3rd International Conference on Information Science and Technology Innovation (ICoSTEC)
July 27, 2024, Yogyakarta, Indonesia

IDENTIFICATION OF COFFEE FRUIT MATURITY LEVEL USING MACHINE LEARNING BASED COLOR CLASSIFICATION WITH COMPARISON OF K-NEAREST NEIGHBOR (K-NN) AND METHOD SUPPORT VECTOR MACHINE (SVM)

Anton Purnama[1], Rika Rosnelly[2], Hartono[3]

Potensi Utama University

[1]antonpurnama515@gmail.com
[2]rika@potensi-utama.ac.id
[3]hartonoibbi@gmail.com

Abstract— **Coffee is one of Indonesia's main export commodities which has high economic value. The maturity of coffee berries is an important factor in determining the quality and price of coffee, therefore, developing a method for identifying the level of maturity of coffee berries using image processing is an effective solution. The aim that the author wants to achieve is to obtain a comparison of the performance of K-Nearest Neighbor (KNN) and Support Vector Machine (SVM) to obtain better accuracy values in determining the ripeness of coffee cherries. It was found that for the accuracy value of the image of ripe, quite ripe and raw coffee fruit, namely, the accuracy value obtained using KNN was 98.40%, providing better accuracy compared to SVM which had an accuracy value of 86.90%.**

Keywords*— *Classification, Coffee Fruit, Color, KNN, SVM

I. INTRODUCTION

Coffee is one of Indonesia's main export commodities which has high economic value. Coffee fruit ripeness is an important factor in determining the quality and price of coffee.

No	Class	Definition	Figure
1	Raw	Image of raw coffee fruit	
2	Moderately ripe	Image of moderately ripe coffee fruit	
3	Ripe	Image of ripe coffee fruit	

Figure 1. Image of Coffee Fruit

This research aims to develop a system for identifying coffee fruit ripeness levels using image processing with machine learning-based color classification by making a comparison between the K-Nearest Neighbor (K-NN) and Support Vector Machine (SVM) methods. The K-NN and SVM methods are used to compare color classification in coffee fruit images and identify the level of ripeness of the coffee fruit.

Identifying the maturity level of coffee beans using color-based machine learning classification with a comparison of the K-Nearest Neighbor (K-NN) and Support Vector Machine (SVM) methods has its respective strengths and weaknesses.

Identifying the maturity level of coffee beans using color-based machine learning classification offers the advantage of high accuracy. (e.g.[1]) In identifying the maturity level of coffee beans, it can save time and costs as it eliminates the need for human labor, and it can be used in real-time. However, the identification of the maturity level of coffee beans using color-based machine learning classification has drawbacks. It requires a large and representative dataset for training an accurate model, needs adequate hardware for image processing and data analysis, and the model's performance may decline if environmental conditions change.

II. RLATED RESEARCH

Research conducted (eg. [2]) The title of the research is "Recognition of Coffee Beans Based on Color Parameters" using the Backpropagation algorithm and the Support Vector Machine (SVM) Algorithm. Based on the conducted tests, the results from Experiment I produced the best classification model, achieving an accuracy rate of 86%.

Research conducted (e.g. [3]) The research title is "Comparison of Backpropagation and Support Vector Machine Algorithms in Recognizing Types of Corn Seeds." The Support Vector Machine algorithm achieved an average accuracy of 97.1% based on the experiments.

Research conducted (e.g. [4]) With the research title "Soybean leaf disease detection and severity measurement using multiclass SVM and KNN classifier," the K-Nearest Neighbor (KNN) classification is employed for disease classification. The proposed classification system is capable of categorizing various diseases such as blight, brown spot, frogeye leaf spot, and healthy samples with accuracies of 87.3% and 83.6%.

Research conducted (e.g. [5]) With the research title "Blockchain technology-based FinTech banking sector involvement using adaptive neuro-fuzzy-based K-nearest neighbors algorithm," the proposed algorithm is compared with existing approaches to demonstrate its efficiency. The findings indicate that this method achieves an accuracy of 91%.

Research conducted (e.g. [6]) With the research title "Automatic Classification for Fruits' Types and Identification of Rotten Ones using k-NN and SVM," to differentiate between fresh and rotten fruits, Linear and Quadratic Support Vector Machine (SVM) algorithms distinguish between them

based on color segmentation and texture feature values of each fruit image. The accuracy of Linear SVM is 96%, and Quadratic SVM is 98%.

Research conducted (e.g. [7]) The study titled "Parijoto Fruits Classification using K-Nearest Neighbor Based on Gray Level Co-Occurrence Matrix Texture Extraction" explains that the Gray Level Co-Occurrence Matrix (GLCM) method is proposed to extract texture features from Parijoto fruits and then classify them using the K-Nearest Neighbor (KNN) method. GLCM can depict the linear spatial relationships of frequency where the gray values are determined by the gray values in the neighboring area. It can easily employ statistical or histogram-based approaches to image matrix appearance. In this way, information about the relative positions of matching neighboring pixels for the classification process using KNN can be easily obtained. KNN is chosen because it has proven to be applicable to relatively small datasets, but normalization is needed to enhance accuracy. Based on the implementation results of the GLCM and KNN methods for Parijoto fruit classification, the classification accuracy is 80%.

The research entitled "Application of colour, shape, and texture parameters for classifying the defect of Gayo Arabica green coffee bean using computer vision" conducted by (e.g. [8]) The model used in this study employs the K-Nearest Neighbour (K-NN) method and the Support Vector Machine (SVM) method, utilizing 13 parameters such as area, contrast, energy, correlation, homogeneity, roundness, perimeter, and color indices R (red), G (green), B (blue), L*, a*, and b*. A total of 1200 Arabica green coffee beans were captured using a Kinect V2 camera, with 1000 samples for training data and 200 samples for testing data.

The research conducted by (e.g. [9]) The study titled "Analysis K-Nearest Neighbor Method in Classification Of Vegetable Quality Based On Color" reveals that the tests conducted by the author with varying K values in K-Nearest Neighbor (K-NN) such as 3, 4, 5, 6, 7, 8, 9, demonstrate remarkably high accuracy percentages compared to using K-NN alone. The test results indicate that the K-Nearest Neighbor method in classifying data shows excellent accuracy percentages when using random data. The variation percentages in the K-Nearest Neighbor values of 3, 4, 5, 6, 7, 8, 9 show an accuracy percentage of 100%.

The research conducted by (e.g. [10]) The study with the title "Classification of Grape Leaves using KNN and SVM Classifiers" explains that the author proposes a technique to classify grape leaves as healthy or unhealthy. The database consists of 90 manually created images of healthy and unhealthy leaves. The most relevant features for identifying damaged leaves, such as texture and color, are used to train and test the system. K-Nearest Neighbor (KNN) and Support Vector Machine (SVM) classifiers are employed separately to classify grape leaves. KNN classification provides better accuracy compared to SVM classification. The proposed work achieves an accuracy of 96.66% for KNN and 90% for SVM.

The research conducted by (e.g. [10]) The study titled "Fruit Classification System Using Multiclass Support Vector Machine Classifier" aims to classify 18 types of fruits based on feature extraction, reduction, and SVM (Support Vector Machine) classification. The combination of color, GLCM texture features, and shape measurements is utilized for improved results. The experiments demonstrate a significant classification accuracy of 87.06%.

III. Research Methodology

In image processing, there is a methodology applied. The image processing methodology can be seen in Figure 2 below:

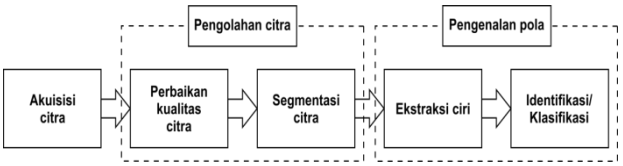


Figure 2. Image Processing Methodology

A. *Image Acquisition*

Digital image acquisition is the process of capturing or scanning analog images to obtain digital images. Devices that can be used for digital image acquisition include digital cameras, webcams, smartphones, scanners, digital microscopes, X-ray machines, MRI machines, CT scanners, or other radiodiagnostic equipment. Factors to consider in the image acquisition process include: the resolution of the acquisition device, distance and angle of image capture, lighting conditions, magnification (zoom), object or camera movement (static or dynamic), and the format of the acquired image.

B. *Image Quality Enhancement*

Image quality enhancement is a preprocessing stage in image processing aimed at improving the quality of the image. The segmentation results serve as indicators of good image quality

C. *Image Segmentation*

Image segmentation is the process of grouping neighboring pixels that have coherence in their properties (such as intensity values). The resulting regions can represent objects or parts of objects and can be further analyzed or modified following image analysis or pattern recognition steps. Image segmentation algorithms are generally based on two properties of intensity values: discontinuity and similarity. For the first property, the approach involves dividing the image based on intensity changes, such as edges in the image. For the second property, the approach involves grouping the image into regions that share similarity based on predefined criteria. Examples of methods based on this property include thresholding, region growing, and region splitting and merging.

D. *Feature Extraction*

Feature extraction is used to recognize objects in an image by identifying parameters that characterize those objects. Features that can be used to differentiate one object from another include shape features, size features, geometric features, texture features, and color features. Each object extracts its features based on specific parameters and is grouped into a particular class.

E. Identification/Classification

In the identification stage, typically two main processes are carried out, namely the training process and the testing process. The training process is conducted using a set of training data containing feature parameters used to distinguish between one object and another. The training process maps the training data to the training target through a formulation (identification/classification algorithm). The next process is the testing process, where the formulation generated from the training process is used to map the test data, resulting in output data that is then compared with the test target to obtain the accuracy level of the testing process.

F. Hue Saturation Value (HSV) Image Model

HSV describes the values of hue, saturation, and value in an image. The color values in the segmentation process are depicted with hue and are used to differentiate one color from another, determining attributes such as redness, greenness, and others from the light. The wavelength of light is also united with hue. The color intensity of an image is represented by saturation.

(K-NN) K-Nearest Neighbor (K-NN) Algorithm

K-Nearest Neighbor is a classification algorithm for objects based on learning data that are closest or have the most similar object characteristics. There are also several steps to calculate the K-NN algorithm, namely:

1. Determining the value of K.
2. Calculating the Euclidean distance (query instance) to each object in the training data.
3. Grouping objects into clusters based on the smallest Euclidean distance.
4. Collecting class labels y (nearest neighbor classification)
5. The distance, whether an object is far or near, can be calculated using the Euclidean distance, where two vector distances of size n, for example, $X=(X_1,X_2,X_3,\ldots,X_n)$ and $Y=(Y_1,Y_2,Y_3,\ldots.Y_n)$ the equation is obtained as follows:

$$\text{Dist}(X,Y)=\sqrt{\sum_{i=1}^{n}(xi-yi)^2}$$

where,

d = distance between x and y
x = cluster center data
y = data on the attribute
i = each data
n = number of data,
xi = data on the cluster center for i
yi = data on each data for i

From the above equation, as depicted in Figure 3

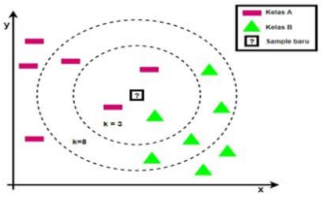


Figure 3. K-Nearest Neighbor Algorithm Method

Support Vector Machine (SVM) Algorithm

Support Vector Machine (SVM) is a discriminative classifier that produces a separating hyperplane. Error tolerance is included to make the separating hyperplane robust in case of class data that cannot be completely separated—a machine learning method that operates based on the principles of Structural Risk Minimization, aiming to find the best hyperplane that separates two classes in the input space, as illustrated in Figure 4.

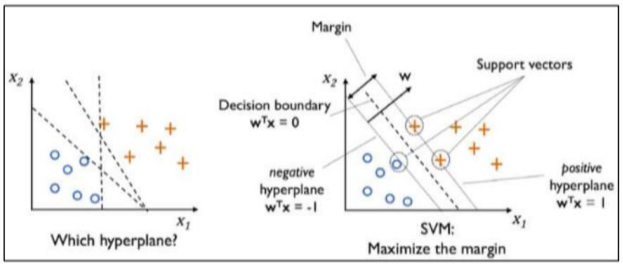


Figure 4. SVM Algorithm Method

The linear SVM classification hyperplane can be seen in the following equation (1):

$$F(x) = w^T x + b$$

According to Vanpik and Cortes (1995), the equations for Support Vector Machine can be seen in the following equations (2) and (3):

$$[(w,x_i) + b)] \geq 1 \text{ untuk } yi = +1$$
$$[(w^T . x_i + b)] \leq = 1 \text{ untuk } yi = -1$$

Where:
W: Weight Vector
b: Bias
yi: Class label

IV. Hasil Results And Discussion

IV.1. Results

The results of the simulation of two prediction models, data that has undergone preprocessing is then tested to obtain the best prediction model. The prediction model has been tested and evaluated using a set of test data in the Orange application, where 1 attribute is the target. The simulation results are obtained as shown in Figure 5.

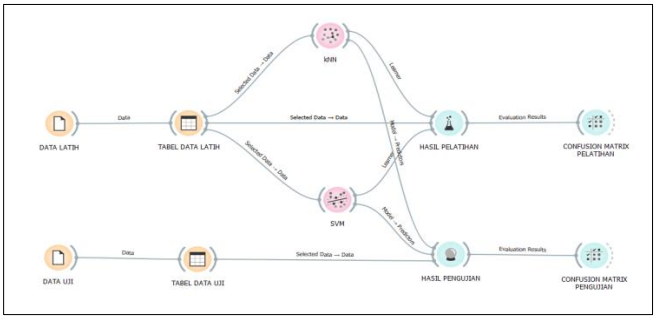


Figure 5. Design of Coffee Fruit Ripeness Prediction Model

After the data preprocessing, the image data is then processed into the prediction model in the Orange software using the K-NN and SVM methods.

IV.2. Discussion

The training data used consists of 249 image data with 7 features, as seen in Figure 6 below:

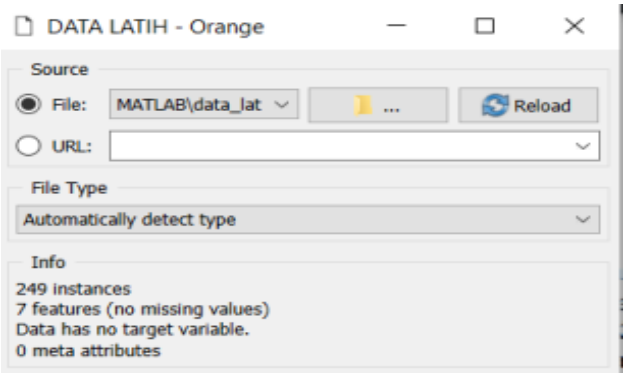
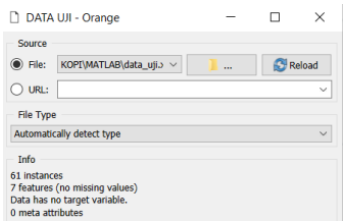


Figure 6. Training Data

The test data used consists of 61 image data, as seen in Figure 7:



Uji Figure 7. Test Data

When using the K-Nearest Neighbors (KNN) algorithm, the main parameter to consider is the 'k' parameter (the number of nearest neighbors to be considered). Additionally, there are several other parameters that can influence the performance of the KNN model. Here are the main parameters of KNN.

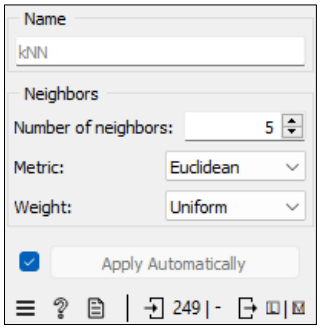


Figure 8. K-Nearest Neighbors (K-NN)

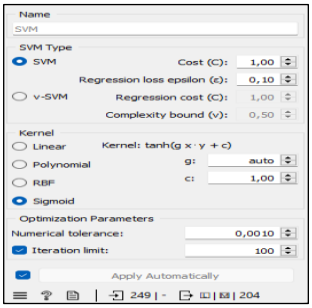


Figure 9. Support Vector Machine (SVM)

The results of the calculations for both methods can be seen in Figure 10. As follows:

Figure 10. Test and Score of K-Nearest Neighbors (K-NN)

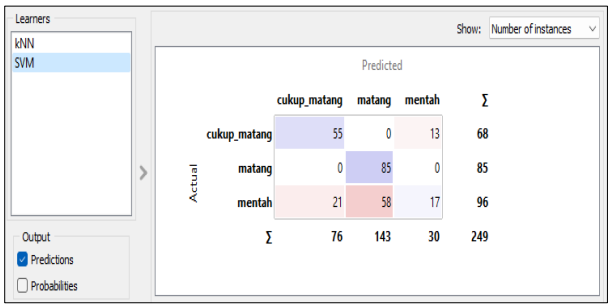


Figure 11. Test and Score of Support Vector Machine (SVM)

V. CONCLUSIONS

Based on KNN training compared with SVM, Precision, Recall, F1 Score from the Support Vector Machine Model Accuracy is worth = 63.10%, Precision is worth = 61.90%, Recall is worth = 63.10%, F1 Score is worth = 56.70%, compared to KNN Precision is worth = 98.89%, Recall value = 98.90 %, F1 Score value = 98.89% becomes value = 98.89% after comparing the results of KNN and SVM training for both models, it can be concluded that KNN has a better accuracy value than the accuracy value of SVM.

TABLE I
COMPARION OF TRAINING RESULTS

Model	CA	F1	Prec	Recall
KNN	0.940	0.980	0.980	0.980
SVM	0.931	0.567	0.619	0.631

Based on KNN training compared with SVM, Precision, Recall, F1 Score from the Support Vector Machine Model Accuracy is worth = 63.10%, Precision is worth = 61.90%, Recall is worth = 63.10%, F1 Score is worth = 56.70%, comparison with KNN Precision is worth = 98.89%, Recall value = 98.90 %, F1 Score value = 98.89% becomes value = 98.89% after comparing the results of KNN and SVM training for both models, it can be concluded that KNN has a better accuracy value than the accuracy value of SVM.

TABLE II
COMPARISON OF TEST RESULTS

Model	CA	F1	Prec	Recall
KNN	0.869	0.867	0.892	0.869
SVM	0.984	0.984	0.984	0.984

Based on the results of KNN testing compared with SVM, Precision, Recall, F1 Score from the Support Vector Machine Model, Accuracy value = 86.90%, Precision value = 89.20%, Recall value = 86.90%, F1 Score value = 86.70%, comparison of test results with KNN Precision value = 98.40%, Recall value = 98.40%, F1 Score value = 98.40% to value = 98.40% after comparing the KNN and SVM test results for the two models, it was concluded that KNN has a better accuracy value than the accuracy value from SVM.

REFERENCES

[1] S. P. Adenugraha, V. Arinal, and D. I. Mulyana, "Klasifikasi Kematangan Buah Pisang Ambon Menggunakan Metode KNN dan PCA Berdasarkan Citra RGB dan HSV," *J. Media Inform. Budidarma*, vol. 6, no. 1, p. 9, 2022, doi: 10.30865/mib.v6i1.3287.

[2] H. A. Sihombing and I. C. Buulolo, "Pengenalan Buah Kopi Berdasarkan Parameter Warna Menggunakan Algoritma Backpropagation Dan Algoritma Support Vector Machine (Svm)," *Seminastika*, vol. 3, no. 1, pp. 26–32, 2021, doi: 10.47002/seminastika.v3i1.234.

[3] M. Rizky and H. Irsyad, "539-Article Text-1311-1-10-20201010," vol. 1, no. 1, pp. 111–120, 2020.

[4] S. B. Jadhav, V. R. Udupi, and S. B. Patil, "Soybean leaf disease detection and severity measurement using multiclass SVM and KNN classifier," *Int. J. Electr. Comput. Eng.*, vol. 9, no. 5, pp. 4077–4091, 2019, doi: 10.11591/ijece.v9i5.pp4077-4091.

[5] H. Rjoub, T. S. Adebayo, and D. Kirikkaleli, "Blockchain technology-based FinTech banking sector involvement using adaptive neuro-fuzzy-based K-nearest neighbors algorithm," *Financ. Innov.*, vol. 9, no. 1, 2023, doi: 10.1186/s40854-023-00469-3.

[6] A. Nosseir and S. E. Ashraf Ahmed, "Automatic classification for fruits' types and identification of rotten ones using k-NN and SVM," *Int. J. online Biomed. Eng.*, vol. 15, no. 3, pp. 47–61, 2019, doi: 10.3991/ijoe.v15i03.9832.

[7] I. U. W. Mulyono *et al.*, "Parijoto Fruits Classification using K-Nearest Neighbor Based on Gray Level Co-Occurrence Matrix Texture Extraction," *J. Phys. Conf. Ser.*, vol. 1501, no. 1, 2020, doi: 10.1088/1742-6596/1501/1/012017.

[8] A. Maghfirah and I. S. Nasution, "Application of colour, shape, and texture parameters for classifying the defect of Gayo Arabica green coffee bean using computer vision," *IOP Conf. Ser. Earth Environ. Sci.*, vol. 951, no. 1, 2022, doi: 10.1088/1755-1315/951/1/012097.

[9] P. H. Putra, M. S. Novelan, and M. Rizki, "Analysis K-Nearest Neighbor Method in Classification of Vegetable Quality Based on Color," *J. Appl. Eng. Technol. Sci.*, vol. 3, no. 2, pp. 126–132, 2022, doi: 10.37385/jaets.v3i2.763.

[10] A. A. Bharate and M. S. Shirdhonkar, "Classification of Grape Leaves using KNN and SVM Classifiers," *Proc. 4th Int. Conf. Comput. Methodol. Commun. ICCMC 2020*, no. Iccmc, pp. 745–749, 2020, doi: 10.1109/ICCMC48092.2020.ICCMC-000139.