
4. N Jobs: -1, which allows the use of all available CPU cores for parallel calculations.
5. N Neighbors: 5, namely the number of nearest neighbors used in prediction.
6. P: 2, which indicates the use of the Euclidean distance metric (p = 2).
7. **Weight**: 'uniform', which means that all neighbors have the same weight in the classification process.

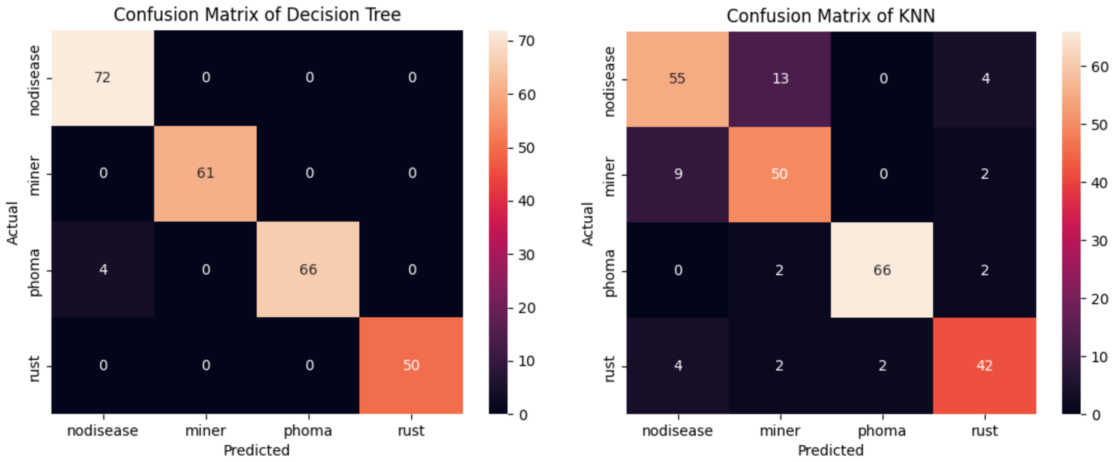
3.2. Performance Results of Decision Tree and KNN Methods

In this research, the Decision Tree model succeeded in achieving a very high level of accuracy, namely 98.20% on validation data as seen in Table 2, showing extraordinary ability in classifying coffee leaf images into the correct disease categories. In addition, the AUC value of 0.9879 illustrates the ability of this model to differentiate disease categories very well, with an ROC graph that is close to perfect. The recall of 0.9820 also indicates that the Decision Tree model is able to detect coffee leaf diseases with high accuracy, with a very low false negatives rate. Precision of 0.9835 confirms that the Decision Tree model has a low classification error rate when predicting disease, so that most of the positive predictions made by this model are correct. In addition, the F1-Score of 0.9819 provides an illustration of the good balance between recall and precision in the Decision Tree model, indicating the ability of this model to identify diseases well while avoiding classification errors. Information on the results of the confusion matrix can be seen in Figure 3.

Table 2. Performance of the Decision Tree and KNN Methods

Model	**Accuracy**	**AUC**	**Recall**	**Precision**	**F1-Score**
Decision Tree	0.9820	0.9879	0.9820	0.9835	0.9819
KNN	0.7501	0.9465	0.7501	0.7569	0.7485

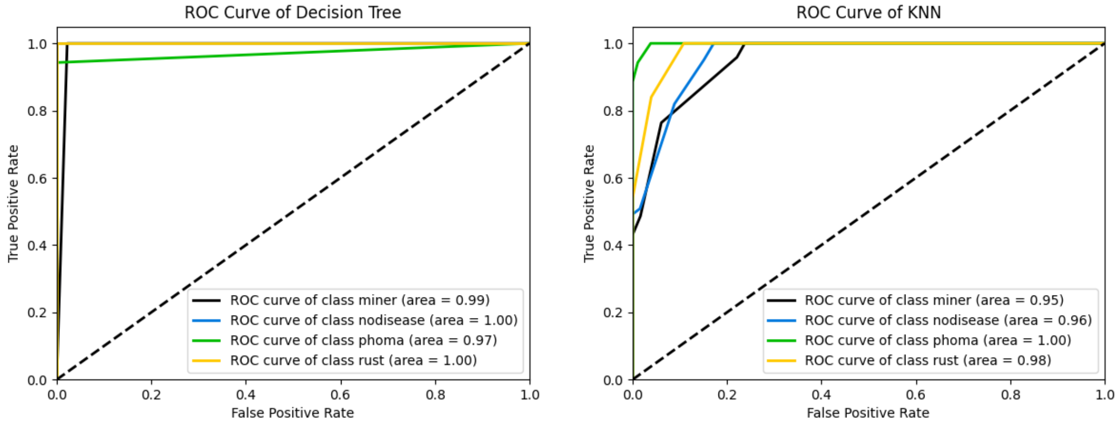
On the other hand, the K-Nearest Neighbors (KNN) model achieved an accuracy level of 75.01% on validation data. Although this is quite good accuracy, it turns out that this model is not as good as Decision Tree in terms of accuracy. The AUC of 0.9465 shows that the KNN model has good ability in distinguishing disease categories, but is less superior than the Decision Tree. A recall of 0.7501 indicates that the KNN model has a tendency to have higher false negatives compared to Decision Tree, which means that this model is less effective in detecting disease correctly. Precision of 0.7569 shows that the KNN model has a good ability to identify disease when making positive predictions, but the accuracy of this model is slightly lower than Decision Tree. The F1-Score of 0.7485 reflects the balance between recall and precision in the KNN model, but overall this result is not as good as Decision Tree. The ROC graph can be seen in Figure 4.



(a) (b)

Figure 3. Confusion Matrix Model: (a) Decision Tree, (b) KNN

Test results consistently show that the Decision Tree model provides better performance than the K-Nearest Neighbors model in the context of coffee leaf disease classification. Some factors that might explain this difference include the data structure and better use of features by Decision Tree. The main advantage of the Decision Tree model is its ability to provide easy-to-understand interpretations, which can be a valuable asset in practical applications in agriculture. Nevertheless, the KNN model still has good values in terms of accuracy and ability to differentiate diseases, but tends to have higher false negatives. These results show that although Decision Trees are superior in this case, KNN can still be a potentially good choice depending on application needs and modeling preferences. In future developments, this research may continue to consider improving the model, using better features, or even exploring deep learning methods to increase accuracy and effectiveness in coffee leaf disease detection.



(a) (b)

Figure 4. ROC Curve Model: (a) Decision Tree, (b) KNN

4. Conclusion

In conclusion, this research succeeded in developing and comparing classification models for disease detection on coffee leaves based on images. Experimental results show that the Decision Tree model significantly outperforms the K-Nearest Neighbors (KNN) model in terms of accuracy and other evaluation metrics. The Decision Tree model achieved an accuracy of 98.20% on validation data, while the KNN model achieved an accuracy of 75.01%. In terms of AUC, recall, precision, and F1-score, the Decision Tree model also shows better performance, with AUC 0.9879, recall 0.9820, precision 0.9835, and F1-score 0.9819. Meanwhile the KNN model has an AUC of 0.9465, recall of 0.7501, precision of 0.7569, and F1-score of 0.7485. These results indicate that the Decision Tree model is a more effective choice for disease detection on coffee leaves in the context of this research. However, the relevance of KNN models remains dependent on application needs and modeling preferences. These findings have the potential to contribute to the development of automated systems for detecting diseases in coffee plants, which in turn could support more sustainable and productive agriculture. In future research, the use of other methods and expanding the dataset could be the next steps to improve the performance of existing models.

References

[1] E. T. Kembaren and M. Muchsin, “Pengelolaan Pasca Panen Kopi Arabika Gayo Aceh,” *J. Visioner Strateg.*, vol. 10, no. 1, 2021.

[2] A. S. Anwar, U. Rahardja, A. G. Prawiyogi, N. P. L. Santoso, and S. Maulana, “iLearning model approach in creating blockchain based higher education trust,” *Int. J. Artif. Intell. Res*, vol. 6, no. 1, 2022.

[3] D. Irfansyah, M. Mustikasari, and A. Suroso, “Arsitektur Convolutional Neural Network (CNN) Alexnet Untuk Klasifikasi Hama Pada Citra Daun Tanaman Kopi,” *J. Inform. J. Pengemb. IT*, vol. 6, no. 2, pp. 87–92, 2021.

[4] F. Faiqotuzzulfa and S. A. Putra, “Virtual Reality’s Impacts on Learning Results in 5.0 Education: a Meta-Analysis,” *Int. Trans. Educ. Technol.*, vol. 1, no. 1, pp. 10–18, 2022.

[5] A. Purnamawati, W. Nugroho, D. Putri, and W. F. Hidayat, “Deteksi Penyakit Daun pada Tanaman Padi Menggunakan Algoritma Decision Tree, Random Forest, Naïve Bayes, SVMdan KNN,” *InfoTekJar J. Nas. Inform. dan Teknol. Jar.*, vol. 5, no. 1, pp. 212–215, 2020, [Online]. Available: https://doi.org/10.30743/infotekjar.v5i1.2934.

[6] A. Fernanda, A. R. F. Geovanni, and M. Huda, “Application of artificial intelligence to the development of playing ability in the valorant game,” *IAIC Trans. Sustain. Digit. Innov.*, vol. 4, no. 1, pp. 22–31, 2022.

[7] B. Wahyuningtyas, I. I. Tritoasmoro, and N. Ibrahim, “Identifikasi Penyakit Pada Daun Kopi Menggunakan Metode Local Binary Pattern Dan Random Forest (Identification Of Disease In Coffee Leaves Using Local Binary Pattern And Random Forest Methods),” *e-Proceeding Eng.*, vol. 8, no. 6, pp. 2972–2980, 2022.

[8] A. Hermawan, W. Sunaryo, and S. Hardhienata, “Optimal Solution for OCB Improvement Through Strengthening of Servant Leadership, Creativity, and Empowerment,” *Aptisi Trans.*

Technopreneursh., vol. 5, no. 1Sp, pp. 11–21, 2023.

[9] M. Odhiambo, “Coffee leaf diseases,” 2021. https://www.kaggle.com/datasets/badasstechie/coffee-leaf-diseases/data.

[10] M. R. Anwar and S. Purnama, “Boarding house search information system database design,” *Int. J. Cyber IT Serv. Manag.*, vol. 2, no. 1, pp. 70–81, 2022.

[11] B. P. K. Bintoro, N. Lutfiani, and D. Julianingsih, “Analysis of the Effect of Service Quality on Company Reputation on Purchase Decisions for Professional Recruitment Services,” *APTISI Trans. Manag.*, vol. 7, no. 1, pp. 35–41, 2023.

[12] M. Hardini, R. A. Sunarjo, M. Asfi, M. H. R. Chakim, and Y. P. A. Sanjaya, “Predicting Air Quality Index using Ensemble Machine Learning,” *ADI J. Recent Innov.*, vol. 5, no. 1Sp, pp. 78–86, 2023.

[13] R. Suhendra, S. Suryadi, N. Husdayanti, A. Maulana, and T. Rizky, “Evaluation of Gradient Boosted Classifier in Atopic Dermatitis Severity Score Classification,” *Heca J. Appl. Sci.*, vol. 1, no. 2, pp. 54–61, 2023, doi: 10.60084/hjas.v1i2.85.

[14] U. Rahardja, “Blockchain Education: as a Challenge in the Academic Digitalization of Higher Education,” *IAIC Trans. Sustain. Digit. Innov.*, vol. 4, no. 1, pp. 62–69, 2022.

[15] R. Windiawan and A. Suharso, “Identifikasi Penyakit pada Daun Kopi Menggunakan Metode Deep Learning VGG16,” *Explor. IT! J. Keilmuan dan Apl. Tek. Inform.*, vol. 13, no. 2, pp. 43–50, 2021.

[16] U. Rahardja, Q. Aini, P. A. Sunarya, D. Manongga, and D. Julianingsih, “The Use of TensorFlow in Analyzing Air Quality Artificial Intelligence Predictions PM2. 5,” *Aptisi Trans. Technopreneursh.*, vol. 4, no. 3, pp. 313–324, 2022.

[17] J. Kusuma, Rubianto, R. Rosnelly, Hartono, and B. H. Hayadi, “Klasifikasi Penyakit Daun Pada Tanaman Jagung Menggunakan Algoritma Support Vector Machine, K-Nearest Neighbors dan Multilayer Perceptron,” *J. Appl. Comput. Sci. Technol.*, vol. 4, no. 1, pp. 1–6, 2023, doi: 10.52158/jacost.v4i1.484.

[18] S. K. Wildah, A. Latif, A. Mustopa, S. Suharyanto, M. S. Maulana, and A. Sasongko, “Klasifikasi Penyakit Daun Kopi Menggunakan Kombinasi Haralick, Color Histogram dan Random Forest,” *J. Sist. dan Teknol. Inf.*, vol. 11, no. 1, p. 35, 2023, doi: 10.26418/justin.v11i1.60985.

[19] N. N. Azizah and T. Mariyanti, “Education and technology management policies and practices in madarasah,” *Int. Trans. Educ. Technol.*, vol. 1, no. 1, pp. 29–34, 2022.

[20] R. Suhendra, F. Arnia, R. Idroes, N. Earlia, and E. Suhartono, “A novel approach to multi-class atopic dermatitis disease severity scoring using multi-class SVM,” in *2019 IEEE International Conference on Cybernetics and Computational Intelligence (CyberneticsCom)*, 2019, pp. 35–39.