Using Random Forest Algorithm to Grading Mango's Quality Based on External Features Extracted from Captured Images

Nguyen Minh Trieu and Nguyen Truong Thinh *

Institute of Intelligent and Interactive Technologies, University of Economics Ho Chi Minh City – UEH, Vietnam
Email: trieunm@ueh.edu.vn (N.M.T.)
*Correspondence: thinhnt@ueh.edu.vn (N.T.T.)

***Abstract*—The grading of mango is still a manual process in agriculture. Nowadays, mangoes are classified based on human experience, which makes the grade not uniform for agricultural product export establishments. Therefore, the automated grading of mango is very important to solve these problems. In this study, a random forest algorithm is proposed for an automated mango grading system based on quality attributes such as density, surface defect, and weight. The internal features including dimensions and surface defects are extracted via the captured image. These features are combined with the weight to estimate density. This study uses 732 mangoes that are collected from several local farms. The experiment of the grading system has high accuracy with 98.3%. Instead of using Non-Destructive Testing (NDT) equipment, this grading method can be used to apply to evaluate the quality of other tropical fruits.**

***Keywords*—mango sorting, machine learning, grade system, random forest**

I. Introduction

Mango is a popular fruit because of its nutritional value and medicinal value, it is rich in vitamins C, E, A, B, and K [1]. Mango is an export item that accounts for a large proportion of Vietnam. The demand for mango is growing in Western markets, so mangoes are a potential export commodity in Vietnam. However, exportation is limited mainly due to the lack of tools and techniques to meet the quality requirements of importing countries. Currently, consumers demand mango not only for external features such as color, shape, and surface but also for internal quality such as sugar content, and acidity. Mango's taste as well as grading quality are extremely important for farmers. The quality of mango depends on the cycle of growth, development, and picking ripeness, but also depends on the post-harvest process such as handling, sorting, and preserving. The parameters affecting the quality of mango fruit under the current assessment are as follows: size, shape, color, Total Dissolved Solids (TSS), acidity, pH, physiological weight, juice, pulp, and moisture. The objective of this paper is to consider the recent work reported on the quality parameters of mango. The effect of harvesting and post-harvest treatments concerning the quality of mango and to explore the potential of available non-destructive techniques for determination of maturity, physical, biochemical, viscoelastic and rheological quality parameters, and internal disorders in mango fruits.

Nowadays, machine learning methods are applied to solve specific problems such as fruit classification, medical, traffic, etc. [2–4]. These papers evaluate the quality of mango following color, shape, defect, and weight. Pandey et al proposed efficient algorithms for color feature extraction to grade mango [5]. Khoje *et al.* [6] used Feed Forward Neural network (FFNN) and Support Vector Machines for size grading mango from Maharashtra, India while the backpropagation neural networks method is employed to classify the mangoes into three classes—SS, S, L. Additionally, Fuzzy is also used effectively to classify Maturity and Quality [7]. Moreover, the perimeter, area, roundness, and percent of the defect were extracted to identify whether the mango's quality for export, local, or removal [8]. There are also some re-searches using machine learning to analyze and estimate the weight of mangoes. Schulze *et al.* [9] used the weights of mangoes to determine the type of mango. Sa'ad *et al.* [10] estimated weight and volume by using the Cylinder approximation analysis method. The process for classification in machine learning always follows the following steps: data collection, data processing, training and optimization of predictive models, and finally practical application of the optimal model. In most previous studies, the quality of mangoes is determined by looking at external or internal features so the quality of mangoes is assessed to be inadequate. Therefore, this study proposes a new grading system based on quality features including density, defect, and weight to evaluate mangoes into three categories G1, G2, and G3 with G1 being the best group.

Manuscript received April 5, 2023; revised May 15, 2023; accepted June 20, 2023.

doi: 10.18178/joig.11.4.391-396

II. Extracting Features of Mango

The size estimation of mango from 2D images has been a challenging problem in the field of computer vision. The process starts by obtaining multiple views of images of the mango.

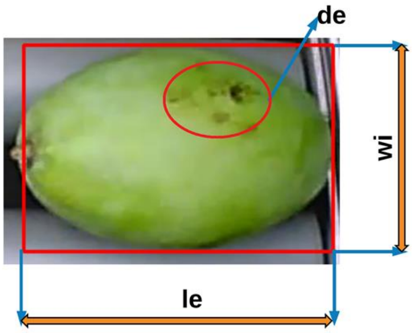

Figure 1. The external features of mango.

After the intrinsic and extrinsic parameters of the camera are calibrated, these parameters are used to estimate the mango's dimension. The external features consist of length, width, and defect (Fig. 1). Length (*le*) is defined as the maximum length mango's x-axis which lies between the tip and the pole of the mango. The width (*wi*) is defined as the length of the mango's y-axis which is the widest line perpendicular to the x-axis. The defect (*de*) of the mango is the damage on the surface caused by insects or collision during the growth of the mango. In the image processing chamber, mangoes are captured in a random direction. The acquired image is preprocessed using different methods such as increased frames per second (*fps*), image noise filter, edge detection, and boundary tracking. Features of mangos are considered as length, width, Surface defect, and weight. The process of extracting features is indicated in Fig. 2.

The performance of the classification mango system depends on the quality of the captured images, hence, if the number of still frames in each mango rises, the accuracy for the extracting features is improved. The still frames are extracted from the video image at the fps of the camera. Therefore, this article suggests an algorithm that increases the number of fps (optical flow-based intermediate frame synthesis) [8], which exceeds the responsiveness of the camera. In a brief description, the new frame f_x is synthesized from two adjacent frames (f_i, f_{i+1}) by the i-directional optical flows $(f_{i\rightarrow i+1},\ f_{i+1\rightarrow i})$, respectively. Where $(f_{x\rightarrow i}, f_{x\rightarrow i+1})$ denote the optical flow from f_x to fi and f_x to f_{i+1}, respectively. The image f_x can be synthesized following Eq. (1).

$$f_x = a \cdot bi\ (f_i, f_{x\rightarrow i}) + (1-a) \cdot bi(f_{i+1}, f_{x\rightarrow i+1}) \qquad (1)$$

where b_i is a bilinear interpolation function and α controls the contribution of (f_i, f_{i+1}) and depends on two factors: temporal consistency and occlusion reasoning.

In each frame, the noises are filtered to improve the image quality by using the Gaussian method which is adaptive as Eq. (2). The kernel matrix slides across each row of mages according to each region of the image, thus the central pixel is the sum of the results. The smooth continuous boundary of the mango is achieved by the filtering process.

$$G(\mathrm{x},\mathrm{y}) = \mathrm{A}e^{\frac{-(\mathrm{x}-\mu_x)^2}{2\sigma_x^2} + \frac{-(\mathrm{y}-\mu_y)^2}{2\sigma_y^2}} \qquad (2)$$

where μ is mean and σ is the variance of Gaussian distribution; A is the Gaussian coefficient based on the standard deviation σ.

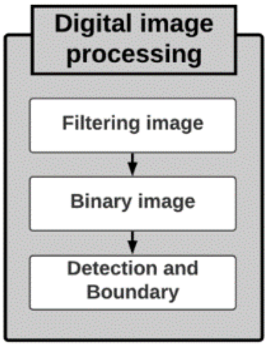

Figure 2. The image processing of mango classification system.

The used images are color images in RBG color spaces namely Red, Green, and Blue that are converted into a binary image using an adaptive threshold for mango. That means the mango is white and the background is black (0, 1 respectively). From the binary image, the edges of the object are highlighted by partial geometric differential equations [11]. By using the maximum principle and rigorous mathematical analysis, the contour of the mango is found effectively. The boundary pixels of the object are detected and interpolated into curves according to Eq. (3), where the elasticity and rigidity coefficients of the curve are represented $\alpha > 0$ and $\beta > 0$, respectively.

$$E_{\mathrm{int}} = \int_0^1 (\alpha|v'(t)|^2 + \beta|v''(t)|^2)dt \qquad (3)$$

The edges of the mango are also detected based on Eq. (4), $\nabla I(v(t))$ is the maximum curve possible.

$$E_{ext} = -\lambda\int_0^1 |\nabla I(v(t))|\,dt \qquad (4)$$

Eqs. (3) and (4) implement the edge and boundary through Eq. (5).

$$E(v,\alpha,\beta,\lambda) = \int_0^1 (\alpha|v'(t)|^2 + \beta|v''(t)|^2)dt - \lambda\int_0^1 |\nabla I(v(t))|\,dt \qquad (5)$$

From the mango boundaries in the image, the actual size is estimated by using Eq. (6). The ratio K depends on the distance between the camera and the object, the shorter the distance, the actual length is displayed with a larger number of pixels.

$$L = KA_{pixels} \tag{6}$$

where *L* means length, *A* pixels means the number of pixels, and *K* means the ratio between pixel size and actual size.

Let *L*, *A*, and *n* are the length, the number of pixels, and the number of images, respectively. The average of length $\bar{L}$ and the average of pixels $\bar{A}$ are calculated by Eqs. (7) and (8).

$$\bar{L} = \frac{1}{n}\sum_{i=1}^{n} L_i \tag{7}$$

$$\bar{A} = \frac{1}{n}\sum_{i=1}^{n} A_i \tag{8}$$

From Eqs. (7) and (8), *K* ratio factor is shown in Eq. (9), where ε is the error of *K*.

$$K = \frac{\sum_{i=1}^{n} A_i.L_i - n\bar{A}\bar{L}}{\sum_{i=1}^{n} A_i^2 - n\bar{A}^2} + \varepsilon \tag{9}$$

Density (*ds*) depends on weight and volume which is calculated based on *le* and *wi*. Therefore, the task is to predict the volume (*V*) which is shown by Eq. (10).

$$V = b_0 + b_1 le + b_2 wi + \varepsilon^V \tag{10}$$

where the coefficients of the variables are b_0, b_1, b_2, and means the error of volume.

Based on *V* and *we*, density (*ds*) is calculated by Eq. (11).

$$ds = \frac{we}{V} + \frac{1}{n}\sum_{i=1}^{n}(\varepsilon_i^{we} + \varepsilon_i^{V}) \tag{11}$$

In this section, based on a series of calculation formulas, the size of the image is determined to the actual size with its error. The estimation process is calibrated depending on the hardware of the machine. A method of determining density is proven to be effective, so this method can be used for some other tropical fruits.

III. Applying Machine Learning

Random Forest (RF) model is considered to classify types of mangoes. This is a very popular and efficient classification model to solve the problem of classification when predictive variables have nonlinear relationships. RF model is an ensemble learning method of many decision-tree models to obtain predictive results with small variance. The training process of RF model is shown in Fig. 3. This approach is performed following three steps.

A. Apply Bootstrap Aggregating to Create k Subsets from Training Set

Let $F = \{\vec{f}_i : 0 < i < n+1\}$ is feature set of *n* mango samples that have been labeled, where ds_i, we_i, de_i, are density, weight, and defect of mango respectively, and $\vec{f}_i = [de_i, ds_i, we_i]^T$.

Let $T = \{t_i : 0 < i < n+1\}$ is type set according *n* element in set F, where $t_i = \{1, 2, 3\}$ with G1, G2, and G3 are types of mango respectively. RF method creates *k* subset by selecting a random sample with replacement of *[F, T]*. Hence, the set of *k* subset is $B = \{b_i : 0 < i < k+1\}$, where bi is the i^{th} subset. The cardinality of bi is equal to *S* or it could be symbolized $|b_i| = S$. Moreover, b_i has $(1-1/e)$ unique examples of *S*.

B. Training RF Model

Nodes labeled with input features are chosen and decentralized leads to a subordinate decision node. there are two ways to implement this process: Gini or Entropy but in this study, Gini is chosen, because Gini can minimize misclassification and Gini will tend to find the largest class while Entropy tends to find groups of classes that make up approximately 50% of the data. In other words, Gini's computing time is faster than Entropy, this will be for reduced data training time.

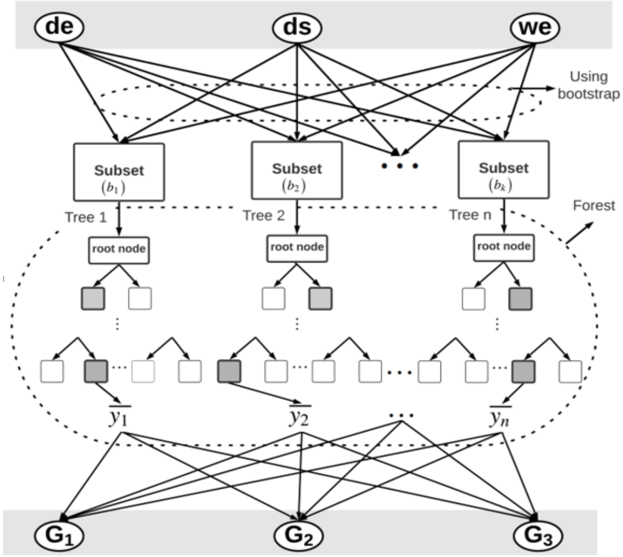

Figure 3. The training process of random forest model.

The Gini formula is given by Eq. (12). Gini impurity is a measure of how often a randomly chosen F would be incorrectly labeled if it was randomly labeled according to the distribution of labels in b_i.

$$G(F) = 1 - \sum_{j=1}^{t} P_j^2 \tag{12}$$

C. Result Selection

The result of random forest (y_{FR}) is selected from the result of trees in the forest by the majority vote method. In this section, the RF model has been generalized about the theory and how to apply them to the dataset. The results are implemented clearly on the existing data set. RF model is used in this case because of some reason. Firstly, the random forest's benefits include effective noise management and faster processing of significant variables.

Moreover, the fact that only two parameters need to be tuned contributes to the growing popularity of RF [12].

IV. Experiments and Results

With the development of science and technology, smart algorithms are applied in many fields such as medicine, robotics, material, classification, and so on [13–16]. In this work, an algorithm is proposed for an automated mango grading system based on quality attributes such as density, surface defect, and weight. The dataset of 732 mangoes is harvested from November to June in several orchards. The original dimension is measured by a Mitutoyo tool with an accuracy of 0.05 mm. Besides, the weight of each mango is measured by electronic scales based on load-cell with an accuracy of 0.01 g. Besides, the volume (V) of the mango is measured by the overflow method with the device being a 1,000 mL glass jar and the 0.4 mL error. The mango classification system has been designed and applied in practice to examine many orchards (Fig. 4).

There are two parts in the mango classification system comprising a computer vision system and a sorting system. First, in the image processing chamber, dimensions and defects are extracted. Second, mangoes are moved to the tray by a roller conveyor system, then the weight is measured by the load cell in the sorting system. Density is calculated based on weight and volume. The volume of the mango is estimated by the length and width. Finally, the data is processed and analyzed to reduce the noise for clear data. The mango is graded into three groups with Grade 1 being the highest quality. The steps of image processing have been experimented in Fig. 5.

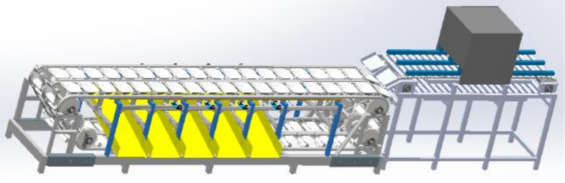

Figure 4. Mango classification system.

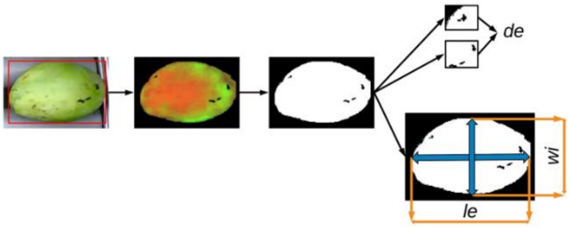

Figure 5. Extraction of external features.

The dimensions (length and width) are estimated indirectly through a series of algorithms including identify the mango, convert RGB image, binary image, then extracted features. Because the dimensions are calculated with the pixel unit on the binary image, there are error between the actual and predicted result. Hence, the figures are calculated and calibrated. This is an important step for classification [17]. The volume is predicted from the height and width of the mango. Data are taken from real data to create a volume prediction model. From the graph, the features: length, Width have a linear relationship with *V*. Therefore, the volume is predicted by a linear model with variables: length, and width that is defined in Eq. (13).

$$V = -1087.97 + 4.02le + 11.74wi \tag{13}$$

The comparison between the actual and predicted volume is shown in Table I. During processing, the signal is always interfered with, causing the measurement results of the load cell to be inaccurate. All signals from the load cell are passed through the Kalman noise filter, where too large values of variation are discarded.

TABLE I. The Comparison of Actual Values and Estimated Values

Index	Density (kg/m^3)		Defect (cm)	
	Actual size	Estimated size	Actual size	Estimated size
1	0.72	0.73	2.4	2.8
2	0.56	0.59	0.3	0.1
3	0.86	0.90	0.4	0.9
4	1.02	1.00	1.2	1.0
5	0.67	0.71	2.4	2.8
…	…	…	…	…
37	0.63	0.61	5.1	5.5
38	0.79	0.80	2.2	2.7
39	0.88	0.86	3.5	3.9
40	0.94	0.95	1.5	1.7

After obtaining the mass signal from the load cell, these signals are decoded and estimated to show the true weight result of the mango. This weight result is affected by the position of the mango on the tray. Therefore, the position of mangoes on the tray will be checked by a camera. The estimated weight values will synchronize with height, width, defect giving results of the process to form a closed loop.

The data is divided into three parts including training data, validation data, and testing data. Data set with 732 data samples extracted and aggregated from images and load-cell. The dataset was divided into three parts with 459 samples used for the training model, 136 samples used for validation, and 137 samples used for testing. A visualization of the training dataset is shown in Fig. 6. In two ranges of the defect (0, 2) and (6, 8), the mango's types are quite obvious but become more complicated with the middle defect = (4, 6).

When the mango defect is too large or too small, the mango is easy to identify its type but at the average level, the mango classification becomes difficult much more. The weight and density are recognized easily. However, there are still significant amounts of mangoes that are rated as poor quality since the rest variables do not satisfy the standard. The number of trees in the forest given in Fig. 7 was controlled to find the best model. Increasing the number of trees, the accuracy of the RF model is between 98% to 98.3% from the 50th tree onwards. Therefore, to ensure the stability and training speed of RF model, the number of trees is selected to 50 for the parameter of the random forest model. The problem that affects the accuracy of the classification process is that the boundaries between mango groups are not fixed and intertwined.

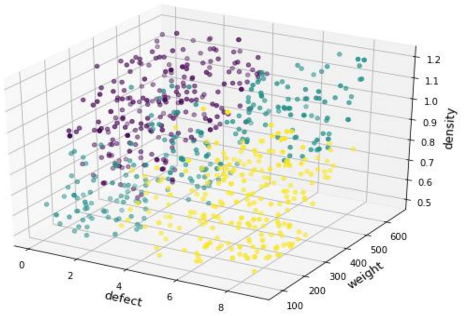

Figure 6. Visualization of the training dataset.

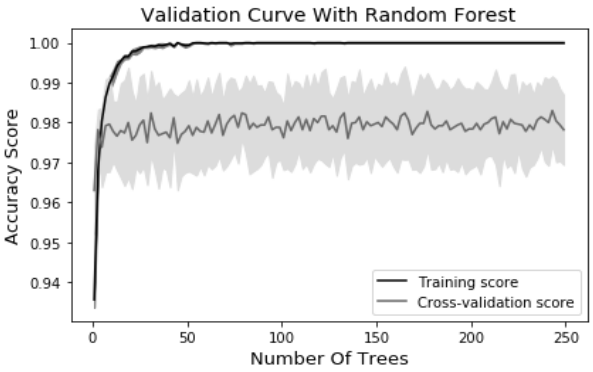

Figure 7. The influence of the number of trees on the accuracy of the RF model.

In Fig. 8, the area of density (0.5, 1.1) is explicitly and accurately defined by the classification system according to the law. The classification of the mango and the classification method of the RF model tend to be identical when it is categorized according to a set of laws which is the reason for the high accuracy. The reason for the error in G2 is too much interference between G1, G2, and G3 in training data. The relationship between categories 1, 3 and 2 is not evident in the separation of category 2 from the other two types. In the RF model, the ratio is significantly different when the rate of guessing data of G2 to G1 is smaller than that of G2 to G3.

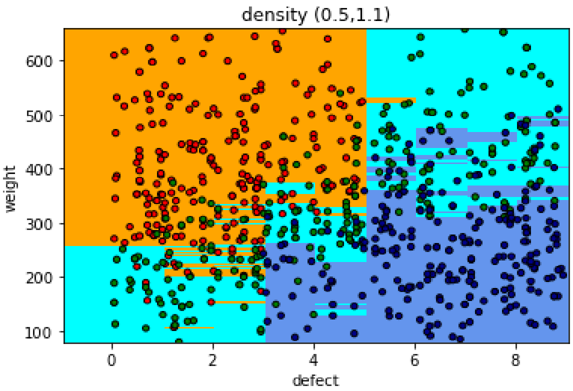

Figure 8. Predicted results of RF model.

V. Conclusions

The RF model is used to grade the automatic mango. The system applied computer vision technology to extract features from captured images and signal load-cell. There are several conclusions drawn as follows:

Firstly, the accuracy of the RF model in this study is highly accurate at 98.3%. Secondly, Throughout the classification process, a sequence of analysis methods in computer vision is used to transform the captured image to estimate density which is an important feature for grading mango. Finally, Since the category of mango is categorized on the basis of the rules and the relationship of the mango feature, therefore, the random forest approach has an advantage over other methods when classifying the based on rules generated from input variables.

In future work, other algorithms will be applied to improve the accuracy and optimization of the system. The training dataset will be collected more diverse mango varieties in many seasons of the year to increase the adaptability of the automatic mango classification system.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Nguyen Minh Trieu: methodology, wring original draft, visualization, validation, configured. Nguyen Truong Thinh: writing, review and editing, methodology project administration. All authors had approved the final version.

Funding

This research is funded by University of Economics Ho Chi Minh City—UEH University, Vietnam.

References

[1] T. A. Zafar and J. S. Sidhu, "Composition and nutritional properties of mangoes," in *Handbook of Mango Fruit: Production, Postharvest Science, Processing Technology and Nutrition*, 2017, vol. 217.

[2] N. M. Trieu and N. T. Thinh, "Development of grading system based on machine learning for dragon fruit," in *Proc. Regional Conference in Mechanical Manufacturing Engineering*, Springer, Singapore, 2022, pp. 230–243.

[3] B. Kompa, J. Snoek, and A. L. Beam, "Second opinion needed: Communicating uncertainty in medical machine learning," *NPJ Digital Medicine*, vol. 4, no. 1, pp. 1–6, 2021.

[4] M. Shafiq, Z. Tian, A. K. Bashir, A. Jolfaei, and X. Yu, "Data mining and machine learning methods for sustainable smart cities traffic classification: A survey," *Sustainable Cities and Society*, vol. 60, 102177, 2020.

[5] R. Pandey, S. Naik, and R. Marfatia, "Image processing and machine learning for automated fruit grading system: A technical review," *International Journal of Computer Applications*, vol. 81, no. 16, pp. 29–39, 2013.

[6] S. Khoje and S. Bodhe, "Comparative performance evaluation of size metrics and classifiers in computer vision based automatic mango grading," *International Journal of Computer Applications*, vol. 61, no. 9, 2013.

[7] C. S. Nandi, B. Tudu, and C. Koley, "A machine vision technique for grading of harvested mangoes based on maturity and quality," *IEEE Sensors Journal*, vol. 16, no. 16, 2016.

[8] T. U. Ganiron, "Size properties of mangoes using image analysis," *International Journal of Bio-Science and Bio-Technology*, vol. 6, no. 2, pp. 31–42, 2014.

[9] K. Schulze *et al.*, "Development and assessment of different modeling approaches for size-mass estimation of mango fruits," *Computers and Electronics in Agriculture*, vol. 114, 2015.

[10] F. S. A. Sa'ad, M. F. Ibrahim, A. M. Shakaff, A. Zakaria, and M. Z. Abdullah, "Shape and weight grading of mangoes using visible imaging," *Computers and Electronics in Agriculture*, vol. 115, pp. 51–56, 2015.

[11] D. J. Williams and M. Shah, "A fast algorithm for active contours and curvature estimation," *CVGIP: Image Understanding*, vol. 55, no. 1, pp. 14–26, 1992.

[12] T. N. Phan, V. Kuch, and L. W. Lehnert, "Land cover classification using google earth engine and random forest classifier—The role of image composition," *Remote Sensing*, vol. 12, no. 15, 2411, 2020.

[13] Y. Chen, W. Du, X. Duan, Y. Ma, and H. Zhang, "Squeeze-and-Excitation convolutional neural network for classification of malignant and benign lung nodules," *Journal of Advances in Information Technology*, vol. 12, no. 2, 2021.

[14] S. Bunrit, N. Kerdprasop, and K. Kerdprasop, "Improving the representation of CNN based features by autoencoder for a task of construction material image classification," *Journal of Advances in Information Technology*, vol. 11, no. 4, 2020.

[15] N. M. Trieu and N. T. Thinh, "The anthropometric measurement of nasal landmark locations by digital 2D photogrammetry using the convolutional neural network," *Diagnostics*, vol. 13, no. 5, 891, 2023.

[16] A. S. Chakraborty, T. Choudhury, R. Sille, C. Dutta, and B. K. Dewangan, "Multi-view deep CNN for automated target recognition and classification of synthetic aperture radar image," *Journal of Advances in Information Technology*, vol. 13, no. 5, pp. 413–422, October 2022.

[17] N. M. Trieu and N. T. Thinh, "A study of combining KNN and ANN for classifying dragon fruits automatically," *Journal of Image and Graphics*, vol. 10, no. 1, pp. 28–35, 2022.