Agr. Nat. Resour. 57 (2023) 615–624

AGRICULTURE AND NATURAL RESOURCES

Journal homepage: http://anres.kasetsart.org

Research article

Optimization of information gain interval for determining artificial ripeness of banana using image data with imbalanced class

Candra Dewi[a,b,*], Endang Arisoesilaningsih[a,†], Wayan Firdaus Mahmudy[b,†], Solimun[c,†]

[a] *Biology Department, Faculty of Science, Universitas Brawijaya, Malang 64145, Indonesia*

[b] *Informatics Department, Faculty of Computer Science, Universitas Brawijaya, Malang 64145, Indonesia*

[c] *Statistics Department, Faculty of Science, Universitas Brawijaya, Malang 64145, Indonesia*

Article Info

Article history:
Received 5 April 2023
Revised 2 June 2023
Accepted 14 August 2023
Available online 31 August 2023

Keywords:
Banana artificial ripeness,
k-Means,
Optimized IG,
Scott rule,
Sturgess rule

Abstract

Importance of the work: Precise determination of ripeness is crucial for post-harvest processing of fruits. Moreover, distinguishing between natural and artificial ripeness of banana fruits requires a specific feature because they have similar physical appearance.

Objectives: To optimize the information gain (IG) interval to obtain the optimal features in identifying artificial banana ripeness with imbalanced class data.

Materials & Methods: The test was based on six Indonesian banana cultivars using 11,593 images. In total, 78 features were extracted using morphological descriptors, a convex hull, a local binary pattern and a gray-level co-occurrence matrix. IG optimization was based on the Sturgess rule, Scott rule and K-means clustering. Oversampling based on the synthetic minority oversampling technique (SMOTE) was used to handle imbalanced data.

Results: The results of the identification using extreme learning machine classification of imbalanced data showed higher accuracy based on IG optimization using the Sturgess and Scott rules rather than the use of IG. The implementation of SMOTE also substantially increased the accuracy from 20% to 40% compared to the result with imbalanced data. Most of the accuracy (80%) resulted from using selected features for four cultivars (ambon lumut, hijau, kepok and raja). The two other banana cultivars (morosebo and susu) had accuracy levels of more than 71% and 76%, respectively.

Main finding: Due to the complexity of choosing the optimum number of IG bin intervals for data with very high similarity characteristics, optimization using Sturgess and Scott rules produced greater accuracy, especially with imbalanced data.

† Equal contribution.

* Corresponding author.
E-mail address: dewi_candra@ub.ac.id (C. Dewi)

online 2452-316X print 2468-1458/Copyright © 2023. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/), production and hosting by Kasetsart University Research and Development Institute on behalf of Kasetsart University.

https://doi.org/10.34044/j.anres.2023.57.4.06

Introduction

Bananas are widely consumed in Indonesia and are readily found in traditional markets or supermarkets because Indonesia has a high level of banana diversity (De Langhe et al., 2009), with very high production (BPS, 2022). The increasing level of consumption of bananas has spurred efforts to accelerate the ripening process of the fruit using chemicals. However, the regular consumption of artificially ripened banana fruits based on the excessive use of chemicals in the ripening stage could harm consumer health. Field observations suggest that fruits that are ripened naturally (natural ripening) compared to those that have been chemically cured (artificial ripening) have almost the same physical appearance and it is very difficult to distinguish the difference. Thus, it is very difficult for consumers to consciously select quality fruit with natural ripening.

Currently, computer vision using various algorithms is widely used in the identification of banana cultivars (Sabilla et al., 2019; Dewi et al., 2022c), fruit ripeness (Harel et al., 2020; Pourdarbani et al., 2020; Behera et al., 2021;) and banana ripeness (Sanaeifar et al., 2016; Kipli et al., 2018; Mazen and Nashat, 2019). Nevertheless, the identification of fruit (and banana ripeness in particular) has been limited to the ripeness level without identifying artificial ripeness. However, identification of artificial ripeness was carried out by Rahman et al. (2021) and Maheswaran et al. (2018) on mango fruits, with very good results. These researchers showed the potential to identify artificial ripeness of fruits. However, each fruit had quite different physical characteristics. Therefore, Vaviya et al. (2019) identified the artificial ripeness of banana fruit using a convolution neural network (CNN) with a high level of accuracy. However, they only distinguished natural and artificial ripeness, without considering ripeness levels, with such refinement being difficult using CNN. Thus, further research is required related to the artificial ripeness of banana fruits.

Distinguishing between natural and artificial ripeness of banana fruits requires a specific feature because they have similar physical appearance. Dewi et al. (2022b) succeeded in identifying the ripening of susu bananas at the ripening level with high similarity using features selected using information gain (IG) and principal component analysis (PCA). The success of IG in feature selection has been reported in various studies (Di Franco, 2019; Elmaizi et al., 2019; Tang et al., 2019; Singer et al., 2020). However, it is necessary to simplify the calculation of IG with continuous data, to categorize each feature and select the optimal number of intervals. This process requires testing large number of intervals; therefore, an additional method is needed to speed up the calculation automatically. IG optimization with genetic algorithms (GA) for feature selection was investigated by Jadhav et al. (2018), with GA requiring complex calculations and a long time to achieve convergence. The weighted information-gain ratio (WIGR) was introduced by Singer et al. (2020) and was only optimal if used in an ordinal ordering type of variable class and with non-monotonic dependencies between attributes and class values. Research conducted by Dewi et al. (2022b) used the Sturgess rule for IG optimization. The Sturgess rule is efficient for determining the histogram bin size based on mutual information (Legg et al., 2007). Nevertheless, the Sturgess rule is not appropriate for use on data with an abnormal distribution or with large amounts of data (Hyndman, 1995). Due to non-normal distribution of banana fruit data, the current was carried out based on optimization using the Scott rules.

In addition, the division of bins into a number of intervals is similar to grouping data into several clusters, where the data in one group are similar in some way and dissimilar to the data in other groups. K-means is one of the clustering methods that is quite widely used (Govender and Sivakumar, 2020) because it has low complexity, is fast and can handle large amounts of data (Jain, 2010). Based on this, another approach with K-means clustering was also carried out to determine the number of intervals by assuming the number of clusters was analogous to the number of intervals.

This study investigated six local Indonesia banana cultivars, where each cultivar had different characteristics during ripening. Based on observations, not all banana fingers on some cultivars were evenly yellow at the ripe level. Furthermore, an evenly yellow color change is apparent after the yellowish-green (level 2). Thus, the number of fruits with a greenish-yellow color (level 3) becomes smaller, which produces an imbalanced class and affects recognition in multi-class classification (Tarekegn et al., 2021) because features with minority classes are difficult to recognize (Jeatrakul et al., 2010). The current study aimed to overcome this problem by oversampling based on the synthetic minority over-sampling technique (SMOTE). Oversampling techniques have proven to be more effective than under-sampling (Xie and Qiu, 2007). Studies have also shown that SMOTE could improve recognition accuracy using neural networks in imbalanced classes (Dewi et al., 2020; Mahmudy et al., 2021). Furthermore, a comprehensive analysis conducted by Elreedy and Atiya (2019) showed that SMOTE could improve accuracy for both small and large datasets.

Based on the description above, the purpose of the current study was to optimize IG to select optimal features in the identification of artificial ripeness of banana fruit using a dataset of banana finger images with imbalanced classes. Specifically, the objectives of the study were: 1) performance analysis of IG optimization using the Sturgess rule, Scott rule and K-means clustering; 2) performance analysis of imbalanced class handling using SMOTE; and 3) identification of the artificial ripeness and maturity levels of banana fruit.

Materials and Methods

The banana fruits used were ordered from farmers located in Tawangrejo village, Binangun district, Blitar regency, East Java, Indonesia. Six cultivars used: ambon lumut, hijau, kepok, morosebo, raja and susu. The fruit harvest period for each cultivar includes two different maturities, namely "high three-quarters" (end/old harvested) and "three-quarter round" (mid/rather old harvested) proposed by Serrano and Rolle (2018).

Approximately six tiers were taken from each bunch of bananas to represent the upper, middle and lower portions. Each tier was divided into two treatments: naturally ripened (without the use of chemicals) and ripened using chemicals. The chemical substance used as a fruit ripener was ethephon 10% in the ratio of 1:1.5 mL of ethephon diluted with 1 L of water. Each banana finger was separated from the tiers and cured from green peel to yellow-brown spots, according to the observed level of ripening, as described in Table 1.

Each cultivar with a different treatment requires a different amount of time for peel discoloration. In addition, the time for each cultivar to reach a certain level does not occur simultaneously for each finger of the fruit. Generally, banana fingers located on the upper tier (1 and 2) will ripen first, while the lower tier takes a longer time to deepen its color. Based on observations, the time difference between natural and chemical ripening is in the range 0–9 d. The old harvest fruit generally takes the same amount of time whether it has been naturally or chemically ripened. However, the use of chemicals allows more fruits to ripen simultaneously. And can accelerate the curing and shelf life of rather old harvested fruits.

Images were captured for each finger of the banana at each level of ripeness. The fruit was placed on white paper and the images were captured using a mobile phone camera with a distance of 15–20 cm from the top of the fruit. All images were captured in a room with sufficiently bright sun light to reduce any shadow effect. Each group of data collected was labeled to distinguish the age of fruit harvest with natural or chemical curing. Images of the raja cultivar with high traits at each ripeness level with natural and chemical ripening are provided in Fig. 1.

The number of banana fingers in each tier of a cultivar varies and it influenced the numbers in the datasets. In addition, the technique of curing and the age of the harvest affected the number of datasets at each level. For example, cultivars harvested at a rather old age with natural curing had a very small dataset that reached level 4. The kepok banana that was harvested rather old had almost no images of evenly yellow. The fruit is considered ready to be consumed when the peel color is yellow with a green shoot and slightly-soft-to-soft pulp. In most fruits, brown spots will immediately appear with a brown base of the fruit, even though the fruit remains yellowish-green. This is due to the length of time of curing so that the peel becomes dull. The number of datasets at each ripeness level is shown in Table 2 for naturally ripened bananas and in Table 3 for artificially ripened bananas.

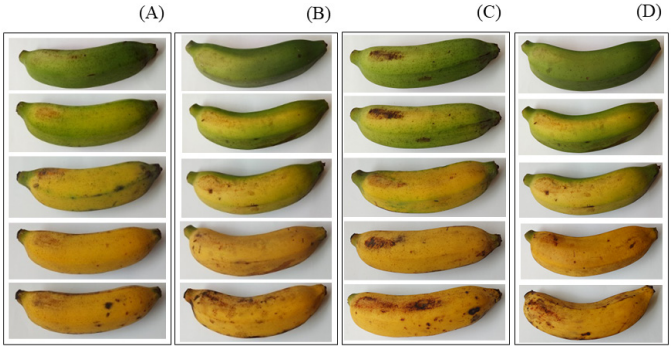


Fig. 1 Images of raja cultivar at all ripeness levels: (A) natural ripening and old harvested; (B) artificial ripening and old harvested; (C) natural ripening and rather old harvested; (D) artificial ripening and rather old harvested

Table 1 Banana ripeness levels

Level	Description
1	Green peel color and hard fruits
2	Green peel color with slightly yellow and rather hard fruit
3	Yellow peel color and rather soft fruit
4	Evenly yellow peel color (some evenly yellow with green shoots/bases) and soft fruit
5	Yellow peel color with brown spots and very soft fruit

Table 2 Number of images in datasets at each level with natural ripening

Cultivar	Harvest time	Number of datasets at each level					Total
		1	2	3	4	5	
Ambon Lumut	End	52	59	96	91	94	392
	Mid	96	112	105	109	124	546
Hijau	End	98	110	45	79	116	448
	Mid	118	102	183	115	133	651
Morosebo	End	54	53	51	63	54	275
	Mid	84	58	67	60	82	351
Kepok	End	84	26	104	100	51	365
	Mid	120	52	172	2	117	463
Raja	End	84	78	74	159	101	496
	Mid	113	62	204	138	134	651
Susu	End	89	66	55	278	73	561
	Mid	87	24	88	184	92	475

Table 3 Number of images in datasets at each level with artificial ripening

Cultivar	Harvest time	Number of datasets at each level					Total
		1	2	3	4	5	
Ambon Lumut	End	51	68	116	88	94	417
	Mid	98	103	147	137	106	591
Hijau	End	94	112	50	92	139	487
	Mid	116	73	112	157	142	600
Morosebo	End	53	55	63	50	99	320
	Mid	81	63	61	110	67	382
Kepok	End	91	35	53	117	52	348
	Mid	109	42	134	174	132	591
Raja	End	96	80	75	232	163	646
	Mid	112	69	162	118	140	601
Susu	End	89	34	85	150	86	444
	Mid	86	58	62	192	94	492

The process of identifying banana ripeness consists of several stages, as shown in Fig. 2. The input of the process was banana finger image with a size of 4,160×3,120 pixels. The images were resized to 20% of the original size (832×624 pixels), because the use of large images required a long computational time and large resources. The size of this image has been tested and reported to not reduce the quality of the recognition results (Dewi et al., 2022a). The process continued with image segmentation using Gaussian blur with a combination of binary inverse and Otsu thresholds to separate the fruit part from the background (Dewi et al., 2022a).

Feature extraction collected information on color features, shape features and texture features. Referring to previous research, the color moment feature extraction method was used to extract color features (Dewi et al., 2022b), while morphological feature extraction and convex hull were used to

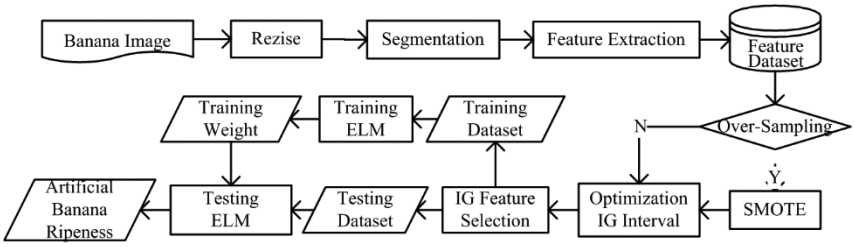


Fig. 2 Flow diagram of optimization of IG in artificial banana ripeness identification with imbalanced classes, where ELM = extreme learning machine classifier, IG = information gain feature selection and SMOTE = synthetic minority over-sampling technique

extract shape features (Dewi et al., 2022c), as well as applying local binary pattern (LBP) and grey level co-occurrence matrix (GLCM) texture feature extraction. Each level of banana ripeness has a difference in peel color ranging from green to yellow with brown spots. Consequently, the color feature was used to define the color change. The shape feature is usually successfully used to recognize banana cultivars, because some cultivars have very different shapes and sizes. Although it is a bit difficult to distinguish artificial ripeness based on the shape of the fruit, artificial ripeness generally was done for rather old harvested banana samples. According to Serrano and Rolle (2018), banana fruits harvested at old maturity (full/high three quarters) have a full round shape, while fruits harvested at rather old maturity (light full three-quarters) are generally still angular. Nevertheless, the corners of the fruit are not sufficiently recognizable by shape features and so this needs to be combined with the texture of the peel surface.

The research conducted by Dewi et al. (2022b) used a combination of color and LBP texture features to recognize the ripeness level of susu bananas at the ripened level (3 sub-levels were characterized based on an evenly-yellow peel color). Nevertheless, naturally ripe banana fruits and chemicals often have a similar color and shape, especially in old harvested bananas, with differences in the texture of the peel surface very difficult to distinguish; thus the combination of all three features was used in the current study. Additionally, the texture features were not only extracted with LBP, which only uses local information of neighborliness from each pixel observed, but also texture features were extracted based on the grey level co-occurrence matrix (GLCM), which utilized neighboring pixel information globally on the entire image. Test results have shown that the optimal number of LBP features was 26. The six GLCM features introduced by Haralick et al. (1973) were extracted: contrast, energy, homogeneity, correlation, similarity and angular second moment (ASM; Yao et al., 2009). These GLCM features were extracted from the matrix for each RGB image channel at a distance of 1 and using four orientation angles of 0°, 45°, 90°, and 135°. The total number of GLCM features extracted was 24, with 78 total features, consisting of 15 color moment features, 13 descriptor morphology and convex hull features, 26 LBP feature and 24 GLCM features.

The normality and distribution of each feature were checked using histograms and Kolmogorov Smirnov tests. The p-value of normality obtained at the 95% confidence level was very low and below the cyclical level of 0.05; therefore, H_0 was rejected, indicating that the data distribution was abnormal. The data distribution of most color moment features was skewed to the left, while the morphological feature data and convex hull data tended to be close to the average value. In the LBP feature, some data were skewed to the right (close to the average value) and some data ware close to the minimum value. The GLCM feature had a very left-leaning data distribution.

Furthermore, feature selection was carried out to obtain features having a high influence on optimal recognition. The current research uses entropy-based IG (Leung et al., 2011; Shaltout et al., 2015). To obtain the optimal number of IG intervals, optimization was carried out using the Sturgess rule, Scott rule and the k-Means clustering method. The bin number was calculated using the Scott rule, as shown in Equation 1:

$$w = 3.49 \frac{s}{\sqrt[1/3]{n}} \quad (1)$$

where s is the standard deviation of the sample from n data. This equation was used to minimize bias in histogram variants. The Sturgess rule produced the same number of intervals for each feature, while the Scott rule produced a different number of intervals for each feature. This study also applied K-means clustering to divide the dataset into certain groups. The number of clusters formed was then used to express the number of intervals. Partition-based K-means clustering is widely used because it has low complexity, is fast and can handle large amounts of data (Jain, 2010; Govender and Sivakumar, 2020). K-means was used to divide the data with dimension n into k partitions with each partition containing at least 1 piece of data. The goal of K-means is to minimize the number of error quotients on all K clusters, using Equation 2:

$$J = \sum_{j=1}^{k} \sum_{i \in c_j} \left\| x_i - c_j \right\|^2 \quad (2)$$

where J is the criteria function, x_i is the ith data, c_j is the center of the jth cluster, c_j is the data set on cluster j and k is the number of clusters. The number of intervals was assumed to be equal to the number of clusters and the data as members of the cluster were also considered members of the interval. Therefore, each feature had the same number of intervals.

Once the features dataset had been obtained, over-sampling was carried out to handle imbalanced classes using SMOTE. The synthetic data based on the k-nearest neighbors was created to increase the amount of minor class data. This paper used the SMOTE method as proposed by Chawla et al. (2002).

In order to determine the effectiveness of using the feature selection, the classification was done using ELM. Based on research conducted by Sun et al. (2016), ELM was effectively used for multilabel classification and proven to

have better performance than another method (Huang et al., 2006; Chorowski et al., 2014; Ding et al., 20014). In addition, ELM provided optimal results for both the identification of the cultivar and the ripeness levels of banana fruits with high similarity (Dewi et al., 2022a, b, c). The ELM training and testing process in this study was carried out with a number of 3 k-folds to divide the training and test datasets.

Results and Discussion

The simulation and testing was performed using a python program and utilized several libraries (opencv-python, scipy. stats, Python.scikit-image, Extreme Learning Machine and KMeans. Each ELM hyperparameter test scenario with the input of all features and a selected feature was repeated five times to determine the consistency of the test and then the average accuracy was calculated to obtain the final value.

Feature selection using an IG manual interval was done on 3, 5, 7 and 15 intervals, while IG optimization with K-means was carried out on 3, 5 and 7 clusters. Furthermore, both IG manual interval and optimized IG using either the Sturgess rule (IG-Sturgess), Scott rule (IG-Scott) or K-means (IG-K means) were conducted for thresholding to reduce the number of features by as much as 75%, 50% and 25% of the total number of features. The tests showed that the optimal number of features was 75% of the total features. The best selected features for each cultivar are displayed in Table 4.

The results of feature selection showed that the shape features were not dominant, with no shape features being selected for the ambon lumut cultivar using IG-Sturgess and for the hijau cultivar using IG-Scott. In addition, the results showed that the color features were quite dominant in the selections, while the LBP and GLCM features were also quite dominant, although the percentage in each method varied. These results revealed that the color and texture features were quite dominant in the selection process using an IG manual interval, IG-Sturgess and IG-Scott.

Result of information gain optimization

This study carried out testing on six banana cultivars that had different characteristics, as well as different amounts of data. The cultivars ambon lumut, hijau, kepok, morosebo and raja consisted of 20 classes (10 classes for old and mid harvested bananas with natural ripening and 10 classes the same timing of harvest with artificial ripeness. The susu bananas consists of 24 classes, namely 12 classes each on natural and artificial ripening. The test results using all features and features selected using an IG manual interval and an optimized IG are shown in Fig. 3.

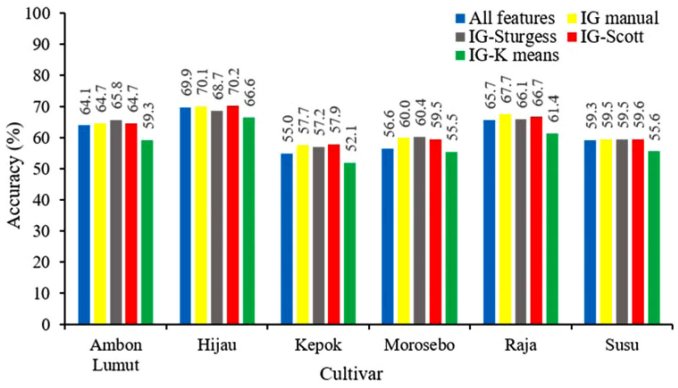


Fig. 3 Maturity identification results using all features and selected features

Table 4 Features selected using information gain (IG) feature selection and IG optimization

Cultivar	Method	Selected features
Ambon Lumut	IG – Sturgess	0,1,2,3,4,5,6,7,8,9,10,11,14,29,30,31,33,34,35,36,37,38,39,40,41,42, 43,44,45,46,47,52,53,54, 55,56,57, 58,59,60,61,62,63,64,65,66, 67,68, 69,70,71,72,73,74,75,76,77
Hijau	IG - Scott	0,1,2,3,4,5,7,8,10,11,14,28,29,30,31,32,33,34,35,36,37,38,39,40, 42,43,44,45,46,47,48,49,50,52,53,54,5 5,56,57,58,59,60,61,62,63, 64,65, 66,67,69,70,71,72,73,74,75,76,77
Kepok	IG - Scott	0,1,2,3,4,6,7,8,9,10,12,13,14,15,16,17,18,19,28,32,33,34,35,36, 37,38,39,40,41,42,43,44,45,46,47,48,49 ,51,52,55,57,58,59,60,61, 63, 64,65,66,67,69,70,71,72,73,75,76,77
Morosebo	IG – Sturgess	1,3,6,7,8,9,10,12,13,15,16,17,18,19,20,25,31,32,33,34,35,36,37, 38,39,40,41,42,43,44,46,48,50,52,53,5 4,55,56,57,58,59,60,61,62, 63,64,65,66,67,68,69,70,71,72,73,74,75,76
Raja	IG - Manual	0,1,2,3,4,5,6,7,8,9,10,11,15,16,17,18,19,20,21,25,28,29,30,31,33, 34,35,36,37,38,39,40,41,42,43,44,45, 46,47,50,51,53,54,55,58,59, 60,61,64,65,66,67,70,71,72,73,76,77
Susu	IG - Scott	0,1,3,4,7,9,10,11,12,13,16,17,19,20,26,28,30,31,32,33,35,36,37, 41,42,43,44,45,47,48,49,50,51,53,54,5 5,56,57,58,59,60,61,62,63, 64,65,66,67,68,69,70,71,72,73,74,75,76,77

Code values of 0–77 were used to refer to the 78 features (0–14 for color features, 15–27 for shape features, 28–53 for LBP features and 54–77 for GLCM features.)

The very high similarity observed between natural and artificial ripening bananas was reflected in the low recognition accuracy. Observations on the old harvested bananas showed that both natural and artificial ripening took almost the same time at all five levels of ripening, resulting in the changes in color and peel surface texture being very similar. Although rather old harvested bananas in the lower tier required a longer time for natural ripening, they still had almost the same color as artificially ripened ones. Consequently, they were still very difficult to distinguish. Cultivars of ambon lumut, hijau and raja produced higher recognition accuracy (above 60%) compared to kepok, morosebo and susu (below 60%).

Accuracy of identification of all cultivars using IG manual interval feature selection resulted in higher levels of accuracy compared to using all features (Fig. 3). A moderate increase in accuracy (greater than or equal to 2%) was recorded for the kepok, morosebo and raja cultivars, while the increase in accuracy for the ambon lumut, hijau and susu cultivars was not more than 0.6%.

The elimination of all shape features and the presence all GLCM features on selection using IG-Sturgess for ambon lumut produced the best accuracy. The IG manual interval and other IG optimizations still selected most of the shape features, suggesting that the shape characteristics had very little influence on the identification of the artificial ripeness of the ambon lumut banana sample.

The highest accuracy obtained for the hijau bananas was from using Scott rule optimization. According to Table 4, the results of the feature selection with this rule also eliminated all shape features. Unlike for the ambon lumut banana, more shape features for the hijau cultivar sample did not reduce the accuracy of recognition. The optimal combination of color, LBP and GLCM features could improve recognition accuracy.

Testing the kepok banana cultivar showed that the shape characteristics of the fruit had a moderate influence on the recognition. This was indicated by the presence of a shape feature in IG-Scott which produced the highest accuracy and likewise, with using an IG manual interval. Most of the shape features that had a considerable influence were combinations of the five basic morphological features (width, length, diameter, perimeter, area).

Removal of all shape features with IG-Scott for the raja cultivar was not effective at improving the accuracy, whereas the use of the IG manual interval resulted in higher accuracy. However, the use of all the features with IG-K means was ineffective, producing the lowest accuracy. As for other cultivars, combinations of color features, LBP and GLCM predominated in recognition.

As for the kepok and raja properties, the influence of basic morphological features on the morosebo banana was quite high. This was demonstrated in the use of IG-Sturgess that resulted in the highest accuracy. However, the use of all the shape features did not produce optimal results, as was indicated with the use of IG-K means.

Susu bananas were somewhat different from the other cultivars, because the influence of shape features did not show a specific pattern. The selection of IG interval manual, IG-Sturgess and IG-Scott produced almost the same accuracy; however, the shape features selected in these three methods were quite diverse.

The results of this test showed a combination of most color features, BLP and GLCM was optimal in the recognition of naturally and artificially ripened banana harvested at different maturity levels. The influence of shape features was low for the ambon lumut and hijau bananas. However, the use of shape features, especially basic morphological features, was still necessary for some cultivars.

Fig. 3 shows that IG optimization with the Sturgess and Scott rules could increase accuracy. However, clustering K-means produced lower accuracy compared to the IG manual interval and even the use of all features. In ambon lumut and morosebo bananas, IG-Sturgess provided a higher accuracy compared to the IG manual interval, whereas in hijau, kepok, and susu bananas, the IG-Scott rule produced a higher level of accuracy than the IG manual interval. Only IG optimization with raja bananas resulted in lower accuracy compared to the IG manual interval. However, IG-Sturgess and IG-Scott still produced a higher accuracy compared to the use of all features. Thus, it was supposed IG optimization with Sturgess and Scott rules was more effective than IG interval manual. This optimization was successful in reducing the number of features by 25% of the total number of features (20 features).

IG uses the probability of features to the class, so that the greater the number of data frequencies, the higher the possibility of entering a certain class. The histogram showed the intervals by firstly sorting data, so that the data included in same interval has values in sequential order. K-means clustering grouped data based on the proximity of data to the center of the cluster without considering the label class and the order of values. This allowed for clusters containing data with different classes, even though they do not necessarily have almost the same value. For this reason, clustering techniques were not effective in determining the number of IG intervals.

Result of handling imbalanced classes

Feature selection with IG interval manual and optimized IG succeeded in increasing the accuracy by reducing the number of features by up to 25% of the total features. Nevertheless, the results obtained were quite low (on average below 70%) due to the presence of an imbalanced class (Tables 2 and 3). Some maturity levels had only a very small amount of data. One example is kepok banana with natural curing, for which there were only 2 datapoints (rather old harvest at level 4) and 26 datapoints (old harvest at level 2). Therefore, SMOTE was implemented to handle this problem. The accuracy of the test results after performing SMOTE is shown in Fig. 4.

Handling imbalanced data using SMOTE very significantly increased the accuracy and produced almost the same level of accuracy as from using all features, feature selection with IG manual intervals, IG-Sturgess and IG-Scott. However, the IG-K means results had lower accuracy than the imbalanced class.

IG optimization with the Sturgess and Scott rules only provided the best results for ambon lumut, while for kepok only the IG-Sturgess resulted in higher accuracy. For the other bananas, feature selection using the IG manual interval resulted in higher accuracy than for IG optimization. In fact, the use of the susu bananas whole feature resulted in the highest accuracy.

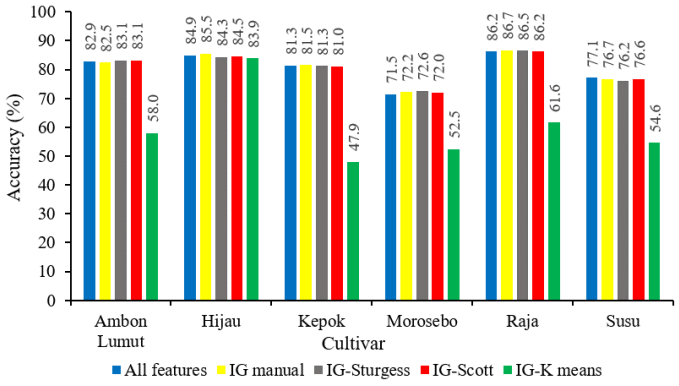


Fig. 4 Performance of maturity identification results using all features and selected features with balanced class

However, the difference in accuracy was not substantial (no more than 0.6%).

Overall, handling the imbalanced classes using SMOTE improved recognition accuracy. The addition of datasets to minority classes produced more data that could be used in training. A substantial increase in accuracy occurred in testing using all features and feature selection with the IG manual interval and IG optimization using the histogram technique (Table 5). However, an anomaly occurred in testing using IG optimization with K-means, where the increase in accuracy only occurred in the hijau banana dataset, with decreases for the other datasets.

A very substantial increase (about 40% or more) occurred in testing with the kepok. Referring to Tables 2 and 3, the kepok banana dataset, especially with natural ripeness, was very unbalanced. The improvement in accuracy in testing using the other banana datasets ranged from 20% to 30%. These datasets had differences in the amount of data at each level below 50%.

There has been little reported research related to the identification of artificial ripeness in bananas. Vaviya et al. (2019) reported 98% accuracy from using a CNN. Although their result was higher than for the current study, they only distinguished two classes (natural and artificial ripeness) without distinguishing the ripeness level. Thie study eliminated datasets containing replicas, noise and redundancy, which would greatly reduce the complexity and the occurrence of misidentification. Furthermore, their study did not mention the cultivars used in the testing. Though each cultivar has different characteristics, testing on various banana cultivars is essential. In addition, they did not mentioned the maturity of the fruit. Naturally and artificially ripened bananas that were harvested in the middle of fourth quarter were more easily distinguished. Based on the observations of the current result testing, most of the misidentification that occurred was from not being able to distinguish between natural and artificial ripeness, especially for old harvested bananas. This error mainly occurred at levels 2, 3 and 4, where in some datasets, there was almost the same color, shape and surface texture.

Table 5 Improved accuracy after over-sampling using synthetic minority over-sampling technique

Cultivar	Accuracy Improvement (%)				
	All features	IG manual	IG-Sturgess	IG-Scott	IG-K means
Ambon Lumut	29.4	27.5	26.3	28.5	-2.2
Hijau	21.5	22.0	22.7	20.5	26.1
Kepok	48.0	41.2	42.2	40.0	-7.9
Morosebo	26.3	20.3	20.3	21.0	-5.5
Raja	31.2	28.0	30.9	29.3	0.4
Susu	29.9	28.9	28.0	28.6	-1.8

Other studies paid more attention to identifying quality and ripeness levels without distinguishing whether there was natural or artificial ripening. The study conducted by Dewi et al. (2022b) identified the ripeness of susu bananas at three levels with the peel equally evenly yellow. This was almost in line with the current research in terms of having high characteristic similarities. However, the current research had higher accuracy after the problem of imbalanced classes had been addressed.

Mostly, the results of classification using ELM showed that IG optimization with histogram techniques improved the recognition accuracy, even though the obtained accuracy was not as high due to imbalanced data. A substantial increase in accuracy occurred after the handling of imbalanced classes. Nevertheless, the accuracy of IG optimization on some tests with balanced classes was slightly lower compared to the IG manual interval. However, in general, IG optimization based on the Sturgess and Scott rules was optimal because it produced almost similar levels of accuracy, with some even higher than from using the IG manual interval, as well as all features. Further study should investigate using deep learning for both feature selection and the classification of artificial ripeness of banana fruit.

Conflict of Interest

The authors declare that there are no conflicts of interest.

Acknowledgements

The Faculty of Computer Science, Universitas Brawijaya funded the research. The staff of the Faculty of Computer Science and Biology Department, Faculty of Science, Universitas Brawijaya provided support during data collection and the experiment.

References

Behera, S.K., Rath, A.K., Sethy, P.K. 2021. Maturity status classification of papaya fruits based on machine learning and transfer learning approach. Inf. Process. Agric. 8: 244–250. doi.org/10.1016/j.inpa.2020.05.003

BPS. 2022. Production of fruit plants 2021. https://www.bps.go.id/indicator/55/62/1/ produksi-tanaman-buah-buahan.html, 23 January 2023. [in Indonesian]

Chawla, N.V, Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P. 2002. SMOTE : Synthetic minority over-sampling technique. J. Artif. Intell. Res. 16: 321–357. doi.org/10.1613/jair.953

Chorowski, J., Wang, J., Zurada, J.M. 2014. Review and performance comparison of SVM- and ELM-based classifiers. Neurocomputing 128: 507–516. doi.org/10.1016/j.neucom.2013.08.009

De Langhe, E., Vrydaghs, L., De Maret, P., Perrier, X., Denham, T. 2009. Why bananas matter: An introduction to the history of banana domestication. Ethnobot. Res. Appl. 7: 165–178.

Dewi, C., Arisoesilaningsih, E., Mahmudy, W.F., Solimun. 2022a. Identifying of unripe Ambon and Hijau banana fruits using computer vision and extreme learning machine classifier. IOP Conf. Ser. Earth Environ. Sci. 951: 012031. doi: 10.1088/1755-1315/951/1/012031

Dewi, C., Arisoesilaningsih, E., Mahmudy, W.F., Solimun. 2022b. Performance of information gain and PCA feature selection for determining ripen susu banana fruits. 2022 IEEE International Conference on Cybernetics and Computational Intelligence. Malang, Indonesia, pp. 323–328. doi: 10.1109/CyberneticsCom55287.2022.9865623

Dewi, C., Mahmudy, W.F., Arifando, R., Arbawa, Y.K., Ahmadie, B.L. 2020. Improve performance of extreme learning machine in classification of patchouli varieties with imbalanced class. In: Proceedings of the 5th International Conference on Sustainable Information Engineering and Technology. Malang, Indonesia, pp. 16–22. doi.org/10.1145/3427423.3427424

Dewi, C., Mahmudy, W.F., Solimun, Arisoesilaningsih, E. 2022c. Automatic differentiating of postharvest banana fruits with high traits using imagery data. AGRIVITA Journal of Agricultural Science 44: 276–289. doi.org/10.17503/agrivita.v44i2.3648

Di Franco, A.2019. Information-gain computation in the FIFTH system. Int. J. Approx. Reason. 105: 386–395. doi.org/10.1016/j.ijar.2018.11.013

Ding, S., Xu, X., Nie, R. 2014. Extreme learning machine and its applications. Neural. Comput. Applic. 25: 549–556. doi.org/10.1007/s00521-013-1522-8

Elmaizi, A., Nhaila, H., Sarhrouni, E., Hammouch, A., Nacir, C. 2019. A novel information gain based approach for classification and dimensionality reduction of hyperspectral images. Procedia Comput. Sci. 148: 126–134. doi.org/10.1016/j.procs.2019.01.016

Elreedy, D., Atiya, A.F. 2019. A comprehensive analysis of synthetic minority oversampling technique (SMOTE) for handling class imbalance. Inf. Sci. 505: 32–64. doi.org/10.1016/j.ins.2019.07.070

Govender, P., Sivakumar, V. 2020. Application of *k*-means and hierarchical clustering techniques for analysis of air pollution: A review (1980–2019). Atmos. Pollut. Res. 11: 40–56. doi.org/10.1016/j.apr.2019.09.009

Haralick, R.M., Shanmugam, K., Dinstein, I. 1973. Textural features for image classification. IEEE Transactions on Systems, Man and Cybernetics SMC-3: 610–621. doi: 10.1109/TSMC.1973.4309314

Harel, B., Parmet, Y., Edan, Y. 2020. Maturity classification of sweet peppers using image datasets acquired in different times. Comput. Ind. 121: 103274. doi.org/10.1016/j.compind.2020.103274

Huang, G.-B., Zhu, Q.-Y., Siew, C.-K. 2006. Extreme learning machine: Theory and applications. Neurocomputing 70: 489–501. doi.org/10.1016/j.neucom.2005.12.126

Hyndman, R.J. 1995. The problem with Sturges'rule for constructing histograms. Unpublished note. robjhyndman.com, 12 January 2023.

Jadhav, S., He, H., Jenkins, K. 2018. Information gain directed genetic algorithm wrapper feature selection for credit rating. Appl. Soft Comput. 69: 541–553. doi.org/10.1016/j.asoc.2018.04.033

Jain, A.K. 2010. Data clustering: 50 years beyond K-means. Pattern Recogn. Lett. 31: 651–666. doi.org/10.1016/j.patrec.2--009.09.011

Jeatrakul, P., Wong, K.W., Fung, C.C. 2010. Classification of imbalanced data by combining the complementary neural network and SMOTE algorithm. In: Wong, K.W., Mendis, B.S.U., Bouzerdoum, A. (Eds.). Neural Information Processing. Models and Applications. ICONIP 2010. Lecture Notes in Computer Science, Vol. 6444. Springer. Berlin, Heidelberg.

Kipli, K., Zen, H., Sawawi, M., et al. 2018. Image processing mobile application for banana ripeness evaluation. 2018 International Conference on Computational Approach in Smart Systems Design and Applications. Kuching, Malaysia, pp. 1–5. doi: 10.1109/ICASSDA.2018.8477600

Legg, P.A., Rosin, P.L., Marshall, D., Morgan, J.E. 2007. Improving accuracy and efficiency of registration by mutual information using Sturges' histogram rule. In: Zwiggelaar, R., Labrosse, F. (Eds.). Proceedings of Medical Image Understanding and Analysis 2007. Cambridge, UK, pp. 26–30.

Leung, K.S., Lee, K.H., Wang, J.F., et al. Data mining on DNA sequences of hepatitis B virus. IEEE/ACM Transactions on Computational Biology and Bioinformatics 8: 428–440. doi: 10.1109/TCBB.2009.6

Maheswaran, S., Sathesh, S., Priyadharshini, P., Vivek, B. 2018. Identification of artificially ripened fruits using smart phones. 2017 International Conference on Intelligent Computing and Control. Coimbatore, India, pp. 1–6. doi: 10.1109/I2C2.2017.8321857

Mahmudy, W.F., Dewi, C., Arifando, R., Ahmadie, B.L., Rahman, M.A. 2021. Combination of morphology, wavelet, and convex hull features in classification of patchouli varieties with imbalance data using artificial neural network. Journal of Applied Research and Technology 19: 633–643.

Mazen, F.M.A., Nashat, A.A. 2019. Ripeness classification of bananas using an artificial neural network. Arab. J. Sci. Eng. 44: 6901–6910. doi.org/10.1007/s13369-018-03695-5

Pourdarbani, R., Sabzi, S., Kalantari, D., Paliwal, J., Benmouna, B., García-Mateos, G., Molina-Martínez, J.M. 2020. Estimation of different ripening stages of Fuji apples using image processing and spectroscopy based on the majority voting method. Comput. Electron. Agr. 176: 105643. doi.org/10.1016/j.compag.2020.105643

Rahman, M.M., Islam, S.M.T., Biswas, M.A., Bepery, C. 2021. Detection of artificially ripen mango using image processing. In: Prateek, M., Singh, T.P., Choudhury, T., Pandey, H.M., Gia Nhu, N. (Eds.). Proceedings of International Conference on Machine Intelligence and Data Science Applications. Algorithms for Intelligent Systems. Springer. Singapore, pp. 759–769. doi.org/10.1007/978-981-33-4087-9_61

Sabilla, I.A., Wahyuni, C.S., Fatichah, C., Herumurti, D. 2019. Determining banana types and ripeness from image using machine learning methods. 2019 International Conference of Artificial Intelligence and Information Technology. Yogyakarta, Indonesia, pp. 407–412. doi: 10.1109/ICAIIT.2019.8834490

Sanaeifar, A., Bakhshipour, A., De La Guardia, M. 2016. Prediction of banana quality indices from color features using support vector regression. Talanta 148: 54–61. doi.org/10.1016/j.talanta.2015.10.073

Serrano, E.P., Rolle, R. 2018. Post-Harvest Management of Banana for Quality and Safety Assurance. Food and Argriculture Organization of the United Nations. Rome, Italy.

Shaltout, N., Moustafa, M., Rafea, A., Moustafa, A., El Hefnawi, M. 2015. Comparing PCA to information gain as a feature selection method for Influenza-A classification. 2015 International Conference on Intelligent Informatics and Biomedical Sciences. Okinawa, Japan, pp. 279–283. doi: 10.1109/ICIIBMS.2015.7439550

Singer, G., Anuar, R., Ben-Gal, I. 2020. A weighted information-gain measure for ordinal classification trees. Expert Syst. Appl. 152: 113375. doi.org/10.1016/j.eswa.2020.113375

Sun, X., Xu, J., Jiang, C., Feng, J., Chen, S.-S., He, F. 2016. Extreme learning machine for multi-label classification. 18: 225. doi.org/10.3390/e18060225

Tang, L., Sun, K., Liu, L., Wang, G., Liu, Y. 2019. A reduced-reference quality assessment metric for super-resolution reconstructed images with information gain and texture similarity. Signal Process. Image 79: 32–39. doi.org/10.1016/j.image.2019.08.004

Vaviya, H., Yadav, A., Vishwakarma, V., Shah, N. 2019. Identification of artificially ripened fruits using machine learning. 2nd International Conference on Advances in Science & Technology. Mumbai, India.

Xie, J., Qiu, Z. 2007. The effect of imbalanced data sets on LDA: A theoretical and empirical analysis. Pattern Recogn. 40: 557–562. doi.org/10.1016/j.patcog.2006.01.009

Yao, Q., Guan, Z., Zhou, Y., Tang, J., Hu, Y., Yang, B. 2009. Application of support vector machine for detecting rice diseases using shape and color texture features. 2009 International Conference on Engineering Computation. Hong Kong, China, pp. 79–83. doi: 10.1109/ICEC.2009.73