Building of Informatics, Technology and Science (BITS)
Volume 5, No 2, September 2023 Page: 576−586
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v5i2.4104

Classification of Fruits High in Vitamin C Using Self-Organizing Map and the K-Means Clustering Algorithm

Nuke L Chusna[1,*], Nurhasan Nugroho[2], Umbar Riyanto[3], Ahmad Ari Aldino[4]

[1] Informatics Engineering Study Program, Faculty of Engineering, Universitas Krisnadwipayana, Bekasi, Indonesia
[2] Computer Science Study Program, Faculty of Computer Science, Bina Bangsa University, Banten, Indonesia
[3] Informatics Engineering Study Program, Faculty of Engineering, Universitas Muhammadiyah Tangerang, Indonesia
[4] Centre for Learning Analytics, Monash University, Australia
Email: [1,*]nukelchusna@unkris.ac.id, [2]nurhasan.nugroho@binabangsa.ac.id, [3] umbar@ft-umt.ac.id, [4]ahmad.aldino@monash.edu
Correspondence Author Email: nukelchusna@unkris.ac.id
Submitted: **17/08/2023**; Accepted: **30/09/2023**; Published**: 30/09/2023**

Abstract−Vitamin C-rich fruits not only taste fresh and delicious but also have the potential to increase the body's resistance to various diseases and maintain a proper nutritional balance. Information about fruits high in vitamin C is very important in order to increase public knowledge about which fruits contain high levels of vitamin C. However, to classify fruits high in vitamin C based on their image, a model is needed that is able to analyze the characteristics present in the image of the fruit. The purpose of this study is to build a classification model for high-vitamin C fruits with a combination of the Self-Organizing Map (SOM) artificial neural network algorithm and K-Means Clustering. Prior to classification, an image segmentation process is carried out using the K-Means Clustering algorithm, which will separate the image into parts that have similar visual characteristics. After the segmented image, the features of the object are extracted based on shape and texture. After the features of the image have been obtained, proceed with classifying images using the SOM algorithm by mapping multidimensional data into a lower-dimensional spatial representation to obtain the appropriate group or class. The accuracy test results for the built model produce an accuracy value of 93.33% and are included in the good category.

Keywords: Artificial Neural Networks; Fruits High in Vitamin C; K-Means Clustering; Self-Organizing Map; Shape Features;

1. INTRODUCTION

The fruit is the part of the plant that develops from the flower and contains the seeds. Fruits come in different shapes, sizes, and tastes depending on the type of plant. Each type of fruit has different nutritional content but generally contains vitamins, minerals, fiber, water, and bioactive substances that provide health benefits [1]. Of the several ingredients present in fruit, one of the vitamins that is important and needs to be considered by someone who consumes fruit is vitamin C. Vitamin C, also known as ascorbic acid, is an essential nutrient that has an important role in maintaining a strong immune system and promoting good health. good skin and acts as an antioxidant that protects the body from damage caused by free radicals [2]. Fruits rich in vitamin C not only provide a fresh and delicious taste but also have the potential to increase the body's resistance to various diseases and maintain a proper nutritional balance [3]. Each fruit has different contents, but not all fruits contain high levels of vitamin C. Fruits that are included in the type of fruit that contains vitamin C are those that contain vitamin C above 50 milligrams in 100 grams of fruit [4]. Information about fruit that has a high vitamin C content is very important for someone who needs this vitamin for their health. Images of fruit are a valuable source of visual information, especially in identifying and selecting fruits that are rich in certain nutrients, such as vitamin C. However, to identify fruits high in vitamin C based on their images, a model is needed that is able to analyze the characteristics contained in the image of the fruit. For this reason, the application of digital image processing can be a solution to help people identify fruits high in vitamin C based on their images.

Digital image processing is a series of techniques and methods used to manipulate digital images for a specific purpose [5]. It involves a series of steps designed to improve image quality, remove noise, enhance important features, and extract useful information from visual data [6]. One of the tasks of image processing is image classification, in which this process involves grouping images into certain categories or classes based on the characteristics possessed by the image. This involves using computational algorithms to differentiate images into related groups based on certain patterns, features, or attributes extracted from the images [7]. There are several previous studies related to the classification or identification of fruit images. The first research is about the classification of pear images using the K-Nearest Neighbor (K-NN) algorithm [8]. In this study, it was able to produce an accuracy rate of up to 87.5%. The K-NN algorithm has a simple concept where grouping is done based on the majority of labels from the nearest neighbors in the training data. However, this algorithm has the disadvantage that it is vulnerable to data that is not standardized and contains noisy values [9]. Subsequent research on fruit classification using the Multiclass Support Vector Machine (Multiclass SVM) approach [10]. The model developed in this study produces an accuracy value of 87.06%. The Multiclass SVM algorithm can address linearly non-separable data in the original feature space. However, this algorithm is sensitive to outlier data, which can affect the formation of boundaries in class grouping. [11]. The next previous research was regarding the identification of bananas based on their level of ripeness using the Extreme Learning Machine (ELM) artificial neural network [12]. In this study, the accuracy rate was 93%. The use of artificial neural networks is considered effective in this case because artificial neural networks are based on artificial neurons that are connected in layers and are able to process information by going through non-linear computation

Building of Informatics, Technology and Science (BITS)
Volume 5, No 2, September 2023 Page: 576−586
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v5i2.4104

stages. However, artificial neural networks using the ELM algorithm use one hidden layer and generate weights between the input and hidden layers randomly, causing overfitting if the determination is not correct.

Based on previous research, the difference with the research that will be carried out is that in this study, the focus is on the classification of high-vitamin C fruit images by implementing a classification model using Self-Organizing Map (SOM) artificial neural networks with K-Means Clustering used for the process of image segmentation. Self-Organizing Map (SOM), also known as Kohonen Map, are a type of unsupervised learning algorithm used to map multidimensional data into a lower-dimensional (usually two-dimensional) representation of space. [13]. The SOM algorithm has the ability to reduce high data dimensions into lower dimensions [14]. This algorithm can group similar data into contiguous groups, which helps in identifying patterns, relationships, and structures in the data [15]. However, in image processing before the object is classified, information is needed regarding the characteristics of the object or feature extraction. The feature extraction process can work well if the object needed can be separated from the background. For this reason, the segmentation process is very important in image classification. Separation between foreground and background in an image can use clustering techniques. So, in this research the K-Means Clustering algorithm was used in the segmentation process. The reason is, this algorithm can divide pixels in an image into several groups or segments based on similarity in color or other characteristics. The use of the SOM and K-Means Clustering algorithms in this case is applied in order to obtain a model that has good accuracy in classifying images of fruit that contain high vitamin C.

As explained in the previous explanation, the aim of this research is to build a classification model for fruits high in vitamin C using a combination of the Self-Organizing Map (SOM) and K-Means Clustering algorithms. Before classification is carried out, an image segmentation process is carried out; this is necessary to improve the quality of feature extraction. So, in the image segmentation process, the K-means clustering algorithm is used, which functions to separate the image into parts that have similar visual characteristics. After the image is segmented, the characteristics of the object are extracted based on shape and texture. Shape features are assessed based on metric and eccentricity calculations, while texture features are assessed based on the Gray Level Co-Occurrence Matrix (GLCM) approach. After the image characteristics are obtained, proceed with classifying the image using the SOM algorithm by mapping multidimensional data into a lower-dimensional space representation to obtain the appropriate groups or classes.

2. RESEARCH METHODOLOGY

2.1 Research Stages

Research stages refer to the sequence of steps or phases followed in planning, collecting data, analyzing, and interpreting information in a study. This stage helps direct the course of the research, ensure the right methodology, and achieve the research objectives [16]. The phases applied to this study are presented in Figure 1.

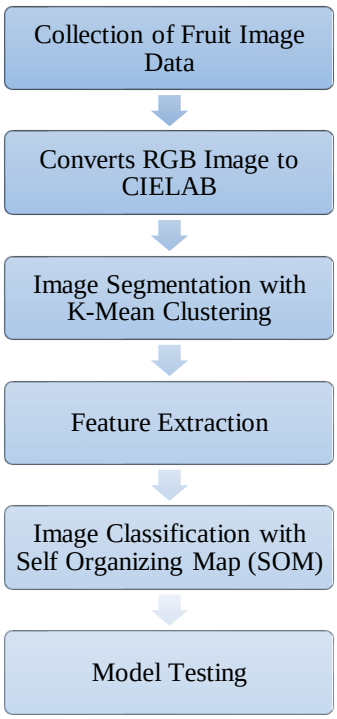


Figure 1. Research Stages

Based on Figure 1, the phases of the research carried out in detail can be explained as follows:

Building of Informatics, Technology and Science (BITS)
Volume 5, No 2, September 2023 Page: 576−586
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v5i2.4104

a. Collection of Fruit Image Data

In this phase, images of fruit data will be collected, which will be used as a dataset. A dataset refers to a collection of data collected or compiled for the purposes of analysis, processing, or research [17]. Dataset quality is very important to ensure accurate and reliable analytical results [18]. The fruit used in this study was based on the top 5 fruits rich in vitamin C taken from the UPaae article [4], these types of fruit include: Guava, Kiwifruit, Papaya, Strawberries and Orange. Images obtained from the internet, for a total of 500 images. All data is then divided into training data and testing data with a percentage of 70% and 30%, respectively. This is because small amounts of data can be used to train the model and make learning patterns more relevant [19]. So, the amount of training data used is 350 images and the test data is 150 images.

b. Converts RGB Image to CIELAB

Image conversion from the RGB (Red-Green-Blue) color space to CIELAB is the process of changing the image color representation from the commonly used RGB color system to the CIELAB color system, which is more intuitive and based on human perception of color [20]. The CIELAB color space (L*a*b) was developed based on parameters such as brightness (L*), red-green tone (a*), and yellow-blue tone (b*), which better suit human perception of color [21]. Before RGB is converted to L* a* b*, it is converted first into the XYZ color space, where the colors are represented as values that are always positive. Calculation of the transformation of the RGB color space to XYZ through the calculation of the transformation matrix using equations (1), (2), and (3).

$$[X] = [0.412453 \quad 0.357580 \quad 0.180423][R] \tag{1}$$

$$[Y] = [0.212671 \quad 0.715160 \quad 0.715160][G] \tag{2}$$

$$[Z] = [0.019334 \quad 0.119193 \quad 0.950227][B] \tag{3}$$

Then the results of the XYZ values are converted to the L* a* b* color space with equations (4), (5), (6) and (7).

$$L^* = 116\left(\frac{Y}{Y_n}\right)\frac{1}{3}16 \; for \; \frac{Y}{Y_n} > 0.008856 \tag{4}$$

$$L^* = 903.3 \; \frac{Y}{Y_n} \; otherwise \tag{5}$$

$$a^* = 500\left(f\left(\frac{X}{X_n}\right)\right) - f\left(\frac{Y}{Y_n}\right) \tag{6}$$

$$b^* = 200\left(f\left(\frac{Y}{Y_n}\right)\right) - f\left(\frac{Z}{Z_n}\right) \tag{7}$$

c. Image Segmentation with K-Mean Clustering

The purpose of image segmentation is to facilitate further analysis, object identification, or other applications by separating different areas in an image into distinct groups [22]. The image segmentation technique used is the K-Means Clustering algorithm. The K-Means Clustering algorithm performs the process of dividing the image into groups or segments based on similarity in color intensity or other features [23]. This is done so that the image can be divided between background and foreground so that it can focus on objects that contain information.

d. Feature Extraction

Feature extraction is the process of transforming complex data or information into simpler and informative representations that can be used for further analysis or data processing tasks [24]. In this study, shape and texture feature extraction was used, where shape features were assessed based on metric and eccentricity calculations, while texture features were assessed based on the Gray Level Co-Occurrence Matrix (GLCM) approach.

e. Image Classification with Self Organizing Map (SOM)

Self-Organizing Map (SOM) is a type of unsupervised learning algorithm used for mapping multidimensional data into a lower-dimensional (usually two-dimensional) spatial representation [13]. The SOM algorithm serves to stimulate topological structures hidden in data, similar to how neurons in the human brain interact [14]. The SOM algorithm can group or divide classes based on the same features through a pattern-learning process.

f. Model Testing

At the testing stage, it is carried out by testing the accuracy of the model that has been built. The main purpose of accuracy testing is to measure the extent to which a model or system is reliable in performing certain tasks and to assess its quality [25]. To measure accuracy using equation (8).

$$Accuracy = \frac{NP}{TP} \times 100\% \tag{8}$$

where NP is the number of correct predictions and TP represents the total number of predictions.

2.2 Feature Extraction on Shape and Texture

Building of Informatics, Technology and Science (BITS)
Volume 5, No 2, September 2023 Page: 576−586
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v5i2.4104

Feature extraction is the process of transforming complex data into a simpler and more informative representation [24]. In this study, two feature extractions were used, namely shape feature extraction and texture feature extraction. Shape feature extraction involves identifying and describing the geometric characteristics of objects in an image. Two parameters that are often used in shape feature extraction are the matric and eccentricity parameters [26]. By comparing the length of the contour with the square root of the area that the contour shades, the matrix parameter assesses the complexity of an object's shape. Meanwhile, eccentricity measures the extent to which an object approaches a circle shape. The eccentricity value is close to 0 if the object is a perfect circle and closer to 1 if the object is elongated. With this value, we can recognize whether the object has a round or more oval shape. In order to obtain values for the matric and eccentricity parameters, equations (9) and (10) can be used.

$$M = \frac{4\pi \times A}{c} \tag{9}$$

$$e = \sqrt{1 - \frac{b^2}{a^2}} \tag{10}$$

where *a* denotes the minor axis and *b* denotes the major axis. For the notation, *A* shows the area and *C* shows the circumference.

As for texture feature extraction, the Gray-Level Co-occurrence Matrix (GLCM) approach is used. GLCM forms a matrix that represents the frequency of occurrence of adjacent pixel intensity value pairs in a certain direction and distance [27]. It provides insight into image texture by identifying patterns such as coarse, fine, grainy, or grainy. In the GLCM approach, there are several texture feature parameters that can be generated. These parameters include:

a. Contrast

Contrast is a measure of the intensity variation between pairs of pixel values. The value of the contrast can be obtained through equation (11).

$$Contrast = \sum_i \sum_j (i-j)^2 pd(i,j) \tag{11}$$

b. Correlation

Correlation is related to the size of the dependency between neighboring pixel intensities. In order to obtain a correlation value, you can calculate it using equation (12).

$$Correlation = \sum_i \sum_j \frac{(i-\mu_i)(j-\mu_j)p(i,j)}{\sigma_i \sigma_j} \tag{12}$$

c. Energy

Energy shows a measure of the clarity of patterns in an image. The energy value is obtained by equation (13).

$$Energy = \sum_i \sum_j p_2^d(i,j) \tag{13}$$

d. Homogeneity

Homogeneity aims to measure how uniform the distribution of pixel intensity values is. To get the homogeneity value of an image, it can be calculated using equation (14).

$$Homogenity = \sum_i \sum_j \frac{pd(i,j)}{i + |i-j|} \tag{14}$$

2.3 K-Means Clustering

K-Means Clustering is a grouping or clustering algorithm in data analysis that aims to divide a set of data into homogeneous groups based on the similarity of certain features or attributes [23]. This algorithm can also be used for image segmentation, namely separating the image into parts that have similar visual characteristics. The K-Means Clustering algorithm works by grouping data points into k clusters based on their distance to the cluster center (centroid) [28]. The data points that have been formed are then grouped around the cluster center (centroid) by reducing the objective function obtained by equation (15).

$$\overline{V_{ij}} = \frac{1}{N} \sum_{k=0}^{n_i} x_{kj} \tag{15}$$

where V_{ij} denotes the cluster center or cluster average *i* on attribute j. Then, n_i denotes the number of cluster members. Meanwhile, for x_{kj}, it shows the data value k in the cluster on attribute *j*.

Building of Informatics, Technology and Science (BITS)
Volume 5, No 2, September 2023 Page: 576−586
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v5i2.4104

2.4 Self-Organizing Map (SOM) Algorithm

Algorithmic Self-Organizing Map (SOM), also known as Kohonen Map, are a type of unsupervised learning algorithm used to map multidimensional data into a lower-dimensional (usually two-dimensional) representation of space called SOM maps [29]. Teuvo Kohonen created this algorithm in 1982, and it is well-known for its use in data analysis, grouping, data visualization, and dimension reduction. The basic principle of the SOM algorithm is to stimulate the topological structure of the data hidden in the original data [30]. It does this by creating artificial neuron networks similar to the human brain, where each neuron represents a location on a map. The architecture of the SOM algorithm is visualized in Figure 2.

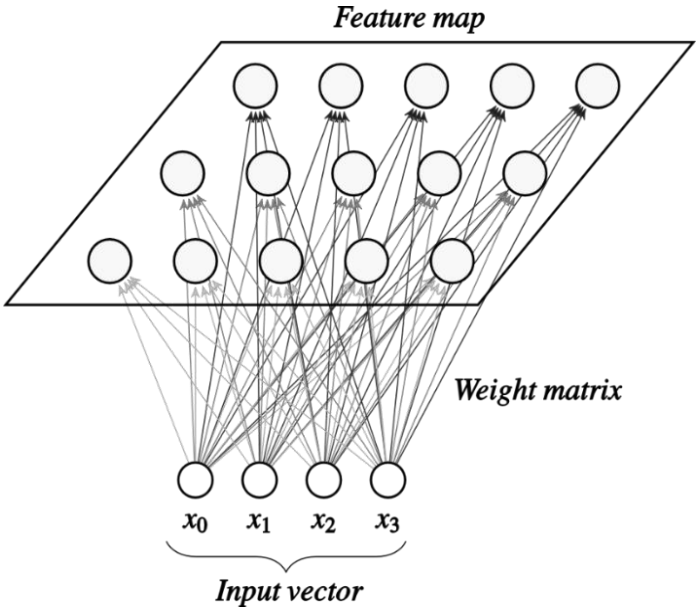


Figure 2. Architecture of Self-Organizing Map (SOM) Artificial Neural Networks

In Figure 2, it can be seen that in map formation each neuron gets a random weight and each neuron has a weight vector that has the same dimensions as the input data. Then the neuron with the closest weight to the data is considered the winner. In addition to reducing data, SOM also divides data based on groups that have similar data, so that it can be said that SOM solves 2 problems, namely reducing data dimensions and displaying data similarities. The following is the training and classification process using the SOM algorithm:

a. Each input is calculated by calculating the distance to each neuron on the map. The distance calculation is done by equation (12).

$$J_{(x,w_m)} = \sum_{i=1}^{n} (x_i - w_{mi})^2 \quad (12)$$

where x is the input data and w_m are the neurons in m in the map.

b. The smallest distance in calculating each data becomes the winning neuron, where the winning neuron will be evaluated for its weight by equation (13).

$$w_{new} = w_{old} + lr \times \theta \times (x - w_{old}) \quad (13)$$

where w_{new} is the new weight that will replace w_{old} which is the old weight on the winning neuron with lr is the learning rate, θ is the neighbor width and x is the training data.

c. After steps 1 and 2 are carried out on each piece of data, the learning rate decreases as it iterates with equation (14).

$$\alpha(t) = \alpha_i \left(1 - \frac{t}{t_{max}}\right) \quad (14)$$

where $\alpha(t)$ is the learning rate at each iteration (t) with αi is the initial learning rate.

d. Besides that, the neighbor width (θ) is changed using equation (15).

$$\theta(t) = \theta_i \left(\frac{\theta_f}{\theta_i}\right)^{\frac{t}{t_{max}}} \quad (15)$$

Where $\theta(t)$ is the width of the neighbors which will decrease with iteration (t). Where θi is the initial value of the neighbor's width and θf is the final value of the neighbor.

Building of Informatics, Technology and Science (BITS)
Volume 5, No 2, September 2023 Page: 576−586
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v5i2.4104

3. RESULT AND DISCUSSION

3.1 Preparing Models for Training

The completion of the classification of high-vitamin C fruits through a combination of the Self-Organizing Map (SOM) algorithm and K-Means Clustering begins with building a model as training. This training model is used to obtain learning patterns for grouping and mapping data into groups that represent different classes or categories. This is used so that the developed algorithm can assist in organizing and classifying data into appropriate groups. To carry out training, it requires a dataset to get the desired patterns. The fruit used in this study was based on the top 5 fruits rich in vitamin C taken from the UPaae article; the types of fruit include: Guava, Kiwifruit, Papaya, Strawberries, and Orange. Images obtained from the internet, for a total of 500 images. All data is then divided into training data and testing data with a percentage of 70% and 30%, respectively. So, the amount of training data used is 350 images. The training and testing processes are carried out using MATLAB software. The training process is carried out using MATLAB software. The training function is to understand and extract patterns or relationships in the training data, so that the model can be used to carry out its tasks.

3.2 Converts RGB Image to CIELAB

The training model that has been created begins with the image conversion process from the RGB (Red-Green-Blue) color space to the CIELAB (L* a* b*) color space. This is done so that the color representation of CIELAB images is more intuitive and based on human perception of color. The CIELAB color model is designed to more closely approximate the human perception of color. The results of converting RGB colors to the CIELAB color space (L* a* b*) are presented in Figure 3.

(a) (b)

Figure 3. (d) RGB Image and (b) CIELAB (L*a*b) Color Space Conversion Results

Figure 3 (b) shows the transformation results from an RGB image to a CIELAB image (L*a*b).

3.3 Image Segmentation with K-Mean Clustering

The next step is to perform image segmentation using K-Mean Clustering. The K-Means Clustering algorithm is used to separate an image into parts that have similar visual characteristics. This algorithm will separate the images into groups based on color similarity, which can assist in further analysis. The way the K-Means Clustering algorithm works is by creating groups of points based on their distance to the center of the cluster, or centroid. An illustration of how the K-Means Clustering algorithm works is presented in Figure 4.

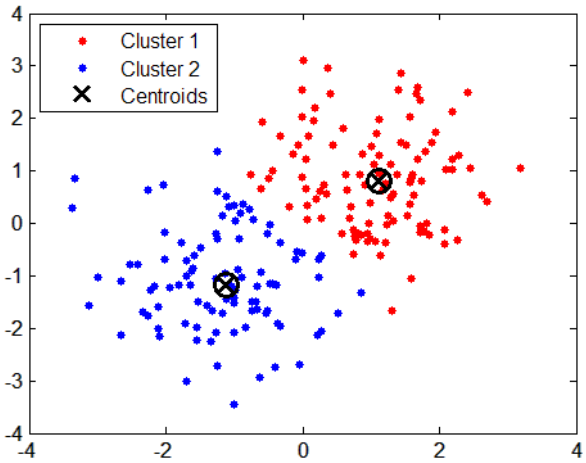
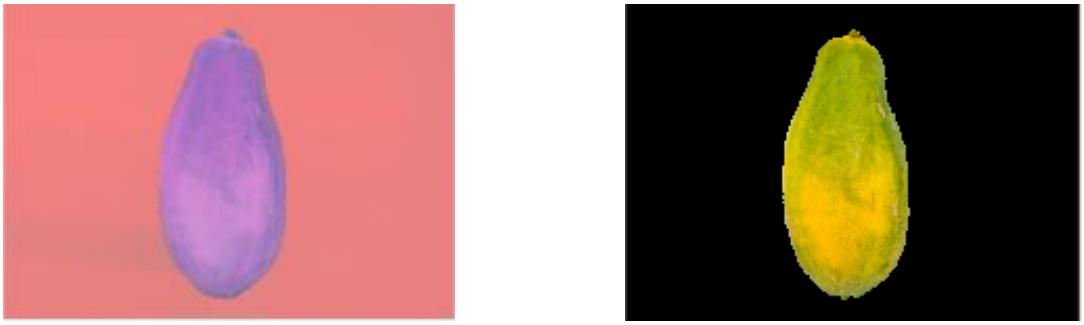


Figure 4. Illustration of the K-Means Clustering Algorithm

As can be seen in Figure 4, the K-Means Clustering algorithm groups data based on proximity to the centroid, where in the image segmentation process it will be separated into two parts, namely the background and foreground

Building of Informatics, Technology and Science (BITS)
Volume 5, No 2, September 2023 Page: 576−586
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v5i2.4104

images. This process is carried out in order to get a foreground object so that it can focus on objects that contain only information. The results of image segmentation using the K-Means Clustering algorithm can be seen in Figure 5.



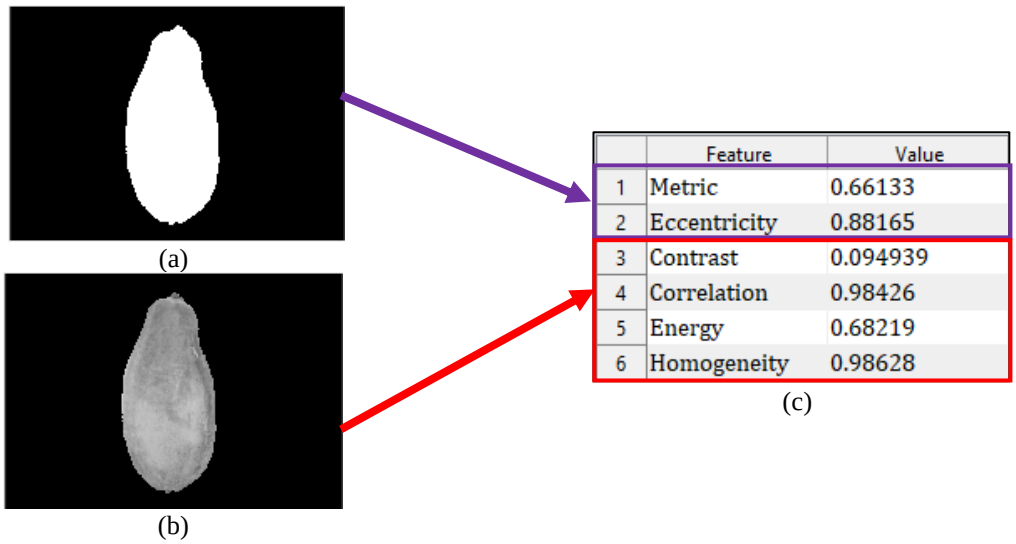
(a) (b)

Figure 5. (a) CIELAB Color Space Image (L*a*b) and (b) Segmented Image Results

The results of the segmented image are shown in Figure 5 (b), where the foreground in the image is separated from the background so that an object containing information is obtained. The K-Means Clustering algorithm groups data into clusters based on the centroid.

3.4 Feature Extraction

After image segmentation is carried out, the next process is to carry out feature extraction to obtain the information contained in the foreground image. In the case study, shape and texture features were used to classify fruit containing high levels of vitamin C. This is done because one fruit can be distinguished from another based on its shape and texture. The parameters used in shape feature extraction are matric and eccentricity parameters. The matric value is obtained based on the comparison of the length of the contour with the square root of the area covered by the contour. The matric parameter assesses the complexity of the shape of an object. The eccentricity value measures the degree to which an object approaches a circular shape. Meanwhile, for texture extraction, the Gray-Level Co-occurrence Matrix (GLCM) approach is used with parameters including contrast, correlation, energy, and homogeneity. To obtain shape characteristics, the segmented image will be converted into a binary image so that information about the shape of the object can be obtained. Then, for the texture features, the segmentation image is converted into a grayscale image to obtain parameter values for contrast, correlation, energy, and homogeneity. Examples of images resulting from feature extraction and the values of each parameter obtained from shape and texture features are presented in Figure 6.



(a)

(b)

	Feature	Value
1	Metric	0.66133
2	Eccentricity	0.88165
3	Contrast	0.094939
4	Correlation	0.98426
5	Energy	0.68219
6	Homogeneity	0.98628

(c)

Figure 6. (a) Binary Image; (b) Grayscale Image and (c) Feature Extraction Results

Figure 6 (a) is a segmentation image converted into a binary image so that information can be obtained about the shape of the object being classified. Figure 6 (b) is the result of grayscale image conversion so that information can be obtained regarding the texture of the object being classified. Meanwhile, Figure 6 (c) shows the shape and texture feature extraction values obtained in the image.

3.5 Image Classification with Self Organizing Map (SOM)

At the classification stage, the Self-Organizing Map (SOM) algorithm is used, where this algorithm is used to map high-dimensional data into a two-dimensional map (or more) that represents the data structure. As an illustration, there

Building of Informatics, Technology and Science (BITS)
Volume 5, No 2, September 2023 Page: 576−586
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v5i2.4104

are 4 features, and they are mapped into 8 nodes (neurons), thus forming a 24 grid. The structure can be seen in Figure 7.

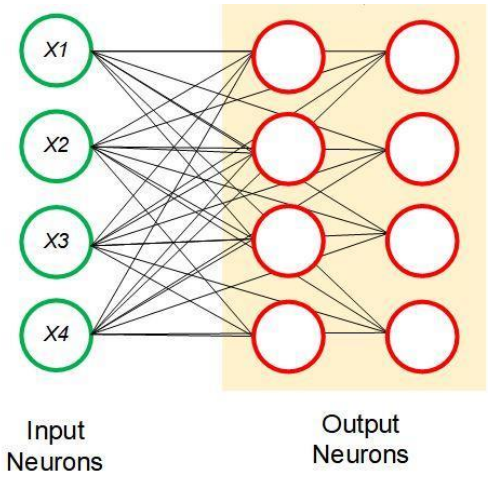


Figure 7. Illustration of the Structure of the Self-Organizing Map (SOM) Algorithm

Figure 7 shows that there are 4 features, so there are 4 dimensions, where each dimension represents its own feature. The number of rows in this tabular dataset represents the number of items. The SOM algorithm reduces many dimensions to only two. The 2-dimensional representation is located in the output neurons (nodes), where the 8 nodes will be mapped. So that the 8 output nodes will be paired into a 2-dimensional field, each input neuron is connected to the selected node (the node that has the shortest distance). The classification model with the SOM algorithm was first built to be used as training. After the training process is carried out, which produces a map with the last weight, the testing process is carried out. The last weight on the map is used as a classification model. The testing process is the same as the training process, but only involves calculating the distance to each neuron on the map. After the winning neuron is obtained, the neuron will become the class for the input data. The developed SOM model is then implemented in MATLAB software, which is used for training. The architecture used in the training uses 350 input images of training data, 5 classes of high-vitamin C fruits, 4 layers, and 200 iterations. Furthermore, the model is used for testing and is built and implemented in MATLAB software, as shown in Figure 8.

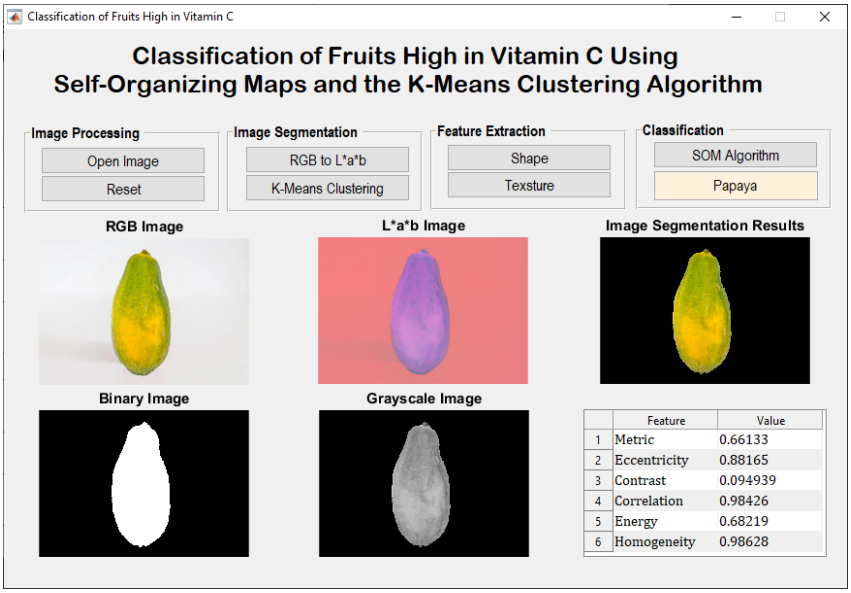


Figure 8. Interface of Classification System for High Vitamin C Fruits

Figure 8 shows the interface of the model that has been implemented in the form of a GUI, which has the following features: opening the image, converting RGB to L*a*b, image segmentation using K-Means clustering, extraction of shape and texture features, and image classification using the SOM algorithm. and see the classification results.

3.6 Model Testing and Discussion of Experimental Results

The test data used were 150 images of high-vitamin C fruits. The 150 images were then divided into 5 classes, namely Guava, Kiwifruit, Papaya, Strawberries, and Orange, so that each class would be tested using 30 images. The accuracy

Building of Informatics, Technology and Science (BITS)
Volume 5, No 2, September 2023 Page: 576−586
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v5i2.4104

test process is carried out by matching the classification results produced by the model to the existing facts. The number of correct classifications is then divided by the total test and multiplied by 100% to get the accuracy value. The results of the acuity test for each class are presented in graphical form, which can be seen in Figure 9.

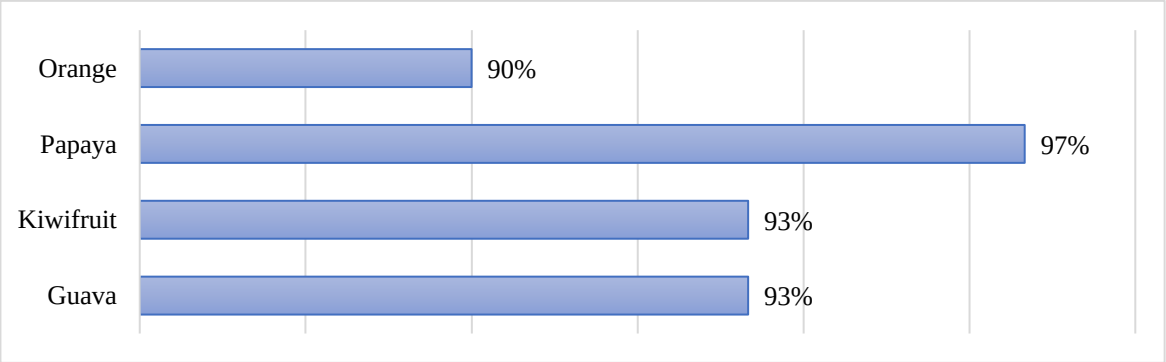


Figure 9. Graph of Accuracy Test Results for Fruits High in Vitamin C

Based on Figure 9, the value of the accuracy test results for each class of fruits high in vitamin C is presented. The Papaya class produces the highest accuracy value at 97%, and the Orange class produces the lowest accuracy value at 90%. The average accuracy value for all classes produces an accuracy value of 93.33%. Then, the results obtained are converted to the category of value criteria with the following guidelines: If the result is between 76% and 100%, it is in the "Good" category; if the result is between 56% and 5%, it is in the enough category; if the result is between 40% and 55%, it is in the category of "Not Very Good", and if the result is below 40%, it is in the category of "Very Bad" [22]. According to these criteria guidelines, the results of the accuracy test obtained fall into the "Good" category. When compared to previous research, the model built produces better accuracy, where in previous studies using the K-Nearest Neighbor (K-NN) approach produced an accuracy rate of 87.5% [8], the Multiclass Support Vector Machine (Multiclass SVM) method obtained an accuracy value of 87.06% [10], and the application of the Extreme Learning Machine (ELM) algorithm obtained an accuracy value of 93% [12]. In the last study, the accuracy was only 0.3%, but in this study, the focus was on identifying the maturity level of bananas. In contrast to the research conducted focusing on fruits high in vitamin C with more classes and data. The research carried out can produce good accuracy because the SOM algorithm includes its ability to perform non-linear mapping, dimension reduction, and grouping of data in an unsupervised manner. These algorithms can help reveal hidden patterns in data and provide a better understanding of the structure and relationships in datasets.

However, it needs to be a concern because the resulting error rate gets a value of 6.67%. There are several factors that affect this. These aspects include the following: 1) The class assigned to each neuron in the SOM algorithm is actually the result of the mapping and clustering performed by the algorithm, so class grouping is very dependent on the parameters set and initial weight initialization; 2) The number of neurons in the SOM algorithm must be determined beforehand, so that if the number of classes or objects in the image is not clear, it is difficult to get optimal results; 3) The K-Means Clustering Algorithm has difficulties getting a foreground that has a variety of backgrounds with several camera angles, making it difficult to process the model, so it needs pre-processing stages; 4) The feature extraction used is only based on the shape and texture; it is necessary to use other features in order to obtain more representative information; 5) The dataset used is still on a small scale, so it is necessary to conduct experiments using larger data in order to obtain a more optimal learning pattern.

4. CONCLUSION

This research has built a model to classify high-vitamin C fruits using a combination of SOM and K-Means Clustering algorithms. Prior to classification, an image segmentation process is carried out using the K-Menas Clustering algorithm, in which this algorithm can differentiate between foreground and background so that it can focus on objects that contain information. After the segmented image of the object obtained, its features are extracted based on shape and texture. Shape features are assessed based on metric and eccentricity calculations, while texture features are assessed based on the Gray Level Co-Occurrence Matrix (GLCM) approach. After the features of the image have been obtained, then proceed with performing image classification using the SOM algorithm through mapping multidimensional data into a lower dimensional representation of space to divide data based on groups that have similar data. The accuracy test results for the built model produce an accuracy value of 93.33% and are included in the good category. However, the research that has been done requires improvement for further research. However, it should also be remembered that the class assigned to each neuron in the SOM algorithm is actually the result of the mapping, so the sensitivity to parameters and sensitivity to initial weight initialization. For this reason, it is necessary to combine it with an algorithm that can properly initialize the weights. In addition, it is necessary to try the application of deep learning so that various features can be resolved and accuracy can be improved. For feature extraction that is used only based on shape and texture, it is necessary to use other features to get more representative information.

Building of Informatics, Technology and Science (BITS)
Volume 5, No 2, September 2023 Page: 576−586
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v5i2.4104

REFERENCES

[1] R. Wildman, *The Nutritionist: Food, Nutrition, and Optimal Health*. Bloomington: Archway Publishing, 2019.
[2] J. G. LeBlanc, *Vitamin C: an Update on Current Uses and Functions*. London: IntechOpen, 2019.
[3] M. Fenech, I. Amaya, V. Valpuesta, and M. A. Botella, "Vitamin C Content in Fruits: Biosynthesis and Regulation," *Front. Plant Sci.*, vol. 9, no. January, pp. 1–21, 2019, doi: 10.3389/fpls.2018.02006.
[4] R. Revealer, "Top 5 Vitamin C Rich Fruits and Health Benefits of Vitamin C," *UPaae*, 2020. https://upaae.com/top-5-vitamin-c-rich-fruits-and-health-benefits-of-vitamin-c/
[5] V. C. S. Rao, S. Venkratamulu, and P. Sammulal, *Digital Image Processing and Applications*. Singapore: Horizon Books (A Division of Ignited Minds Edutech P Ltd), 2021.
[6] R. I. Borman, R. Napianto, N. Nugroho, D. Pasha, Y. Rahmanto, and Y. E. P. Yudoutomo, "Implementation of PCA and KNN Algorithms in the Classification of Indonesian Medicinal Plants," in *ICOMITEE 2021*, 2021, pp. 46–50.
[7] R. I. Borman, F. Rossi, D. Alamsyah, R. Nuraini, and Y. Jusman, "Classification of Medicinal Wild Plants Using Radial Basis Function Neural Network with Least Mean Square," in *International Conference on Electronic and Electrical Engineering and Intelligent System (ICE3IS)*, 2022.
[8] M. A. Hasan and A. S. Budi, "Pears Classification Using Principal Component Analysis and K-Nearest Neighbor," *Sink. J. dan Penelit. Tek. Inform.*, vol. 4, no. 2, pp. 34–41, 2020.
[9] M. C. Untoro, M. Praseptiawan, M. Widianingsih, I. F. Ashari, A. Afriansyah, and O. Oktafianto, "Evaluation of Decision Tree, K-NN, Naive Bayes and SVM with MWMOTE on UCI Dataset," in *ICComSET*, 2019, pp. 1–8. doi: 10.1088/1742-6596/1477/3/032005.
[10] M. Zeeshan, A. Prabhu, C. Arun, and N. S. Rani, "Fruit Classification System Using Multiclass Support Vector Machine Classifier," in *Proceedings of the International Conference on Electronics and Sustainable Communication Systems*, 2020, pp. 289–294.
[11] M. Baldomero-Naranjo, L. I. Martínez-Merino, and A. M. Rodríguez-Chía, "A robust SVM-based approach with feature selection and outliers detection for classification problems," *Expert Syst. Appl.*, vol. 178, p. 115017, 2021, doi: https://doi.org/10.1016/j.eswa.2021.115017.
[12] C. Dewi, E. Arisoesilaningsih, W. F. Mahmudy, and S. Solimun, "Identifying of unripe Ambon and Hijau banana fruits using computer vision and extreme learning machine classifier," in *IOP Conf. Series: Earth and Environmental Science*, 2022, pp. 1–8. doi: 10.1088/1755-1315/951/1/012031.
[13] M. Wong, W. Abeysinghe, and C. Hung, "A Massive Self-Organizing Map For Hyperspectral Image Classification," in *Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing (WHISPERS)*, 2019, pp. 1–5. doi: 10.1109/WHISPERS.2019.8921093.
[14] I. Ahmad, Y. Rahmanto, R. I. Borman, F. Rossi, Y. Jusman, and A. D. Alexander, "Identification of Pineapple Disease Based on Image Using Neural Network Self-Organizing Map (SOM) Model," in *International Conference on Electronic and Electrical Engineering and Intelligent System (ICE3IS)*, 2022.
[15] D. I. Mulyana, A. Saepudin, and M. B. Yel, "Health Detection of Betal Leaves Using Self-Organizing Map and Thresholding Algorithm," *J. Appl. Eng. Technol. Sci.*, vol. 4, no. 1, pp. 180–189, 2022.
[16] I. Ahmad, E. Suwarni, R. I. Borman, A. Asmawati, F. Rossi, and Y. Jusman, "Implementation of RESTful API Web Services Architecture in Takeaway Application Development," in *International Conference on Electronic and Electrical Engineering and Intelligent System (ICE3IS)*, 2022, pp. 132–137. doi: 10.1109/ICE3IS54102.2021.9649679.
[17] A. Mulyanto, E. Susanti, F. Rossi, W. Wajiran, and R. I. Borman, "Penerapan Convolutional Neural Network (CNN) pada Pengenalan Aksara Lampung Berbasis Optical Character Recognition (OCR)," *JEPIN (Jurnal Edukasi dan Penelit. Inform.*, vol. 7, no. 1, pp. 52–57, 2021.
[18] Z. Abidin, R. I. Borman, F. B. Ananda, P. Prasetyawan, F. Rossi, and Y. Jusman, "Classification of Indonesian Traditional Snacks Based on Image Using Convolutional Neural Network (CNN) Algorithm," in *International Conference on Electronic and Electrical Engineering and Intelligent System (ICE3IS)*, 2022, pp. 18–23.
[19] I. O. Muraina, "Ideal Dataset Splitting Ratios in Machine Learning Algorithms: General Concerns for Data Scientists and Data Analysts," in *7th International Mardin Artuklu Scientific Researches Conference*, 2022, pp. 496–505.
[20] J. Pardede, M. G. Husada, A. N. Hermana, and S. A. Rumapea, "Fruit Ripeness Based on RGB , HSV, HSL, L*a*b* Color Feature Using SVM," in *International Conference of Computer Science and Information Technology (ICoSNIKOM)*, 2019.
[21] S.-H. Baek, K.-H. Park, J.-S. Jeon, and T.-Y. Kwak, "Using the CIELAB Color System for Soil Color Identification Based on Digital Image Processing," *J. Korean Geotech. Soc.*, vol. 38, no. 5, pp. 61–71, 2022.
[22] R. I. Borman, Y. Fernando, and Y. Egi Pratama Yudoutomo, "Identification of Vehicle Types Using Learning Vector Quantization Algorithm with Morphological Features," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 6, no. 2, pp. 339–345, 2022, doi: 10.29207/resti.v6i2.3954.
[23] S. Jardim, J. António, and C. Mora, "Graphical Image Region Extraction with K-Means Clustering and Watershed," *J. Imaging*, vol. 8, no. 163, pp. 1–27, 2022.
[24] R. I. Borman, Y. Fernando, and Y. E. P. Yudoutomo, "Identification of Vehicle Types Using Learning Vector Quantization Algorithm with Morphological Features," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 5, no. 158, pp. 339–345, 2022.
[25] R. I. Borman, F. Rossi, Y. Jusman, A. A. A. Rahni, S. D. Putra, and A. Herdiansah, "Identification of Herbal Leaf Types Based on Their Image Using First Order Feature Extraction and Multiclass SVM Algorithm," in *International Conference on Electronic and Electrical Engineering and Intelligent System (ICE3IS)*, 2021, pp. 12–17.
[26] H. Mayatopani, R. I. Borman, W. T. Atmojo, and A. Arisantoso, "Classification of Vehicle Types Using Backpropagation Neural Networks with Metric and Ecentricity Parameters," *J. Ris. Inform.*, vol. 4, no. 1, pp. 65–70, 2021, doi: 10.34288/jri.v4i1.293.
[27] M. H. Santoso, D. A. Larasati, and M. Muhathir, "Wayang Image Classification using MLP Method and GLCM Feature Extraction," *J. Comput. Sci. Inf. Technol. Telecommun. Eng.*, vol. 1, no. 2, pp. 111–120, 2020, doi:

Building of Informatics, Technology and Science (BITS)
Volume 5, No 2, September 2023 Page: 576−586
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v5i2.4104

10.30596/jcositte.v1i2.5131.

[28] S. M. Javidan, A. Banakar, K. A. Vakilian, and Y. Ampatzidis, "Diagnosis of grape leaf diseases using automatic K-means clustering and machine learning," *Smart Agric. Technol.*, vol. 3, pp. 1–14, 2023, doi: 10.1016/j.atech.2022.100081.

[29] A. Banerjee, N. Rakshit, M. Chakrabarty, S. Sinha, S. Ghosh, and S. Ray, "Zooplankton community of Bakreswar reservoir: Assessment and visualization of distribution pattern using self-organizing maps," *Ecol. Inform.*, vol. 72, p. 101837, 2022, doi: https://doi.org/10.1016/j.ecoinf.2022.101837.

[30] J. Zhang, L. Cao, M. Zhang, and W. Fu, "Extracting the Brain-Like Representation by an Improved Self-Organizing Map for Image Classification," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5. doi: 10.1109/ICASSP49357.2023.10095998.