ICIC Express Letters
Part B: Applications
Volume **13**, Number **11**, November **2022**

ICIC International ©2022 ISSN 2185-2766
pp. 1165–1175

CBML: CLASSIFICATION – THAI RED DRAGON FRUIT

Naruwan Yusamran and Nualsawat Hiransakolwong*

Department of Computer Science
School of Science
King Mongkut's Institute of Technology Ladkrabang
Chalongkrung Road, Ladkrabang, Bangkok 10520, Thailand
62605118@kmitl.ac.th; *Corresponding author: nualsawat.hi@kmitl.ac.th

Received March 2022; accepted May 2022

Abstract. *A dragon fruit is nutritious having low sugar levels suitable for diabetics. Its shape is spherical with petals surrounding the bark. In Thailand, 7 species of dragon fruit are grown. Its appearance can be divided into red bark (6 species) or yellow bark (1 species). People can separate between yellow and red bark. Since the red-skinned dragon fruit may have its fruit pulp in one of three possible colors: white, red or pink, it is difficult to guess from the outside what color of the fruit pulp will be. The big problem is that people cannot classify its species of red bark. Some people may be allergic or dislike dragon fruit in some species. Therefore, this research is trying to find the best way to automatically classify the species of Thai red dragon fruit from its image. The CBML stands for content base and Machine Learning. This method uses content base to extract key features (34 attributes), and uses Machine Learning for giving optimization results with the Support Vector Machine (SVM), setting kernel for polynomial in 8 degree. The results showed that the CBML method was able to identify any species of dragon fruits in the red bark group with an accuracy of 98.47%.*
Keywords: Dragon fruit, Classification, Content base, Machine Learning

1. **Introduction.** A dragon fruit is a species of Cactaceae, genus *Hylocercus* spp. and *Selenicereus* sp. [1]. Its bark is yellow or red with petals around the fruit, and at the end of the petals is green. The dragon fruits are divided into 3 groups, which have differences in the shell color (Peel/Skin) and the inner color (fruit pulp), namely Group 1 white pulp with red skin (*Hylocereus undatus*), Group 2 red-pulp-red-skin (*Hylocereus polyrhizus*) or pink-pulp-red-skin (*Hylocereus* spp.), and Group 3 yellow or gold skin-white-pulp (*Hylocereus* sp. and *Selenicereus* sp.) [2].

From visiting the area to explore the cultivation of dragon fruits in the Loei province located at the top of the northeastern region of Thailand, Loei is the area that can harvest the most products in the country [3] with total 7 species of dragon fruits. There are 2 species in Group 1, namely Jumbo White and Vietnamese White, 4 species in Group 2 namely Pink, Siam Red, Taiwan Red, Ruby Red, and 1 species in Group 3 namely Israel Yellow as shown in Figure 1. Dragon fruits in both Group 1 and Group 2 have red bark, with 6 different species. Its morphological features [4] are difficult to collect information because it requires expertise in botany. Therefore, this paper is interested in classifying the species of Thai red dragon fruit using content base and Machine Learning techniques to the specific 6 species of red shell, namely Jumbo White, Vietnamese White, Pink, Siam Red, Taiwan Red and Ruby Red. The correct classification of red dragon fruit species can be very beneficial, for example, making farmers sell them for a better price. A factory can have an automated system for separating and labeling the species. Consumers who are allergic with certain species will be able to avoid it correctly or to choose their favorite.

DOI: 10.24507/icicelb.13.11.1165

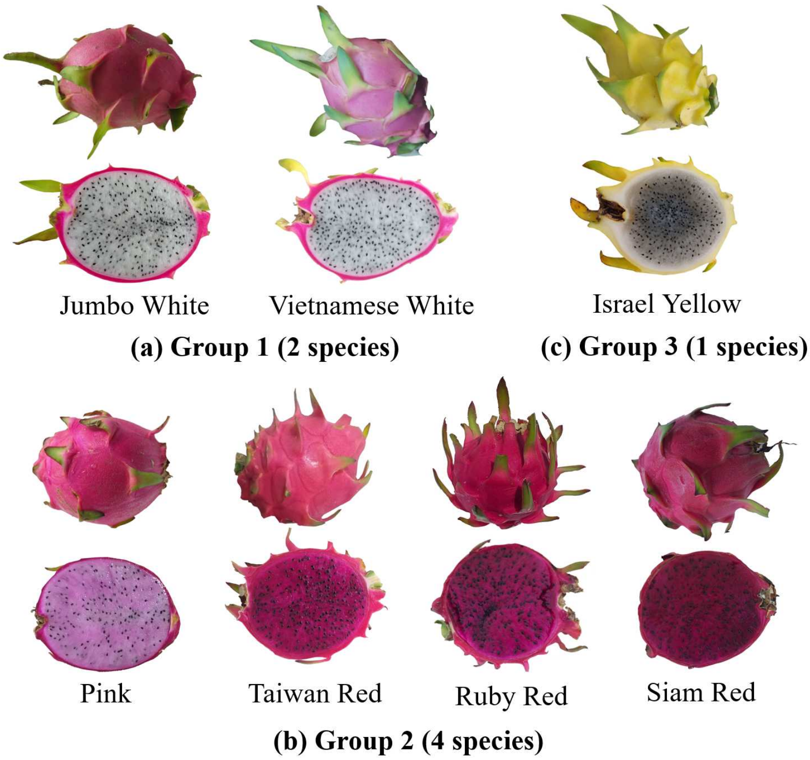

FIGURE 1. Dragon fruits that show its bark color and its fruit pulp in any group and any species

From previous research, it was found popular techniques used in fruit classification, mostly in two prominent methods: Deep Learning and Machine Learning [5,6]. Of course, Deep Learning is probably the simpler method as the features are automatically extracted. However, with Deep Learning processing as a black box [7], it is more difficult for researchers to understand the features used in the classification. For Machine Learning, researchers have to select the attributes that are used in their classification, which can be difficult to know which attributes are the most appropriate. However, with the advantage of selecting this feature itself, researchers are able to understand how the algorithm works. The researchers were able to define the characteristics used for classifying themselves. All parts of the algorithm can be edited. The number of parameters can be set depending on the used algorithm. They can be classified of feature one by one, without confusion. In the past, there had been many successful methods, with mostly an accuracy less than 90% [5,8].

Therefore, in conducting this research, the objectives are to study and search for characteristics finding classification methods suitable for separating 6 species of red bark dragon fruits by using content base and Machine Learning techniques. The results show that our method gave better accuracy. The remainder of this paper will present the relevant research in Section 2, the CBML algorithm in Section 3, the experiment results in Section 4, and remarkable summary discussions in Section 5.

2. **Literature Review.** Popular techniques used in fruit classification are not only Deep Learning, but Machine Learning techniques [5,6]. In the characterization process, the suitable attributes have to be defined to be used for the classification in Machine Learning. However, Deep Learning is automatic feature extraction [7]. If the researchers select the suitable attributes that are well used to classify, the output of the classification will also have good results [6]. Feature extraction is to extract data of an image such as color, shape, size, and texture that can describe all image attributes. It can be all or part of

an image that can represent the entire image [9]. Feature extraction can use a manual selection feature based on the classification problem or use a computer to select a feature automatically [10] to achieve the appropriate feature. These features can be used for recognition, identification, classification, grading, and retrieval [5,6]. The efficiency of these works depends on the appropriate feature extraction. There may be only one or more attributes. The main characteristics that are commonly used to distinguish attributes are color feature, shape feature, and texture feature [6], which can be used to specify statistical features and geometric properties. Most researchers try to select the attributes for the best matching methods used to classify in order to achieve the highest efficiency in their work. Machine Learning is a technique, commonly used to classify. Well-known algorithms have been applied, including Decision Tree (DT), Random Forest (RF), k-Nearest Neighbor (k-NN), Naïve Bayes (NB), and Support Vector Machine (SVM), where these algorithms produce different results [5,6]. The most prominent resultant algorithm is Support Vector Machine (SVM). For example, [11] used color and texture attributes to recognize fruits from photographs of 8 natural fruits. The 16 surface characteristics from the Gray-Level Co-occurrence Matrix (GLCM) and 12 color attributes derived from the RGB color model were used. Using all 28 color and texture attributes can perform better than using just one color or texture feature. The accuracy was 83.33%.

Similarly in [12], the authors used color and texture attributes to recognize 15 supermarket fruits. The fruit images in the RGB color model were converted to the HSV color model. The H Channel and the S Channel were analyzed with 4 statistical properties. There are total 8 attributes. The V Channel are also analyzed to extract surface attributes. The co-occurrence matrices of 5 attributes are contrast, energy, local homogeneity, cluster shade, cluster prominence. Therefore, the combination 13 attributes were used. Fruit image recognition (MDC method) with 13 common color and texture attributes gave an accuracy of 86%. Therefore, color and texture characteristics are considered suitable for species identification of the fruits using datasets of different fruits, with the efficient operation of SVM and MDC.

Color, texture, shape and size attributes [8] were used to distinguish 4 types of date palms. The 3 channels from RGB color model were used to generate histograms using Local Binary Pattern (LBP) and Weber Local Descriptor (WLD) algorithms. There are clearly different patterns, with the features that have a high volume. Therefore, the Fisher Discrimination Ratio (FDR) was used to select important attributes by selecting the first 10 attributes that gave the most FDR values. For shape and size there are totally 4 attributes. There were only 14 attributes used to distinguish the date palms' species. The algorithm used SVM. In experiments, it was found that 10 attributes from LBP or WLD histograms combined with 4 shape and size attributes yielded better results than others up to 99% accuracy. It is very high for fruit classification.

In addition to fruit identification and fruit classification, Machine Learning is also applied to other applications such as detection of bananas based on color and texture characteristics in natural environments [13]. The difficulty of this task is that the color of bananas and the color of the banana trees are the same green. Using the RGB color feature alone is not enough. Therefore, consider using HSV color model in the experiments. The bunch of bananas was isolated with a Histogram of Oriented Gradients (HOG), and texture characteristics were isolated by a Local Binary Patterns (LBP). Support Vector Machine (SVM) algorithm and AdaBoost algorithm were used to compare results. The results found that combination of HOG, LBP and SVM is the most efficient algorithm gaining 100% accuracy. This experiment shows that the Support Vector Machine (SVM) is an algorithm. It provides higher fruit classification performance than other algorithms, but may not work best for data types other than fruit.

The algorithm for multiclass classification has been proposed using differential classifiers [14]. Known as Standardized Variable Distances (SVD), the SVD algorithm is an algorithm that uses the distance between the test data and the mean of the attributes in each class and the Z-score to calculate a score. The results show that SVD is able to classify the Wisconsin Breast Cancer Original (WBCO) and the LED Display Domain (led7digit) datasets more efficiently than the other algorithm. This proved that SVM was not always the best method. Although the performance of SVM is superior to other algorithms, the factor of efficient classification is finding the appropriate attributes to support the classification algorithm to work well.

This paper focuses on studying and searching for characteristics and classification methods that are suitable for separating species of the red-shelled dragon fruit. Based on the above literature review, color and texture attributes were attributes that could be used to differentiate species of red-barked dragon fruit with the dragon fruit image dataset. The algorithm may be experimented with SVM, k-NN, MDC and SVD, which have outstanding performance. Details will be explained in the next section.

3. **The CBML Method.** This paper presents the CBML algorithm which is an automatic method using content base to extract key features and Machine Learning to classify six species of dragon fruits in red bark group. Its main program is presented in Figure 2.

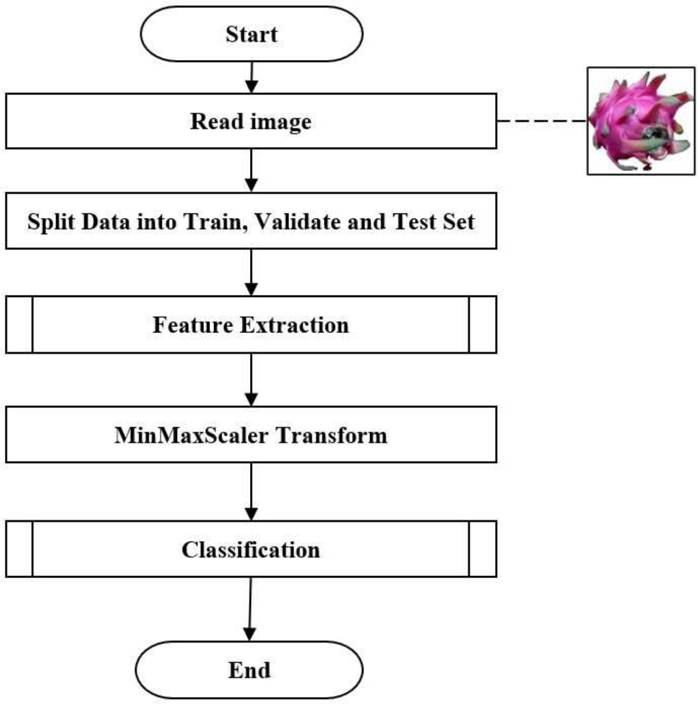

FIGURE 2. The main program of CBML method

This method starts with reading a dragon fruit image which consists of a red-shelled dragon fruit of any species on a white background. An example of a dragon fruit image is shown in Figure 2. Then random split data into 3 groups: Train, Validate, and Test. For each group it will do subprogram, the Feature Extraction, then do MinMaxScaler Transform, and then do subprogram, the Classification. The output result is the species of a dragon fruit image in red bark group.

3.1. **Feature extraction.** From the literature review as mentioned in Section 2, the topic of feature extraction to select appropriate attributes that allow for more accurate categorization can be summarized as Table 1 and thus this provides a guideline for selecting suitable features that enable more accurate classification in CBML method.

From Table 1, [11] and [12] classified fruits and used texture features of GLCM and colors. The possible attributes of GLCM are Dissimilarity (D), Angular Second Moment

Table 1. Comparison of characterization from feature extraction in many methods

Ref.	Data	Features	Classification
[11]	8 types of Mixed Fruits	16 GLCM of RGB (4 Directions from Contrast (Ct), Correlation (Cn), Energy (Ey) and Homogeneity (Hy) and 12 statistical features (Mean, Standard Deviation, Skewness, Kurtosis) of RGB)	SVM (Kernel = not specified)
[12]	15 types of Mixed Fruits	5 GLCM of V channel from HSV (contrast, energy, local homogeneity, cluster shade and cluster prominence) and 8 statistical features of H and S channel from HSV (Mean, Standard Deviation, Skewness, Kurtosis)	Minimum Distance Criterion (MDC) (distance = not specified)
[8]	4 types of Dates	Use Local Binary Pattern (LBP) and Weber Local Descriptor (WLD) from each channel of RGB and then use Fisher Discrimination Ratio (FDR) to select the top-10 highest FDR value combined with 4 shape and size features (Major axis length, Minor axis length, Ellipse eccentricity and Area)	SVM (kernel = RBF)
[13]	detect banana and background	Histogram of Oriented Gradients (HOG) and Local Binary Patterns (LBP)	SVM (Kernel = not specified) compared with AdaBoost
[14]	public dataset from Wisconsin Breast Cancer Original (WBCO) and LED Display Domain (led7digit)	9 attributes from WBCO and 7 attributes from led7digit.	SVM (Kernel = RBF) k-NN (k = 3, distance = Euclidean) MDC (distance = Minkowski, $p = 3$) SVD (distance = Manhattan) SVM (Kernel = RBF) k-NN (k = 3, distance = Chebyshev) MDC (distance = Minkowski, $p = 3$) SVD (distance = Hellinger)
CBML method	6 species of Thai Dragon Fruit from red bark group	2 attributes GLCM of RGB (Dissimilarity (D) and ASM only Directions: 0°) and 32 attributes of color features with RGB, HSV and AB for each channel in any color model using four statistical features (Mean, Standard Deviation, Skewness, and Kurtosis) Total 34 attributes Each attribute used MinMaxScaler to normalization into 0 to 1 value.	SVM (Kernel = Polynomial, degree = 8)

(ASM), Contrast (Ct), Correlation (Cn), Energy (Ey) and Homogeneity (Hy) and used information in 4 directions: 0°, 90°, 45°, and 135°. The colors are mostly RGB, HSV and LAB color models in 9 possible channels: R, G, B, H, S, V, L, A and B using statistical features (Mean, Standard Deviation, Skewness, and Kurtosis). Therefore, the total of possible attributes are 60 attributes. The CBML tries to find out key attributes from these 60 attributes. Each color channel of the red dragon fruit in RGB, HSV and LAB color models is shown in Figure 3, and its binary image is shown in Figure 4. In Figure 4, the G channel image is very similar to the L channel image. Therefore, CBML can reduce L attribute. The texture of red dragon fruit image with petals around it makes the fruit symmetric, and no matter with direction information. Figure 5 shows the mean of dissimilarity in one way ANOVA analysis of texture feature in any direction information. This graph confirms that texture in 0 degree is the most dissimilarity with another direction. The Dissimilarity (D) and ASM are the most dissimilarity with other texture feature. Therefore, the CBML used only 34 attributes: 2 attributes GLCM of RGB (Dissimilarity (D) and ASM only Directions: 0°) and 32 attributes of color features

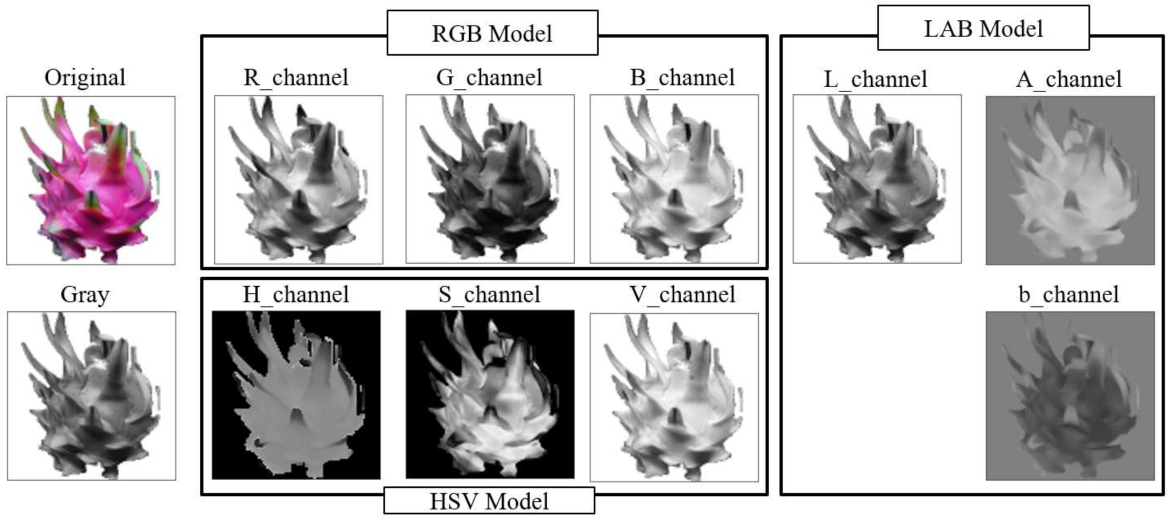

Figure 3. Examples of the red dragon fruit in each color channel of RGB, HSV, and LAB

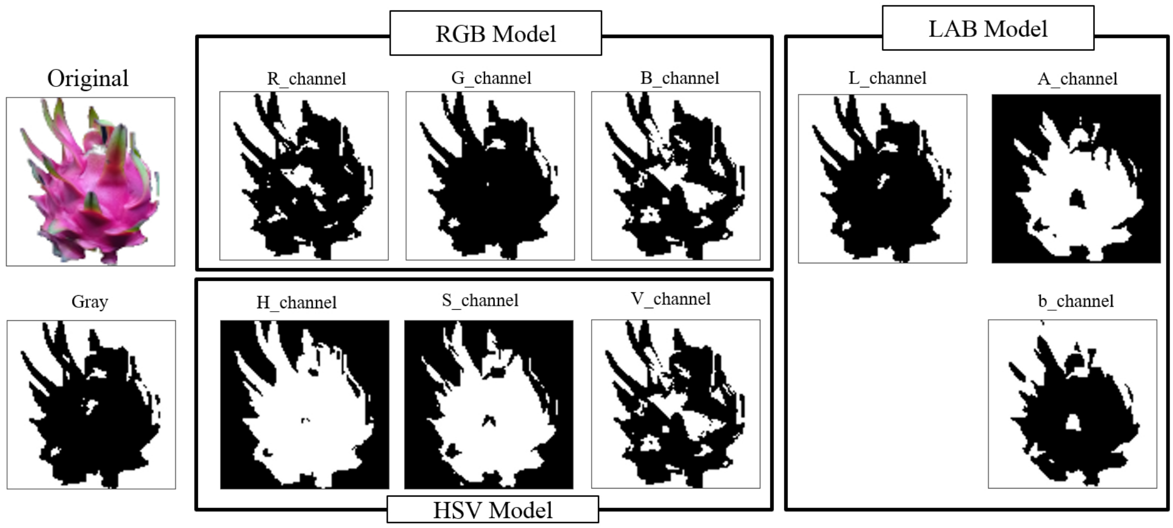

Figure 4. The binary image of Figure 3

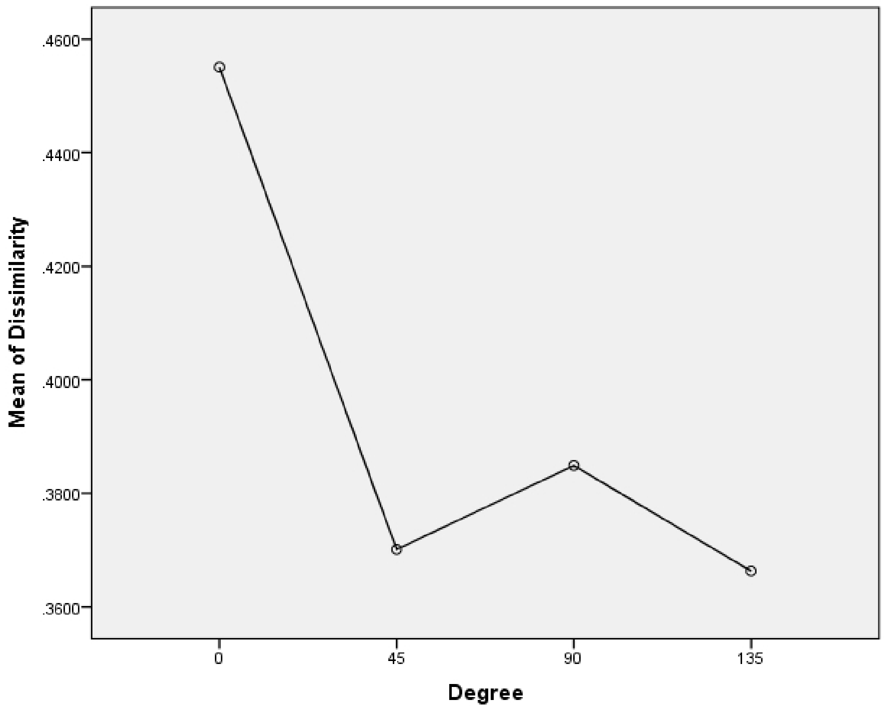

FIGURE 5. The mean of dissimilarity of texture feature in any degree

with RGB, HSV and AB for each channel in any color model using four statistical features (Mean, Standard Deviation, Skewness, and Kurtosis) as shown as the last row in Table 1.

Before the feature data is processed, the scaling value of the data is adjusted to a value in the range of 0-1 with the MinMaxScaler Transform (Equation (1)) using the data in Train Set. The scale transform in Train Set is a reference criterion for adjusting the value of the Validate Set and the Test Set.

$$MinMaxScaler\ Transform = (x - \min)/(\max - \min) \tag{1}$$

where, in each attribute max is not equal to min, x is the value, min is the minimum value in the Train Set, and max is the maximum value in the Train Set. The *MinMaxScaler Transform* is a normalization transform of any value between min and max into 0 to 1.

3.2. **Classification.** For classification in this research, the performance of the SVM, k-NN, MDC [12] and SVD [14] was compared to study and find the most suitable method to classify 6 species of the red bark dragon fruit. The SVM algorithm was defined 4 kernels: Linear, Polynomial (definition degree = 8), Radial Basis Function (RBF), and Sigmoid. All sklearn processed the k-NN, MDC [12] and SVD [14] with 5 effective classification distances [15]: Manhattan, Euclidean, Minkowski (defined $p = 3$), Hellinger [16] and Chebyshev, where k-NN is set to k equal to 3. The results will be described in the next section.

4. **Experiment Results.** Dataset is created by taking a photo of a dragon fruit on white background. Botanists listed the species' validity labels used for validation as baseline information. There are 9,754 images, divided into the red bark group 7,834 images and the yellow bark group 1,920 images. The dataset used in this research is a dataset obtained from randomly selected images of 6 species of red-shell dragon fruit, 1,000 images of each species, for a total of 6,000 images out of 7,834 images. The image size is 100×100 pixels. Some examples of images are shown in Figure 6. All images are a dragon fruit with red bark on white background.

FIGURE 6. Examples of input images

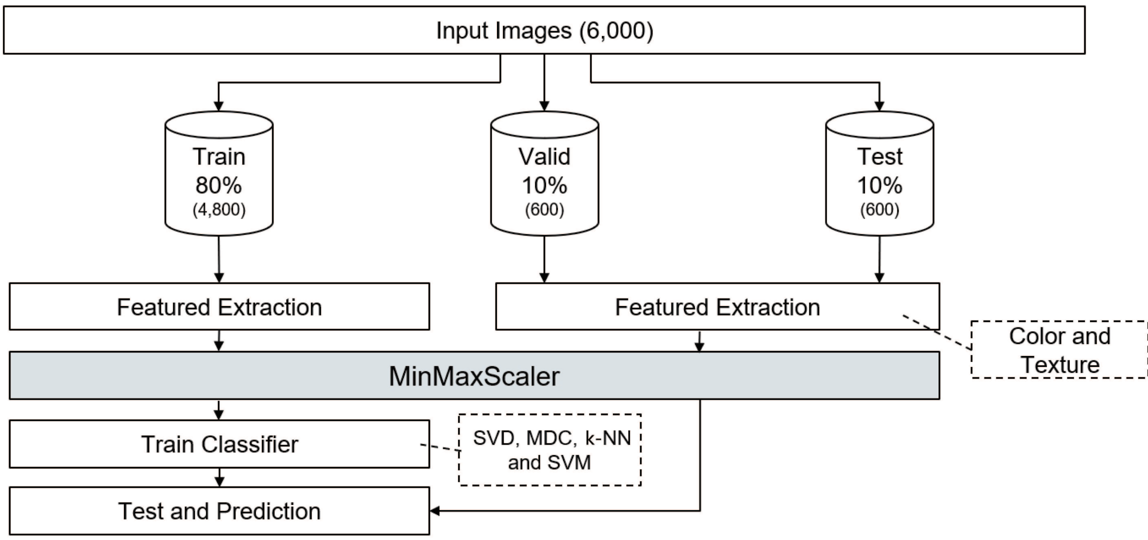

FIGURE 7. Train Set, Validate Set, and Test Set using the 10-Folds Validation technique

The experiment divided the 1,000 images in each species into 800 images for the Train Set, 100 images for the Validate Set and the remaining 100 images for the Test Set, using the 10-Folds Validation technique as shown in Figure 7.

The purpose of the experiment was to study and find the most suitable characteristics and classification methods. Therefore, the dataset was extracted with color and texture attributes, processed with different algorithms, namely SVM, k-NN, MDC [12] and SVD [14], each of which has defined different parameters. All of the experimental results were shown in Table 2. From Table 2, columns are feature attributes of CBML (34 attributes), 28 attributes of [11], and 14 attributes of [8]. Rows are the classification algorithms, that is, SVM, k-NN, MDC [12] and SVD [14]. The results from k-NN, MDC and SVD training cannot reach 100% accuracy. Only the SVM with polynomial, on the second row, training can reach 100% accuracy with the methods between CBML and [11]. The CBML method gave better accuracy of 98.47% than [11] with accuracy rate of 94.20%.

The confusion matrix with SVM algorithm is shown in Figure 8. The accuracy results for Pink, Siam Red, Taiwan Red, Ruby Red, Jumbo White, and Vietnamese White species are 99.5, 99.6, 97.4, 97.4, 98.2, and 98.7, respectively. The Taiwan Red is the most similar

TABLE 2. Accuracy comparison of different features and different classifications

Algorithm	**Distance/ Kernel**	**CBML with 34 attributes color and texture**			**28 attributes color and texture [11]**			**14 attributes texture and shape [8]**		
		Train	**Valid**	**Test**	**Train**	**Valid**	**Test**	**Train**	**Valid**	**Test**
SVM	Linear	91.22	90.90	90.75	83.12	82.65	82.55	57.36	56.12	56.65
	Polynomial	**100.00**	**98.28**	**98.47**	**100.00**	**94.10**	**94.20**	**94.83**	**52.82**	**52.88**
	RBF	92.86	92.33	92.50	87.25	86.17	86.22	62.87	58.98	59.23
	Sigmoid	38.97	38.85	38.75	21.59	21.62	21.87	11.32	11.12	11.30
k-NN	Manhattan	96.51	92.18	92.35	93.72	86.12	85.92	70.80	45.58	46.75
	Euclidean	96.89	92.87	93.23	94.01	87.00	86.87	69.64	45.67	45.13
	Minkowski	97.04	93.12	93.20	94.21	86.70	87.13	69.10	44.62	44.35
	Hellinger	97.05	93.02	92.82	94.25	86.68	86.62	67.19	39.73	40.63
	Chebyshev	93.46	92.18	91.90	93.90	85.93	86.00	67.94	42.27	41.17
MDC [12]	Manhattan	61.95	61.63	62.08	54.99	54.88	55.10	42.95	43.00	43.27
	Euclidean	64.96	64.58	64.52	59.29	59.43	59.13	44.64	44.43	44.27
	Minkowski	65.77	65.60	65.57	61.16	60.80	60.70	43.80	43.02	43.98
	Hellinger	67.15	63.87	64.22	56.86	56.90	57.17	43.98	43.52	43.50
	Chebyshev	64.46	64.17	63.68	58.56	58.45	58.30	37.63	36.93	37.18
SVD [14]	Manhattan	62.42	62.28	62.50	56.65	56.32	56.22	44.05	43.70	44.17
	Euclidean	63.88	63.72	63.88	58.89	59.07	58.70	44.60	44.38	44.50
	Minkowski	64.68	64.33	64.48	59.97	59.93	59.73	44.34	44.10	43.97
	Hellinger	63.90	63.60	64.08	57.73	57.80	57.70	45.06	45.43	45.00
	Chebyshev	69.62	66.43	66.43	60.86	60.87	60.58	43.09	42.48	42.47

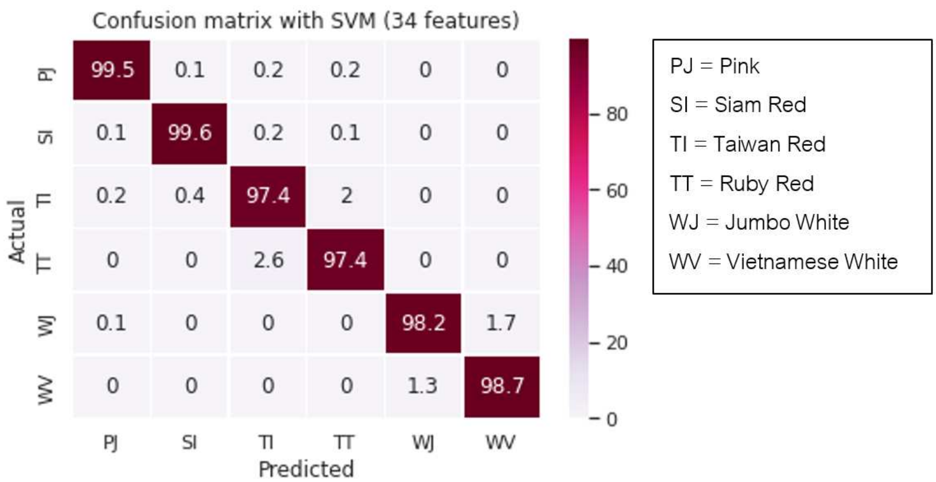

FIGURE 8. The confusion matrix with SVM algorithm

to Ruby Red species. The results were found that SVM (Kernel = Polynomial, degree = 8) had the highest efficiency in classifying the species of red-shelled dragon fruit. The CBML method uses only 34 attributes from 60 attributes, and this can reduce almost 50%. The CBML method using SVM with Polynomial can characterize and classify the most suitable in the species classification of Thai red dragon fruits.

5. **Remarkable Conclusions.** The CBML method can automatically characterize and classify method suitable for the separation of 6 species of Thai red dragon fruits. The features extraction from texture and color with 34 attributes and Machine Learning using SVM with Polynomial in 8 degree can increase the efficiency of classifying species of red dragon fruits. The CBML method can achieve with 98.47% accuracy. The correct classification of red dragon fruit can be very helpful. For example, plants can have automated systems for separating and labeling species, farmers can sell dragon fruits for a better price, and consumers who allergy with the species will be able to avoid them properly or choose their favorite species. The CBML used only the laboratory dataset. In fact, if the Thai dragon fruits can be classified of their species with nature outdoor environment dataset, this will be more useful such as, harvesting dragon fruit with a robot, detecting the ripeness of the fruit to manage for transportation and distribution, counting of produce before harvesting, or other approaches that are expected to enhance human well-being or help relieve labor. This will continue to develop in the future work.

REFERENCES

[1] A. Trivellini, M. Lucchesini, A. Ferrante, D. Massa, M. Orlando, L. Incrocci and A. Mensuali-Sodi, Pitaya, an attractive alternative crop for Mediterranean Region, *Agronomy*, vol.10, no.8, DOI: 10.3390/agronomy10081065, 2020.

[2] National Bureau of Agricultural Commodity and Food Standards, Ministry of Agriculture and Cooperatives, *Thai Agricultural Standard: Dragon Fruit*, https://www.acfs.go.th/standard/download/eng/DRAGON-FRUIT-ENG.pdf, 2015.

[3] Department of Agricultural Extension, Ministry of Agriculture and Cooperatives, *Agricultural Production Information Service System*, https://production.doae.go.th/service/site/login, 2020.

[4] A. Kumaresan, S. Swain, V. Baskaran, K. Venkatesan, S. Krishnan and B. Narayanasamy, Distinguishing three Dragon fruit (*Hylocereus* spp.) species grown in Andaman and Nicobar Islands of India using morphological, biochemical and molecular traits, *Scientific Reports*, vol.11, DOI: 10.1038/s41598-021-81682-x, 2021.

[5] S. K. Behera, A. K. Rath, A. Mahapatra and P. K. Sethy, Identification, classification & grading of fruits using machine learning & computer intelligence: A review, *Journal of Ambient Intelligence and Humanized Computing*, DOI: 10.1007/s12652-020-01865-8, 2020.

[6] K. Hameed, D. Chai and A. Rassau, A comprehensive review of fruit and vegetable classification techniques, *Image and Vision Computing*, vol.80, pp.24-44, 2018.

[7] C. Janiesch, P. Zschech and K. Heinrich, Machine learning and deep learning, *Electronic Markets*, vol.31, pp.685-695, 2021.

[8] G. Muhammad, Automatic date fruit classification by using local texture descriptors and shape-size features, *Proc. of UKSim-AMSS 8th European Modelling Symposium on Computer Modelling and Simulation (EMS2014)*, pp.174-179, 2014.

[9] I. M. Hameed, S. H. Abdulhussain and B. M. Mahmmod, Content-based image retrieval: A review of recent trends, *Cogent Engineering*, vol.8, DOI: 10.1080/23311916.2021.1927469, 2021.

[10] I. A. Siradjuddin, A. Sakinah and M. K. Sophan, Combination of feature engineering and feature learning approaches for classification on visual complexity images, *International Journal of Innovative Computing, Information and Control*, vol.17, no.3, pp.991-1005, 2021.

[11] S. Jana, S. Basak and R. Parekh, Automatic fruit recognition from natural images using color and texture features, *Proc. of the 2nd International Conference on 2017 Devices for Integrated Circuit (DevIC2017)*, pp.620-624, 2017.

[12] A. Selvaraj, N. Shebiah, S. Nidhyananthan and L. Ganesan, Fruit recognition using color and texture features, *Journal of Emerging Trends in Computing and Information Sciences*, vol.1, no.2, pp.90-94, 2010.

[13] L. Fu, J. Duan, X. Zou, G. Lin, S. Song, B. Ji and Z. Yang, Banana detection based on color and texture features in the natural environment, *Computers and Electronics in Agriculture*, vol.167, DOI: 10.1016/j.compag.2019.105057, 2019.

[14] A. Elen and E. Avuçlu, Standardized variable distances: A distance-based machine learning method, *Applied Soft Computing*, vol.98, DOI: 10.1016/j.asoc.2020.106855, 2021.

[15] H. A. A. Alfeilat, A. B. A. Hassanat, O. Lasassmeh, A. S. Tarawneh, M. B. Alhasanat, H. S. E. Salman and V. B. S. Prasath, Effects of distance measure choice on k-nearest neighbor classifier performance: A review, *Big Data*, vol.7, no.4, pp.221-248, 2019.

[16] E. Hellinger, New justification of the theory of quadratic forms of infinite variables, *Journal for Pure and Applied Mathematics*, vol.136, pp.210-271, 1909 (in German).