

Improving the accuracy of green bean palm civet coffee purity classification using wrapper feature selection

Shinta Widyaningtyas^{1*}, Muhammad Arwani², Sucipto³, Yusuf Hendrawan⁴

¹Agriculture Engineering, Politeknik Negeri Jember, Jember, East Java, Indonesia

²Agroindustrial Technology, Universitas Nahdlatul Ulama Indonesia, Central Jakarta, Jakarta, Indonesia

³Agroindustrial Technology, Universitas Brawijaya, Malang, East Java, Indonesia

⁴Biosystem Engineering, Universitas Brawijaya, Malang, East Java, Indonesia

*Corresponding author: shinta_widya@polije.ac.id

Editor: Samuel Pereira de Carvalho

Received in September 11, 2024 and approved in January 14, 2025

ABSTRACT

Palm civet coffee, a highly prized specialty beverage, is increasingly targeted by adulteration due to its limited production and substantial market value. The absence of reliable detection methods necessitates the development of efficient, non-destructive sensing techniques. The aims of this study is a development of machine vision in selecting relevant color and texture features in detecting the purity of civet coffee. The development of computer vision is expected to be more applicable in the creation of Palm Civet Coffee purity detection tools. A comprehensive dataset comprising 528 coffee bean images was extracted and generated 101 image features (11 color and 90 textural). Utilizing a wrapper method, three classifiers k-Nearest Neighbors (kNN), Random Forest (RF), and Support Vector Machine (SVM) were coupled with four optimization algorithms (Bat Algorithm, Cuckoo Search, Genetic Algorithm, and Grey Wolf Optimizer) to identify the most informative features. Comparative performance analysis revealed that the Random Forest classifier, optimized with the Grey Wolf Optimizer achieved the highest accuracy of 0.981 using five features i.e Blue_Mean, Hue_Entropy, Gray_Inverse, S_HSL_Correlation, and Green_Cluster. On the other hand, kNN and SVM classifiers attained maximum accuracies of 0.943 and 0.925, respectively. Additionally, statistical analysis using One-Way ANOVA demonstrated significant performance improvements of the accuracy for each classifier (P-value < 0.01). These findings demonstrate the efficacy of the wrapper method in feature selection for classifying purity of Palm Civet Coffee. This research is one of the stages in the development of automatic civet coffee purity detection to prevent adulteration so that authenticity is guaranteed.

Key words: Classifier; feature selection; learning algorithm; machine vision; palm civet coffee purity

1 INTRODUCTION

Palm Civet Coffee is the most expensive and rarest coffee in the world because of the uniqueness of its production process (Marcone, 2004, Jumhawan et al., 2013, 2015). Palm Civet Coffee consists of coffee beans (either arabica or robusta) that are eaten and then passed through the digestive system of the civet (*Paradoxurus hermaphroditus*). The civet has a keen sense for selecting the best and ripest coffee beans (Suhandy; Yulia, 2017). This process gives Palm Civet Coffee a more exotic taste than regular coffee. In Indonesia, Palm Civet Coffee popularly known as Kopi Luwak, with a selling price of around \$200-250/lbs approximately a hundred times higher than regular coffee, market size at \$6.5 billion in 2021 and is projected to reach \$10 billion in 2030 (Jumhawan et al., 2016). Until a few years ago, the production of Palm Civet Coffee in Indonesia was still limited, making it a prime target for adulteration.

The scarcity of authentic Palm Civet Coffee has led to widespread adulteration, raising doubts about the purity of the product. This problem is compounded by the lack of automated tools to detect such adulteration. Traditional methods of verifying coffee purity are often time-consuming, require sample preparation, are expensive, and necessitate chemical analysis requires experts. This situation presents an opportunity to design simple, fast, accurate, non-destructive,

and real-time equipment to determine the purity of Palm Civet Coffee. Designing tools to determine the purity of Palm Civet Coffee can involve sensors such as machine vision, which has proven effective for analysis in various fields and applications. Machine vision can extract color and texture features, though not all features are relevant to determining the purity of Palm Civet Coffee. One of the main challenges in analyzing big data is detecting relevant variables and eliminating noise. No study has exclusively reported on selecting the best combination of color features and texture features for improving the accuracy of classifying purity of Palm Civet Coffee (Hendrawan; Murase, 2011a). Feature selection, a crucial stage in data preprocessing, helps reduce dimensionality by eliminating redundant and irrelevant data, thereby increasing accuracy and preventing overfitting (Chandrashekar; Sahin, 2014; Mafarja; Mirjalili, 2018; Tsai; Chen, 2019; Zhang; Gao, 2020; Aksakalli et al., 2021; Afshar; Usefi, 2022). According to (Yu; Liu, 2004) feature selection includes four types: filter methods, wrapper methods, hybrid methods, and embedded techniques. Based on (Biswas; Bordoloi; Purkayastha, 2016) the major drawback of the filter method is its lack of interaction with the classification model during feature selection, while embedded techniques cannot identify a small set of features. According to (Tiwari; Chaturvedi, 2022) the hybrid method often becomes stuck in

a local optimal solution. Therefore, the wrapper method has become one of the preferred solutions for feature selection.

The wrapper method comprehensively searches for the best combination of feature subsets. It is conducted using a classification method optimized with a learning algorithm to identify the best features. Classification methods such as k-Nearest Neighbor Classifier (kNN Classifier) (Maleki; Zeinali; Niaki, 2021), Random Forest Classifier and Support Vector Machine can model the non-linear relationship between image features (color and texture) and the purity of Palm Civet Coffee. In this study, the chosen learning algorithms are known for their efficacy in dealing with non-linear data relationships such as Bat Algorithm (Mallidi, 2021, Alyasseri et al., 2022), Cuckoo Search Algorithm (Mousavirad; Ebrahimpour-Komleh, 2014, Kelidari; Hamidzadeh, 2021), Genetic Algorithm (Amini; Hu, 2021) and Grey Wolf Optimizer (Shang et al., 2025) which are used to optimize the model. These algorithms were selected based on their demonstrated performance in previous studies related to feature selection and their compatibility with the classifiers employed. By integrating

these algorithms with kNN, RF, and SVM classifiers, aim to harness their optimization strengths to enhance feature subset quality, thereby improving classification accuracy and model robustness. The evaluation criteria of this study are accuracy and the number of selected features (Mafarja et al., 2023).

2 MATERIAL AND METHODS

2.1 Materials

This study utilized green bean Arabica Palm Civet Coffee and regular Arabica coffee, both sourced from PT. Perkebunan Nusantara XII, Banyuwangi, Indonesia. The Arabica Palm Civet Coffee employed is of the Longan variety, while the regular Arabica coffee was processed using the wet processing method (Hendrawan; Widyaningtyas; Sucipto, 2019). For data collection, each mixture comprised 160 green beans, categorized into four purity levels: very low (0-25%), low (25-50%), medium (50-75%), and high (75-100%) (Hendrawan et al., 2022) as illustrated in Figure 1. The feature

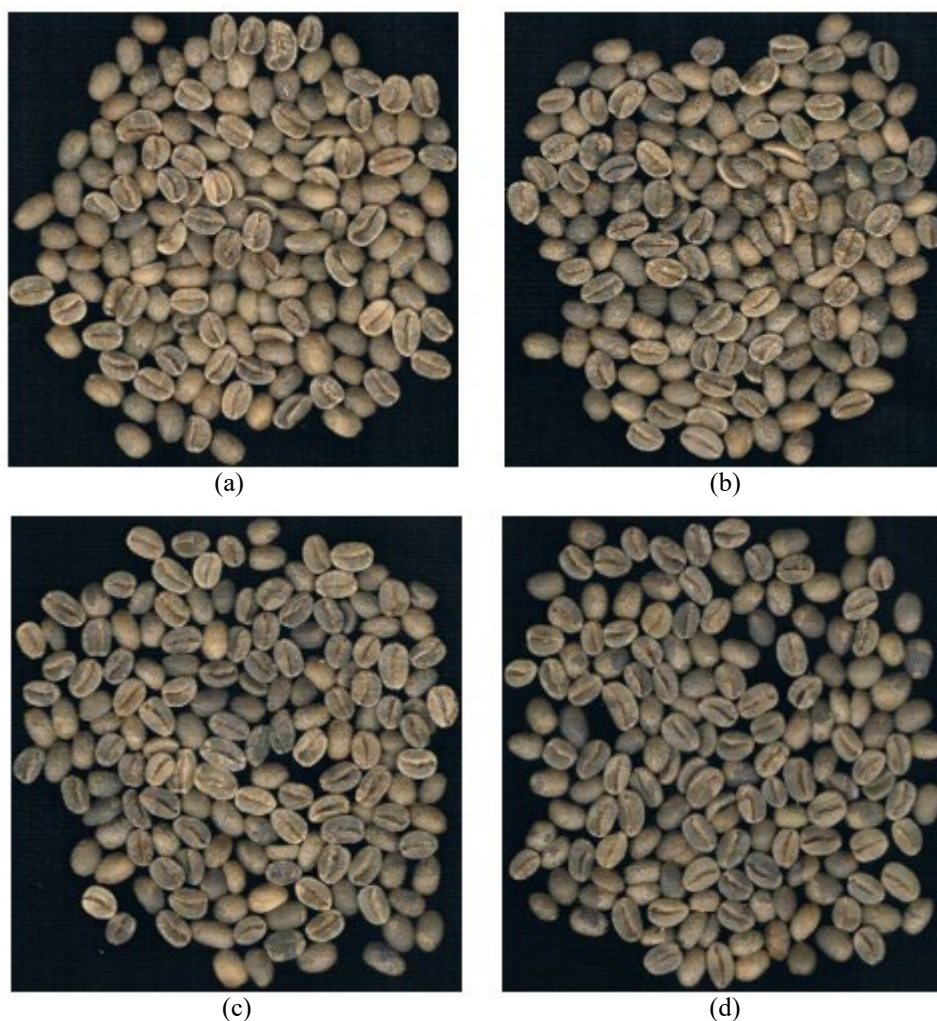


Figure 1: Mixture Purity of Palm Civet Coffee (a) very low (0-25%); (b) low (25-50%); (c) medium (50-75%); (d) high (75-100%).

extraction tool utilizes a self-built Visual Basic 6.0 program based on color and textural analysis software, while feature selection is performed using Python 3.11.5 and Visual Studio Code. The experiments are conducted on a computer system with the following specifications: Intel Core i7-1165G7 2.80 GHz, 8 GB RAM.

2.2 Methods

This study aims to classify the purity of Palm Civet Coffee through a process divided into five steps: image acquisition, feature extraction, model design (including model classification and learning algorithms), and evaluation criteria. The input models are co-occurrence matrices from feature extraction, which include color features, texture features, and combined features (both color and texture). The detailed steps for classifying the purity of Palm Civet Coffee are as follows:

2.2.1 Image Acquisition

Images of the coffee beans were captured using a Nikon Coolpix A10 digital camera (16 megapixels, Japan). The camera was positioned 20 cm above a tray with an area of 256 cm², ensuring consistent lighting conditions via constant fluorescent to minimize variability in color and texture feature extraction. Each image was taken at a resolution of 1343 x 1349 pixels and saved in bitmap (.bmp) format. Pre-processing involved cropping the images to focus solely on the coffee beans, resulting in a total of 528 image datasets for analysis. While this sample size is sufficient to demonstrate the effectiveness of the proposed method, a greater variety of civet coffee samples are needed to improve the generalizability of the model.

2.2.2 Feature Extraction

Feature extraction aims to extract the unique values from image data that distinguish it from other objects. In this study, feature extraction is divided into two parts: color features and textural features. The result of this stage is a co-occurrence matrix consisting of 11 color features and 90 textural features, stored in Microsoft Excel. In machine vision, image data contain three primary colors: red, green, and blue (RGB). From the RGB color model, the image is converted to gray, HSL, HSV, Lab, and other color spaces using equations referenced in (Hendrawan; Murase, 2010). Textural features represent the variations at each level, measuring characteristics such as smoothness, coarseness, and regularity of each surface in different directional orders. Textural features are extracted in the type of color based on Haralick's formula are as follows Equation 1-10 (Hendrawan; Murase, 2010; Hendrawan; Widyaningtyas; Sucipto, 2019):

$$Energy = \sum_i^M \sum_j^N P^2[i, j] \quad (1)$$

$$Entropy = -\sum_i^M \sum_j^N P^2[i, j] \quad (2)$$

$$Contrast = \sum_i^M \sum_j^N (1-j)^2 P[i, j] \quad (3)$$

$$Homogeneity = \sum_i^M \sum_j^N \frac{P[i, j]}{1+|i-j|} \quad (4)$$

$$Inverse \text{ Difference Moment} = \sum_i^M \sum_j^N \frac{P[i, j]}{|i-j|^k} \quad i \neq j \quad (5)$$

$$Correlation = \sum_i^M \sum_j^N \frac{(i-\mu)(j-\mu)P[i, j]}{\sigma^2} \quad (6)$$

$$Sum \text{ Mean} = \frac{1}{2} \sum_i^M \sum_j^N (iP[i, j] + jP[i, j]) \quad (7)$$

$$Variance = \frac{1}{2} \sum_i^M \sum_j^N ((i-\mu)^2 P[i, j] + (j-\mu)^2 P[i, j]) \quad (8)$$

$$Cluster \text{ Tendency} = \sum_i^M \sum_j^N (i+j-2\mu)^k P[i, j] \quad (9)$$

$$Maximum \text{ Probability} = \max_{i,j}^{M,N} P[i, j] \quad (10)$$

where $P[i, j]$ is the element of the normalized co-occurrence matrix, μ is the mean of pixel element and σ is the standard deviation of pixel element. The feature extraction results are divided into 70% training data and 30% testing data.

2.2.3 Model Design

2.2.3.1 Model Classification

Three classifiers were employed to model the relationship between color and texture features and coffee purity levels.

a. k-Nearest Neighbors (kNN)

kNN is a well-known supervised modeling technique for classification in data mining (Saadatfar et al., 2020). In kNN, the k-value represents the number of nearest neighbors used as a basis for determining cluster decisions. In this study using k-values of 1, 3, 5, 7, and 9 based on (Uddin et al., 2022). The k-value in kNN is an odd number, chosen empirically; it should not be too large or too small. A maximum of 100 iterations is used to search for the optimal k-value.

b. Random Forest (RF)

Random Forest (RF) is a classification modeling technique developed by Breiman in 2001 (Fawagreh; Gaber; Elyan, 2014). The main advantage of RF for image classification is its robustness and efficiency (Feng; Liu; Gong, 2015). RF involves constructing a collection of

weakly correlated classification or regression trees and then aggregating them (Iranzad; Liu, 2024). RF models were built using scikit-learn (version 1.5.1). One of the parameters for searching the optimal value in Random Forest is the number of trees (ntree). ntree varied (100, 200, 500, 1000) to identify the optimal value to improve the accuracy of classification.

c. Support Vector Machine (SVM)

SVM is the most used supervised machine learning technique for the classification of linear and non-linear data (Cosma et al., 2017, Kashaf, 2021). The basic concept of this method is to seek a classification hyperplane that can divide the training dataset into a discrete predefined number of classes as the decision surface. The advantage of SVM is good generalization performance (Lee et al., 2010) so become one of the most used classification method (Cervantes et al., 2020). In this study, optimization of SVM using penalty coefficient (c) and gamma. c is effective to reduce the training error and realize model simplification. g describes the nonlinear mapping from input space to certain high-dimensional space (Ren et al., 2021). This study used sensitivity analysis to search the best pair of c and gamma (2^{-10} , 2^{-5} , 2^0 , 2^5 , 2^{10}).

2.2.3.2 Learning Algorithm

Four optimization algorithms were integrated with the classifiers to enhance feature selection and model performance, as detailed in the learning algorithms section.

a. Bat Algorithm (BA)

BA is well known bio-inspired algorithm for multi-objective optimization developed by Dr. Xin-She Yang in 2010 (Kar, 2016, Rostami et al., 2021). Inspired by the echolocation behavior of bats to find the exact path and location of their prey by sending sound waves and receiving reflection (Rostami et al., 2021). Key parameters included a population size of 20 bats, a maximum of 100 iterations, initial values for loudness and pulse rate set to 0.5, and frequency range [0, 2]. The step of BA are follows:

1. Initialization of bats i.e an array with zeros of size to saved frequency for each bat; velocity matrix with zeros of size ; solution matrix with zeros of size to saved update position; Solution matrix with random binary values (0,1) as the feature is selected by each bat; and to saved the fitness of each bat.
2. Initial fitness calculation.
3. Saved the best initial solution.
4. Main loop for generation i.e update each frequency, velocity and position of the bat; perform the local search; evaluate new solution; accept the new solution; and update global best.
5. Return the best solution.

b. Cuckoo Search Algorithm (CSA)

The CSA is a nature-based meta heuristic algorithm which has flexibility and effectiveness in handling optimization

problems introduced by Yang in and Deb in 2009 (Reda et al., 2024). The obligate brood parasite behaviour of cuckoos, who deposit their eggs in the nests of other smaller birds like starlings, is the source of inspiration for this algorithm. The starling cares for the cuckoo chick as if it were its own when the egg hatches. The cuckoo chick starts to outgrow and pushes starlings of the nest completely as it gets bigger (Alzaqebah et al., 2021). The algorithm employed a population of 20 nests, with 100 iterations and a parasitism probability of 0.25, facilitating efficient exploration of the feature space. The steps of CSA are as follows :

1. Initialization of nests and fitness. Initialization the solution matrix 'nests' with random binary values (0 or 1) that represent a features selected or not. Initialization the fitness array 'fitness' to save the fitness values of each nest.
2. Identify the best initial nest. This step divided into two parts i.e. identify and saved the best solution (nest) and saved the fitness value of the best solution.
3. Main loop for generation i.e generate new solution and update nests; evaluate and update global best nest; and parasitism and abandonment.
4. Return the best solution.

c. Genetic Algorithm (GA)

The GA was introduced by Holland in 1992 is a bio inspired algorithm method which has been widely used to solve complex optimization (Goel, 2020). GA is a stochastic approach based on natural genetics and biological evolution for feature optimization (Adhao; Pachghare, 2020). Emulating natural selection, GA evolves feature subsets through processes of selection, crossover, and mutation. Parameters included a population size of 50 individuals, 40 generations, and a mutation probability of 0.1. GA effectively navigates the feature space to identify high-performing feature combinations. The steps of GA are as follows:

1. Population initialization: initialize the population of individuals. Each individual represented by binary vector (0 and 1). 1 indicates that a feature is selected, and 0 indicates that the feature not selected.
2. Individual evaluation: evaluate the fitness of a single individual.
3. Population evaluation: calculate the fitness scores for the entire population.
4. Selection of mating pool: for selecting individual for reproduction based on their fitness.
5. Crossover: generates two offspring from two parent individual through a crossover operation.
6. Mutation: mutate an individual by flipping bits in the binary vector
7. Main genetic algorithm loop: This is the core function that runs the GA. This step including three processes i.e population initialization, generating loop and returning best individual.

8. Final output: The algorithm returns the best individual, representing the most optimal subset of features over all generations.

d. Grey Wolf Optimizer (GWO)

The GWO is a well know Swarm Intelligence method proposed by Mirjalili in 2004. Mimicking the social hierarchy and hunting behavior of grey wolves, GWO optimizes feature subsets by simulating leader-follower dynamics among a population of wolves. The primary benefits of the GWO method are its low computational cost, lack of need for any particular input parameters, and need for less memory than other SI-based algorithms (Wang et al., 2022). The optimizer utilized a population size of 20 wolves and ran for 50 iterations, promoting collaborative exploration of the feature landscape. The steps of GWO are as follows :

1. Population initialization: Initialize the population (pack of wolf) by creating 'pop_size' binary vectors. Each vector represents a possible solution where 1 indicates a selected feature and 0 indicates non selected feature.
2. Fitness evaluation.
3. Leader wolf initialization: Initialize three leader wolf (alpha, beta, delta) as the best, second, and third solution.
4. The main optimization loop consist of three steps i.e. population evaluation, parameter 'a' calculation that controls the exploration and exploitation balance in search space and space update.
5. Return the best solution.

2.2.4 Evaluation Criteria

Evaluation criteria that used in this study is accuracy from the confusion matrix. The confusion matrix can be seen in Figure 2. As the most intuitive performance measure, accuracy is defined as the ratio of the number of correctly classified objects to the total number of objects evaluated (Ali; Neagu; Trundle, 2019). Accuracy is calculated using Equation 11:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (11)$$

where:

1. True Positive (TP) – prediction and actual both are yes
2. True Negatives (TN) – prediction is no and actual is yes
3. False Positives (FP) – prediction is yes and actual is no
4. False Negatives (FN) – prediction is no and actual is No.

The best performance model is selected based on accuracy and the number of selected features. High accuracy indicates the model is getting better because the ratio of the number of correctly classified objects to the total number of objects evaluated is high (Ali; Neagu; Trundle, 2019). In addition, the number of selected features became evaluation criteria because fewer selected features will make it easier

to implement an automatic tool or machine for detecting the purity of Palm Civet Coffee. One-Way ANOVA was conducted to determine the statistical significance of differences in accuracy across feature sets and classifiers.

	Predicted : No	Predicted : Yes
Actual : No	TN	FP
Actual : Yes	FN	TP

Figure 2: Confusion Matrix.

3 RESULTS

Several procedures were performed to determine the selected features using various learning algorithms and classification models. The 528 data points, comprising 101 color and texture features, were categorized into several levels of Palm Civet Coffee purity, including very low, low, medium, and high. The analysis of normalized color features revealed that almost all color features exhibit the same trend at each level. Similarly, the analysis of normalized texture features showed several features with the same data patterns, including Entropy and Contrast; Homogeneity and Inverse Difference; Correlation, Sum_Mean, Variance, and Cluster Tendency. The best features were selected through sensitivity analysis on each color feature, texture feature, and all features; learning algorithm; and classification model. Model performance was evaluated using a confusion matrix by comparing the original class with the predicted class.

4 DISCUSSION

4.1 Purity Level and Image Features

Figure 3 shows the normalized co-occurrence matrix of color features due to the purity level. The result indicate that most all color features (except R_Mean, V_Mean, G_Mean, B_Mean, H_Mean) follow the same pattern due to the purity level; they increase from very low to low, decrease from low to medium and slight increase from medium to high. Figure 4 (a) – (j) shows the normalized co-occurrence matrix of textural features due to the purity level. As seen in Figure 4(a), energy measures the number of repeated pixel pairs. If the frequency of repeated pixel pairs is high, then a high energy level will occur (Hendrawan; Murase, 2011a). The results show all energy co-occurrence matrixs have the same pattern due to the purity level, sharply decreasing from very low to low, sharply increasing from low to medium and decreasing from medium to high. Entropy measures the randomness of a gray-level distribution. The entropy is high if gray level are randomly distribute throughout the image. Figure 4(b) show

most of the entropy textural features (except Hue_Entropy and Value_Entropy) increase from very low to low and decrease gradually from low to high. This result indicate entropy increase in low purity level and decrease as the purity of Palm Civet Coffee increase. Figure 4(c) show most of the co-occurrence matrixs (except Hue_Contrast) increase from very low to low and sharply decrease from low to high. For Hue_Contrast different because from medium to high it slightly decreases to high purity level. Based on (Zhao; Forte; Pipan, 2016), contrast measures the amount of local variation present in an image. If each pair of pixels has a similar gray level, then there should be less contrast. Homogeneity measures the local homogeneity. If each pair of pixels has a similar gray level, then there should be large homogeneity. Figure 4(d) show that all inverse difference moment textural features have the same pattern, decreasing from very low to low and increasing sharply from low to high purity. Inverse difference moment has slightly same with homogeneity pattern. Figure 4(e) shows all inverse difference moment textural features have the same pattern that decreases from very low to low and increases sharply from low to high purity. Inverse difference moment measures the smoothness of the image. If each pair of pixels has a similar gray level, then there should be a high inverse difference moment. Figure 4(f) shows that most of the co-occurrence matrix sum mean textural features (except Green_Correlation, Blue_Correlation, Hue_Correlation, and Gray_Correlation) have the same pattern: from very low to low purity, correlation decrease and increase from low purity to high purity. Correlation measures a correlation between two pixels in the pixel pairs. The correlation is expected to be high if the gray-levels of the pixel pairs are highly correlated. Figure 4(g) show most all co-occurrence matrix sum mean textural features (except Red_SumMean, Hue_SumMean, and S_HSL_SumMean) has same pattern, from very low to low sum mean value slightly increase, decrease from low to medium and increase from medium to high purity. Sum mean value measures the mean of gray levels in the image. The sum mean is expected to be large if the sum of the gray-levels of the image is high. Figure 4(h) shows that most of the co-occurrence matrix variance textural features (except Red_Variance, Hue_Variance, S_HSL_Variance) has the same pattern: from very low to low variance value increase, decrease from low to medium and increase slightly from medium to high purity. Variance characterizes the dispersion degree of gray-level. The variance is expected to be large if the gray-levels of the image are spread out greatly. Figure 4(i) show that most of the co-occurrence matrix cluster tendency textural features (except Blue_Cluster, Red_Cluster, S_HSL_Cluster, Hus_Cluster) have the same pattern: from very low to low cluster value increase, decrease from low to high purity. Cluster Tendency measures the grouping of the pixels that have similar gray-level values. Figure 4(j) shows most all co-occurrence matrix maximum probability (except Hue_Maximum Probability) has

the same pattern, from very low to low maximum probability increases and sharply decreases from low to high purity. Maximum Probability is the result in the pixel pair that is most predominant in the image. The maximum probability is expected to be high if the occurrence of the most predominant pixel pairs is high (Hendrawan; Murase, 2011b).

4.2 Purity Level Classification using Wrapper Method

In this study, the input was separated into color, texture and all features to test the diversity of the input and its accuracy. This is related to the implementation of the tool, so the simplicity of the input is taken into consideration. In general, the extraction of color features is simpler than that of texture features. This will also affect the speed and efficiency of the tool in detecting the purity of civet coffee to prevent adulteration. Table 1 shows sensitivity analysis for color, textural and all features (color and textural features) for kNN classifier. This study using the parameter of kNN besides k-value are kFold validation (kFold) = 5 and maximum number of iterations (T) = 100. 5 kFold is cross validation method that divided dataset into five equal sections or folds for training and testing. Based on (Sejuti; Islam, 2023) 5-fold cross-validation ensures generalization of the dataset and avoids overfitting. Iteration shows the repetition of learning algorithms and classification models to learn data patterns. Iteration will stop when the goal is achieved or reaches the maximum iteration. The selected model for classify purity of Palm Civet Coffee using kNN is textural features with k-value = 5, genetic algorithm, 4 selected features (Gray_Homogeneity, Hue_Correlation, S_HSV_Correlation, Light_Correlation), and accuracy at 0.943. According to (Mafarja et al., 2023), the feature selection approach is to minimize the number of features. For this reason, selecting the best model is also based on the number of selected features besides accuracy. Based on 5 selected models have the same accuracy at 0.943. The fewer features selected will simplify the design of the tool. Furthermore, the combination of genetic algorithm and kNN with texture features can classify the purity of Palm Civet Coffee. Table 1 also show that using only color features is less accurate in classifying the purity of Palm Civet Coffee compared to texture features and all features. This indicates that the combination of kNN and the learning algorithm is not yet able to classify the purity of Palm Civet Coffee based on color features. Table 2, using One-Way ANOVA, shows that there is a significant accuracy difference among color, texture, and all features at a P-value of 0.05 using the kNN classifier.

This study employs random forest with the number of tree (ntree) parameters 100, 200, 500, and 1000 to search optimal value for classification. Based on (Fang; Yu; Zeng, 2022) ntree is one of the random forest parameter that refers to the number of decision trees to be built in a random forest model. Table 3

shows sensitivity analysis for color, texture and all features. The selected model for random forest classifier is all features data using grey wolf optimizer learning algorithm with $n_{tree} = 500$, 5 selected features (Blue_Mean, Hue_Entropy, Gray_Inverse, S_HSL_Correlation, Green_Cluster), and accuracy at 0.981. The

results shows, random forest classifier requires a combination of color and texture features to recognize differences purity of Palm Civet Coffee. Table 4 shows using One Way ANOVA, there is a significant accuracy difference from color, texture and all features at $P\text{-value} = 0.05$ using random forest classifiers.

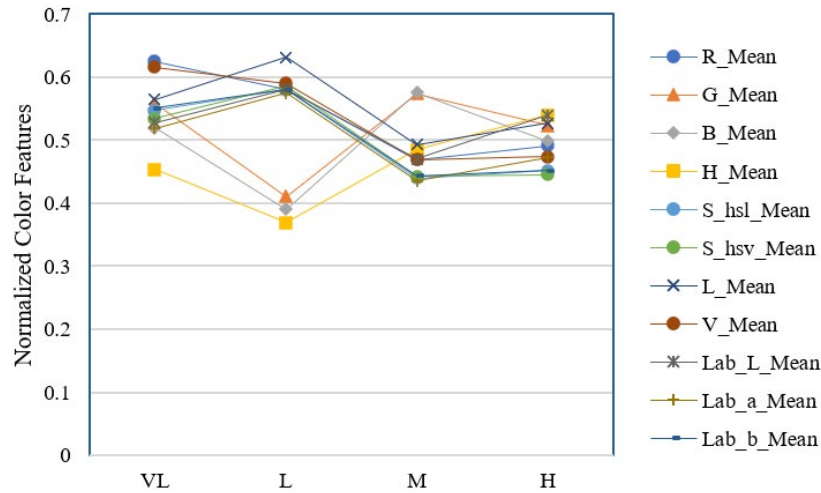


Figure 3: Normalized Color Features.

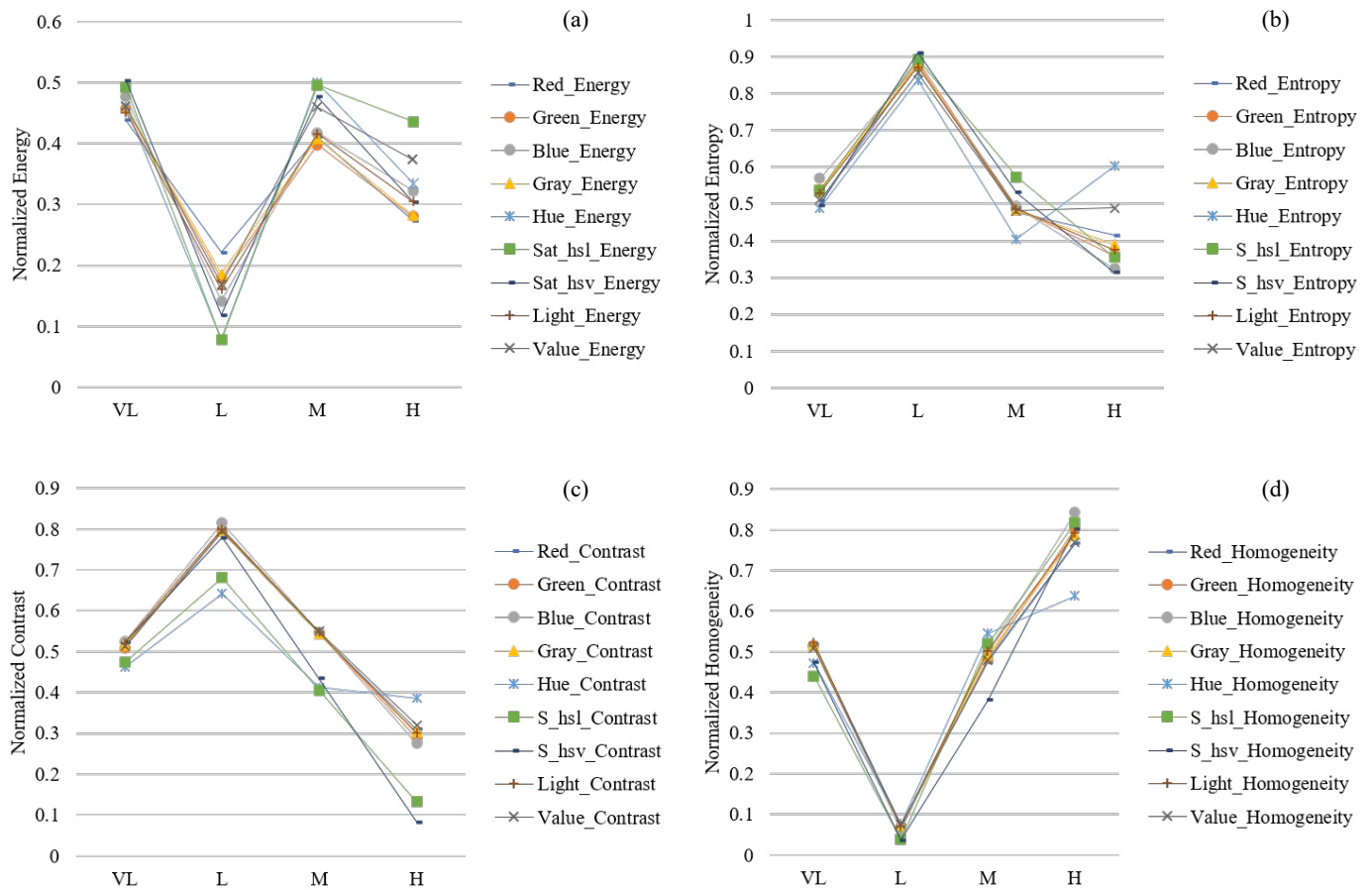


Figure 4: Normalized textural features different purity levels (very low (VL), low (L), medium (M) and high (H)) (a) Energy; (b) Entropy; (c) Contrast; (d) Homogeneity; (e) Inverse Difference Moment; (f) Correlation; (g) Sum Mean; (h) Variance; (i) Cluster; (j) Maximum Probability.

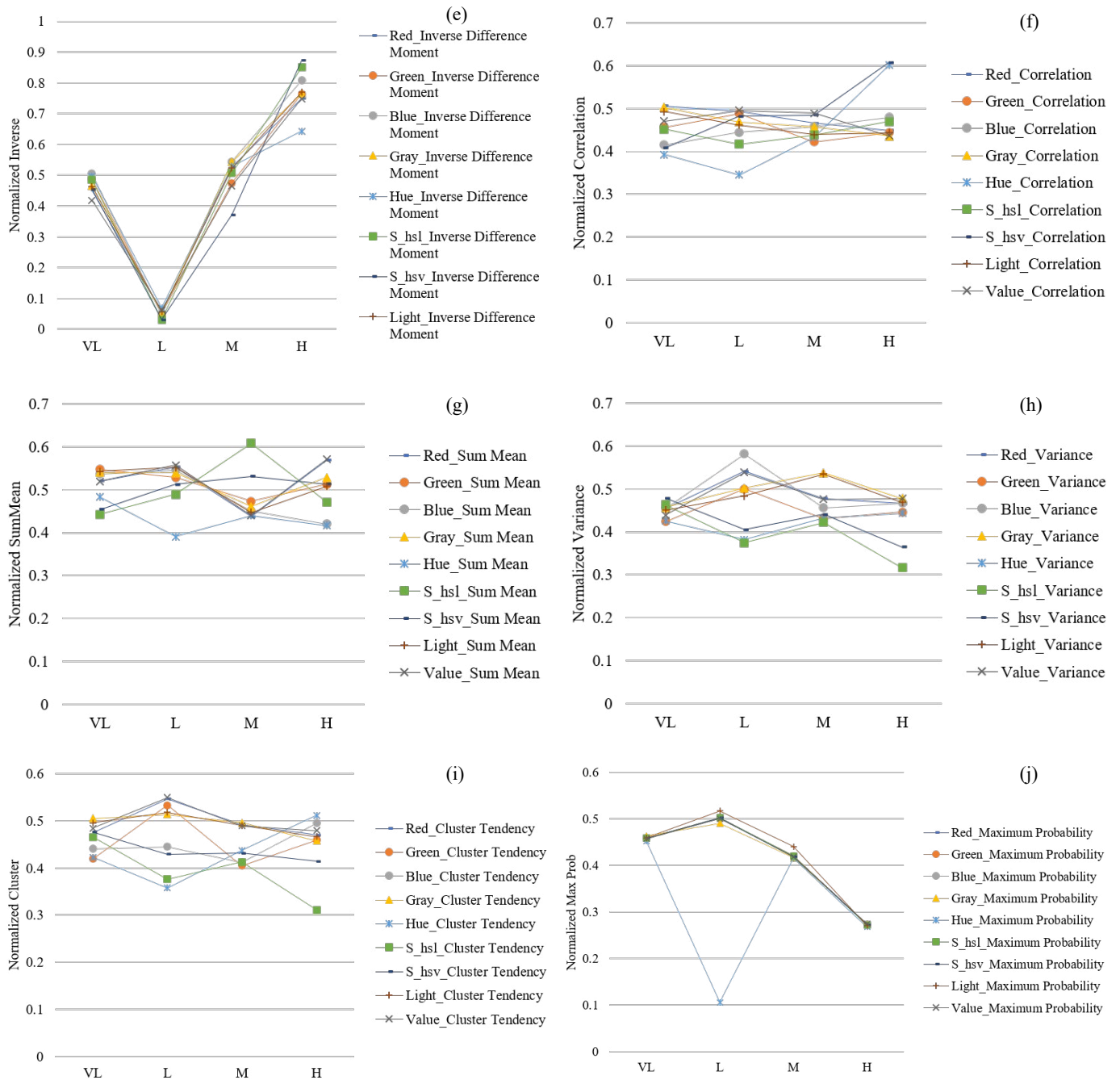


Figure 4: Continuation.

This study utilizes support vector machine classifier combined with the c parameter, gamma, and Radial Basis Function (RBF) kernel functions. Table 5 shows sensitivity analysis for color, textural, and all features. The best model from the support vector machine is a genetic algorithm with 5 selected features (G_Mean, B_Mean, H_Mean, V_Mean, Lab_b_Mean) because it not only provides high accuracy but

also considers the number of selected features. Moreover, the SVM classifier, when coupled with a learning algorithm, requires only a combination of color features to recognize the differences in Palm Civet Coffee. Table 6 shows using One Way ANOVA, there is a significant accuracy difference from color, texture and all features at $P\text{-value} = 0.05$ using SVM classifiers.

Table 1: Sensitivity Analysis for kNN Classifier.

No.	Features	Learning Algorithm	k-Value	No. of Features Selected	Accuracy
1.	Color Features	Bat Algorithm	1	3	0.918
		Cuckoo Search	1	3	0.899
		Genetic Algorithm	5	5	0.899
		Grey Wolf Optimizer	1	3	0.931
2.	Textural Features	Bat Algorithm	7	36	0.943
		Cuckoo Search	1	26	0.943
		Genetic Algorithm	5	4	0.943
		Grey Wolf Optimizer	5	5	0.941
3.	All Features	Bat Algorithm	5	34	0.943
		Cuckoo Search	7	19	0.937
		Genetic Algorithm	5	6	0.943
		Grey Wolf Optimizer	7	4	0.931

Table 2: One Way ANOVA for Accuracy of kNN Classifier.

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	0.020	2	0.010	30.229	1.55E-09	3.168
Within Groups	0.018	54	0.000			
Total	0.038	56				

Table 3: Sensitivity Analysis for Random Forest Classifier.

No.	Features	Learning Algorithm	ntree	No. of Features Selected	Accuracy
1.	Color Features	Bat Algorithm	200	4	0.874
		Cuckoo Search	100	5	0.881
		Genetic Algorithm	200	6	0.843
		Grey Wolf Optimizer	1000	3	0.881
2.	Textural Features	Bat Algorithm	100	3	0.887
		Cuckoo Search	200	36	0.962
		Genetic Algorithm	100	50	0.950
		Grey Wolf Optimizer	200	17	0.969
3.	All Features	Bat Algorithm	500	45	0.969
		Cuckoo Search	200	52	0.962
		Genetic Algorithm	100	51	0.962
		Grey Wolf Optimizer	500	5	0.981

Table 4: One Way ANOVA for Accuracy of Random Forest Classifier.

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	0.087	2	0.043	74.364	1.58E-14	3.220
Within Groups	0.025	42	0.001			
Total	0.111	44				

4.3 The Performance of Different Classifiers

The primary objective of this study was to optimize the purity classification of civet coffee using three classifiers (kNN, random forest, support vector machine) and four learning algorithms (bat algorithm, cuckoo search, genetic algorithm, grey wolf optimizer). It utilized three datasets (color feature, texture feature, and all features) for each proposed classifier and learning algorithm to compare the performance of each model. From each classifier, one best model was obtained, which was then compared with the others. Figure 5(a) shows that the Random Forest (RF) classifier produces the best accuracy compared to kNN and SVM. The results unequivocally demonstrated that the RF classifier, optimized with the Grey Wolf Optimizer (GWO) (RF-GWO) and utilizing a subset of five features, achieved the highest accuracy of 0.981. This superior performance of RF can be attributed to several inherent strengths of the algorithm. RF is renowned for its robustness and efficiency in handling complex classification tasks particularly when dealing with high-dimensional data. In the context of this study, this RF's ability to construct decision trees allowed it to capture intricate patterns and relationships within the color and textural features extracted from the coffee bean images. This ability enhances generalization performance, making RF particularly effective for distinguishing purity levels of Palm Civet Coffee. RF is renowned for its robustness and efficiency in handling complex classification tasks and particularly when dealing with high-dimensional data. In the context of this study, RF's ability to construct an ensemble of decision trees allowed it to capture intricate patterns and relationships within the color and textural features extracted from the coffee bean images. Unlike single decision trees, RF mitigates the risk of overfitting by averaging multiple trees, each trained on different subsets of the features. This ensemble approach enhances generalization performance, making RF particularly effective for distinguishing purity levels of Palm Civet Coffee. Previous research, (Hendrawan; Murase, 2011a) emphasized the importance of feature selection in improving prediction accuracy for agricultural products. As a result, the accuracy of the model increased when feature selection was performed using learning algorithms coupled with Artificial Neural Networks. This study, on the other hand, systematically compares several classifiers and learning algorithms to identify the optimal combination for Luwak Coffee purity classification to prevent adulteration. Unlike studies that focus only on color or texture features, we leverage the complementary strengths of both types of features, resulting in a more robust and accurate classification system. A statistical approach was also conducted to ascertain any significant differences in accuracy for each model.

Figure 5(b) shows that the kNN classifier selects the fewest features, but the random forest classifier is superior because the accuracy value is close to 1. The selected features that can distinguish the purity level of Kopi Luwak Kelapa are Blue_Mean, Hue_Entropy, Gray_Inverse, S_HSL_Correlation, and Green_Cluster.

Figure 5(b) indicates that the kNN classifier selected the fewest features, but the random forest classifier is superior because its accuracy value is close to 1. The selected features from best model play a crucial role in differentiating the purity levels of Palm Civet Coffee i.e Blue_Mean, Hue_Entropy, Gray_Inverse, S_HSL_Correlation, and Green_Cluster. Blue_Mean represents the average intensity of the blue channel in the RGB color space. Variations in Blue_Mean can indicate differences in bean coloration, which may correlate with purity levels. Hue_Entropy measures the randomness in the distribution of hue values. Higher entropy levels suggest greater variability in hue, potentially reflecting adulteration or mixing of different bean types. Gray_Inverse provides insight into the texture and surface properties of coffee beans. This happens because there is a difference in the smoothness of the images on civet and regular coffee beans. S_HSL_Correlation evaluates the correlation between piksel in the HSL color space. This feature helps in assessing the uniformity and consistency of color distribution. Green_Cluster assesses the clustering tendency of green color components, which can be indicative of specific quality attributes related to purity. The combination of these features enables the model to effectively distinguish between very low, low, medium, and high purity categories, demonstrating the nuanced changes in both color and texture associated with each level. Moreover, there is a significant difference in accuracy between the classifiers, as indicated by a P-value of 0.01, shown in Table 7. A sensitivity analysis of feature selection using three different classifiers for classifying Palm Civet Coffee can be carried out to select the best model.

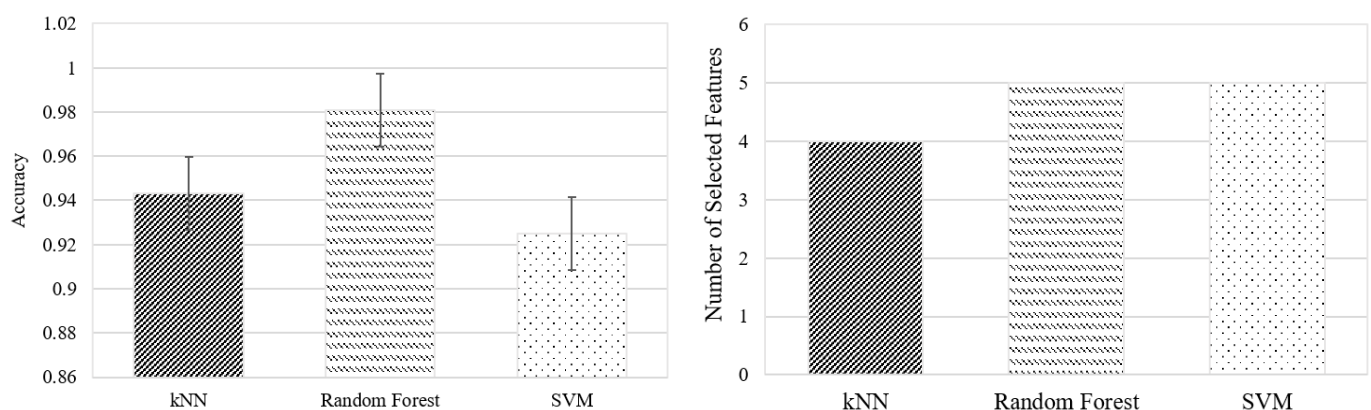
The high accuracy achieved by the RF-GWO model has significant implications for the Palm Civet Coffee industry. Implementing such a non-destructive, rapid, and cost-effective classification system can enhance quality control processes, reduce instances of adulteration, and ensure consumer trust in premium products like Palm Civet Coffee. Despite the promising results, several challenges and limitations must be acknowledged. Firstly, the study's dataset may not capture the full variability present in different batches of Palm Civet Coffee. This limitation could affect the model's generalizability. Future research should aim to address these limitations by expanding the dataset and exploring more sophisticated classification algorithms and learning algorithm.

Table 5: Sensitivity Analysis for Support Vector Machine Classifier.

No.	Features	Learning Algorithm	c	gamma	No. of Features Selected	Accuracy
1.	Color Features	Bat Algorithm	2^{10}	2^{-5}	7	0.925
		Cuckoo Search		2^{-5}	5	0.868
		Genetic Algorithm		2^{-10}	5	0.925
		Grey Wolf Optimizer		2^{-10}	1	0.321
2.	Textural Features	Bat Algorithm	2^{10}	2^{-10}	1	0.849
		Cuckoo Search		2^{-10}	3	0.868
		Genetic Algorithm		2^{-10}	17	0.321
		Grey Wolf Optimizer		2^{-5}	1	0.723
3.	All Features	Bat Algorithm	2^{10}	2^{-10}	13	0.321
		Cuckoo Search		2^{-10}	12	0.899
		Genetic Algorithm		2^5	13	0.321
		Grey Wolf Optimizer		2^0	1	0.635

Table 6: One Way ANOVA for Accuracy of SVM Classifier.

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	2.343	2	1.171	28.062	7.02E-12	3.026
Within Groups	12.272	294	0.042			
Total	14.615	296				

**Figure 5:** Comparison of (a) Accuracy and (b) Number of Selected Features from each classifiers.**Table 7:** One Way ANOVA for Each Classifier.

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	16.001	2	8.000	215.38	2.66E-64	3.018
Within Groups	14.932	402	0.037			
Total	30.933	404				

5 CONCLUSIONS

Machine vision effectively assesses Palm Civet Coffee purity non-destructively. By generating color and texture features, machine vision requires feature selection to eliminate irrelevant features, improve accuracy and prevent overfitting. Our research

shows that the Random Forest optimized with Grey Wolf Optimizer achieved highest accuracy of 0.981 using five features: Blue_Mean, Hue_Entropy, Gray_Inverse, S_HSL_Correlation, and Green_Cluster. These results validate the effectiveness of wrapper feature selection and machine vision in contributing to the authentication and protection of high-value Palm Civet Coffee.

6 AUTHORS' CONTRIBUTION

Conceptual idea: Widyaningtyas, S.; Arwani, M.; Sucipto; Hendrawan, Y. Methodology design: Arwani, M. Data collection: Widyaningtyas, S. Data analysis and interpretation: Widyaningtyas, S.; Arwani, M.; Sucipto; Hendrawan, Y. Writing and editing: Widyaningtyas, S.; Arwani, M.

7 REFERENCES

- ADHAO, R.; PACHGHARE, V. Feature selection using principal component analysis and genetic algorithm. **Journal of Discrete Mathematical Sciences and Cryptography**, 23(2):595-602, 2020.
- AFSHAR, M.; USEFI, H. Optimizing feature selection methods by removing irrelevant features using sparse least squares. **Expert Systems with Applications**, 200:116928, 2022.
- AKSAKALLI, V. et al. Feature selection using stochastic approximation with Barzilai and Borwein non-monotone gains. **Computers & Operations Research**, 132:105334, 2021.
- ALI, N.; NEAGU, D.; TRUNDLE, P. Evaluation of k-nearest neighbour classifier performance for heterogeneous data sets. **SN Applied Sciences**, 1:1-15, 2019.
- ALYASSERI, Z. A. A. et al. Recent advances of bat-inspired algorithm, its versions and applications. **Neural Computing and Applications**, 34(19): 16387–16422, 2022.
- ALZAQBAH, M. et al. Memory based cuckoo search algorithm for feature selection of gene expression dataset. **Informatics in Medicine Unlocked**, 24:100572, 2021.
- AMINI, F.; HU, G. A two-layer feature selection method using genetic algorithm and elastic net. **Expert Systems with Applications**, 166:114072, 2021.
- BISWAS, S.; BORDOLOI, M.; PURKAYASTHA, B. Review on feature selection and classification using neuro-fuzzy approaches. **International Journal of Applied Evolutionary Computation (IJAEC)**, 7(4):28-44, 2016.
- CERVANTES, J. et al. A comprehensive survey on support vector machine classification: Applications, challenges and trends. **Neurocomputing**, 408:189-215, 2020.
- CHANDRASHEKAR, G.; SAHIN, F. A survey on feature selection methods. **Computers & Electrical Engineering**, 40(1):16-28, 2014.
- COSMA, G. et al. A survey on computational intelligence approaches for predictive modeling in prostate cancer. **Expert systems with applications**, 70:1-19, 2017.
- FANG, Z.; YU, X.; ZENG, Q. Random forest algorithm-based accurate prediction of chemical toxicity to *Tetrahymena pyriformis*. **Toxicology**, 480:153325, 2022.
- FAWAGREH, K.; GABER, M. M.; ELYAN, E. Random forests: from early developments to recent advancements. **Systems Science & Control Engineering: An Open Access Journal**, 2(1):602-609, 2014.
- FENG, Q.; LIU, J.; GONG, J. UAV remote sensing for urban vegetation mapping using random forest and texture analysis. **Remote sensing**, 7(1):1074-1094, 2015.
- GOEL, L. An extensive review of computational intelligence-based optimization algorithms: trends and applications. **Soft Computing**, 24(21):16519-16549, 2020.
- HENDRAWAN, Y. et al. Deep learning to detect and classify the purity level of luwak coffee green beans. **Pertanika Journal of Science & Technology**, 30(1):1-18, 2022.
- HENDRAWAN, Y.; MURASE, H. Neural-Genetic algorithm as feature selection technique for determining sunagoke moss water content. **Engineering in Agriculture, Environment and Food**, 3(1):25-31, 2010.
- HENDRAWAN, Y.; MURASE, H. Bio-inspired feature selection to select informative image features for determining water content of cultured Sunagoke moss. **Expert Systems with Applications**, 38(11):14321-14335, 2011a.
- HENDRAWAN, Y.; MURASE, H. Neural-intelligent water drops algorithm to select relevant textural features for developing precision irrigation system using machine vision. **Computers and Electronics in Agriculture**, 77(2):214-228, 2011b.
- HENDRAWAN, Y.; WIDYANINGTYAS, S.; SUCIPTO, S. Computer vision for purity, phenol, and pH detection of Luwak Coffee green bean. **TELKOMNIKA (Telecommunication Computing Electronics and Control)**, 17(6):3073-3085, 2019.
- IRANZAD, R.; LIU, X. A review of random forest-based feature selection methods for data science education and applications. **International Journal of Data Science and Analytics**, 1-15, 2024.

- JUMHAWAN, U. et al. Selection of discriminant markers for authentication of Asian palm civet coffee (Kopi Luwak): a metabolomics approach. **Journal of agricultural and food chemistry**, 61(33):7994-8001, 2013.
- JUMHAWAN, U. et al. Application of gas chromatography/flame ionization detector-based metabolite fingerprinting for authentication of Asian palm civet coffee (Kopi Luwak). **Journal of bioscience and bioengineering**, 120(5):555-561, 2015.
- JUMHAWAN, U. et al. Quantification of coffee blends for authentication of Asian palm civet coffee (Kopi Luwak) via metabolomics: A proof of concept. **Journal of bioscience and bioengineering**, 122(1):79-84, 2016.
- KAR, A.K. Bio inspired computing-a review of algorithms and scope of applications. **Expert Systems with Applications**, 59:20-32, 2016.
- KASHEF, R. A boosted SVM classifier trained by incremental learning and decremental unlearning approach. **Expert Systems with Applications**, 167:114154, 2021.
- KELIDARI, M.; HAMIDZADEH, J. Feature selection by using chaotic cuckoo optimization algorithm with levy flight, opposition-based learning and disruption operator. **Soft Computing**, 25(4):2911-2933, 2021.
- LEE, W.-S. et al. Sensing technologies for precision specialty crop production. **Computers and electronics in agriculture**, 74(1):2-33, 2010.
- MAFARJA, M. et al. An efficient high-dimensional feature selection approach driven by enhanced multi-strategy grey wolf optimizer for biological data classification. **Neural Computing and Applications**, 35(2):1749-1775, 2023.
- MAFARJA, M.; MIRJALILI, S. Whale optimization approaches for wrapper feature selection. **Applied Soft Computing**, 62:441-453, 2018.
- MALEKI, N.; ZEINALI, Y.; NIAKI, S. T. A. A k-NN method for lung cancer prognosis with the use of a genetic algorithm for feature selection. **Expert Systems with Applications**, 164:113981, 2021.
- MALLIDI, S. Bat optimization algorithm for wrapper-based feature selection and performance improvement of android malware detection. **IET Networks (Wiley-Blackwell)**, 10 (3), 2021.
- MARCONI, M. F. Composition and properties of Indonesian palm civet coffee (Kopi Luwak) and Ethiopian civet coffee. **Food Research International**, 37(9):901-912, 2004.
- MOUSAVIRAD, S.; EBRAHIMPOUR-KOMLEH, H. Wrapper feature selection using discrete cuckoo optimization algorithm. **International Journal of Mechatronics, Electrical, and Computer Engineering**, 4(11):709-721, 2014.
- REDA, M. et al. Path planning algorithms in the autonomous driving system: A comprehensive review. **Robotics and Autonomous Systems**, 174:104630, 2024.
- REN, G. et al. Evaluating Congou black tea quality using a lab-made computer vision system coupled with morphological features and chemometrics. **Microchemical Journal**, 160:105600, 2021.
- ROSTAMI, M. et al. Review of swarm intelligence-based feature selection methods. **Engineering Applications of Artificial Intelligence**, 100:104210, 2021.
- SAADATFAR, H. et al. A new K-nearest neighbors classifier for big data based on efficient data pruning. **Mathematics**, 8(2):286, 2020.
- SEJUTI, Z. A.; ISLAM, M. S. A hybrid CNN-KNN approach for identification of COVID-19 with 5-fold cross validation. **Sensors International**, 4:100229, 2023.
- SHANG, Y. et al. An effective feature selection approach based on hybrid Grey Wolf Optimizer and genetic algorithm for hyperspectral image. **Scientific Reports**, 15(1):1968, 2025.
- SUHANDY, D.; YULIA, M. The use of partial least square regression and spectral data in UV-visible region for quantification of adulteration in Indonesian palm civet coffee. **International journal of food science**, 1-7, 2017.
- TIWARI, A.; CHATURVEDI, A. A hybrid feature selection approach based on information theory and dynamic butterfly optimization algorithm for data classification. **Expert Systems with Applications**, 196: 116621, 2022.
- TSAL, C.-F.; CHEN, Y.-C. The optimal combination of feature selection and data discretization: An empirical study. **Information Sciences**, 505:282-293, 2019.
- UDDIN, S. et al. Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction. **Scientific Reports**, 12(1):6256, 2022.

- WANG, J. et al. An adaptively balanced grey wolf optimization algorithm for feature selection on high-dimensional classification. **Engineering Applications of Artificial Intelligence**, 114:105088, 2022.
- YU, L.; LIU, H. Efficient feature selection via analysis of relevance and redundancy. **The Journal of Machine Learning Research**, 5:1205-1224, 2004.
- ZHANG, P.; GAO, W. Feature selection considering uncertainty change ratio of the class label. **Applied Soft Computing**, 95:106537, 2020.
- ZHAO, W.; FORTE, E.; PIPAN, M. Texture attribute analysis of GPR data for archaeological prospection. **Pure and Applied Geophysics**, 173:2737-2751, 2016.