Fast and Efficient Cavendish Banana Grade Classification using Random Forest Classifier with Synthetic Minority Oversampling Technique

Sujitra Arwatchananukul, Rattapon Saengrayap, Saowapa Chaiwong, Nattapol Aunsri*, *Member, IAENG.*

***Abstract*—Cavendish banana is an important export product of many countries, while postharvest banana classification also impacts plantation income. The quality inspection standard classifies Cavendish banana into three groups based on fruit size as large, medium and small. Most factories classify bananas manually which is time-consuming and also prone to errors. To expedite and enhance this process, a new fast and reliable method is proposed for classifying Cavendish banana gradings. The dataset contained 415 records, which were divided into five classes: L1, L2, L3, L4, and Reject. This study proposes employing a Synthetic Minority Oversampling Technique (SMOTE) to address imbalanced data by synthesizing the minority class to generate new minority sample data. Then a Random Forest classifier based on fruit length and fruit diameter parameters is implemented to efficiently classify Cavendish banana grading. Performance of the proposed method was assessed using 10-fold cross-validation. Compared to other machine learning techniques, results revealed that our approach delivered excellent performance with highest classification accuracy of 97.88%. Accuracy of 95.38% was achieved for an unseen testing dataset, illustrating superior performance to the other methods. We finally demonstrate a web application of the Cavendish banana grade classification system that was developed from the proposed model.**

***Index Terms*—Cavendish banana, Random Forest, data classification, banana grading, prediction model, SMOTE, k-fold Cross-Validation**

I. Introduction

For banana production, Thailand is ranked 20th globally and 6th in Asia with annual output of around 1 million tons [1]. Chiang Rai Province in Northern Thailand has large Cavendish banana plantations and exports the product mostly to China. The main banana export varieties are Musa (AAA group) 'Klui Hom Thong' (Gros Michel banana) and 'Klui Hom Khieo' (Cavendish banana) as well as Musa (AA group) 'Klui Khai' [2]. A famous banana variety for both domestic and export markets is Gros Michel banana, particularly for modern trade and retail markets. However, Cavendish banana plantations in Chiang Rai have gradually increased to around 200 hectares, with 90% of the product exported to the Chinese market. Recently, the lack of trained workers has been the main factor impacting postharvest handling of commercial bananas for sorting, selecting and sizing by visual appearance. The main typical fruit characteristics for commercial Cavendish for export are considered such as defect, fruit size, classifies the fruit shape, maturity and disease. The Philippine National Standard (PNS/BAFPS 64:2008) [3], classifies the fruit size of Cavendish banana into three size groups (large, medium, small) in terms of length and diameter. Due to the commercially exported grade, the large size was considered for data analysis in this report. In Thailand, most agricultural business models of banana production and supply chain involve contract farming (smallholder farmers) who have limited technology, tools and poor infrastructure development to control the uniformity of fruit quality. After collecting banana fruit, time is required at the packing house for grading and sorting, and this increases postharvest handling. To date, very limited banana grade classification has been conducted using machine learning to promote efficient Cavendish standards classification for export, particularly from the fruit size. Also, no previous study has reported on the machine learning approach that is low cost, non-destructive, and fast for banana fruit classification in Thailand.

Nowadays, emerging technologies as advanced engineering and computational methods, especially machine learning techniques, play a crucial role in improving quality of life through healthcare, the environment, document processing, linguistic, agriculture, food, and ameliorating postharvest problems [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16]. Research by [17] developed a technique for classifying ripeness from banana images into three types as unripe, ripe, and overripe. The system was based on a histogram of RGB value components, and an artificial neural network (ANN) was used as a classifier, achieving accuracy of 70%. In addition, [18] created an intelligent identification system for grading banana fruit. Two groups containing healthy and defective fruits were considered. An artificial neural network was utilized as a classifier by selecting extracted banana image features composed of RGB to gray conversion, image filtering, thresholding, edge detection, and morphological operation of dilation. Results showed the highest

Manuscript received July 2, 2021; revised November 2, 2021. This research was supported in part by Mae Fah Luang University, and in part by Integrated AgriTech Ecosystem Research Group (IATE), Mae Fah Luang University, Chiang Rai, Thailand. The publication of this paper was funded by Mae Fah Luang University.

S. Arwatchananukul is a lecturer at School of Information Technology and with the Integrated AgriTech Ecosystem Research Group (IATE), Mae Fah Luang University, Chiang Rai, 57100, Thailand. (e-mail: sujitra.arw@mfu.ac.th)

R. Saengrayap is an Assistant Professor of Food and Agricultural Engineering in the Postharvest Technology Programme at School of Agro-Industry and with the Integrated AgriTech Ecosystem Research Group (IATE), Mae Fah Luang University, Chiang Rai, 57100, Thailand. (e-mail: rattapon.sae@mfu.ac.th)

S. Chaiwong is an Assistant Professor of Postharvest Technology at School of Agro-Industry and with the Integrated AgriTech Ecosystem Research Group (IATE), Mae Fah Luang University, Chiang Rai, 57100, Thailand. (e-mail: saowapa@mfu.ac.th)

N. Aunsri is an Assistant Professor of Computer Engineering at School of Information Technology and with the Integrated AgriTech Ecosystem Research Group (IATE), Mae Fah Luang University, Chiang Rai, 57100, Thailand. (e-mail: nattapol.aun@mfu.ac.th)

* Author to whom correspondence should be addressed.

recognition rate of 97.00%. Moreover, [19] implemented an image preprocessing technique for classifying three cultivars as Cavendish, Lady Finger, and Pisang Awak. This study focused on types of banana because they are herbaceous plants that produce fruit all year round. Bananas are an important economic crop for domestic consumption and also for export. Scale-invariant shape analysis was used for prediction by a Bayesian network, and the performance was presented at an overall accuracy of 84%. In 2017, [20] described a model for banana ripeness classification. This is an important task in banana post-harvesting. A model was developed that utilized a fuzzy classifier to categorize the level of banana fruit into unripe, ripe, and overripe stages. Performance determined an average classification accuracy of 93.11%. Furthermore, [21] used various techniques such as k-Nearest Neighbor, Support Vector Machine, Decision Trees, Naive Bayes, and Discriminant Analysis to compare the classifying performance and grade of banana fruit ripening stage. Their proposed model delivered the best results by implementing a system to identify the ripening stage of bananas using an artificial neural network based on color, development of brown spots, and Tamura statistical texture features. A recognition rate of 97.75% was obtained. Moreover, Random Forest was used by [22] for modeling the detection of apple fruit diseases and classification. Color and textural features extracted from the fruit image were combined. Experimental results showed that the accuracy of disease classification using combined features was better than that obtained from a single feature. Due to its excellent performance, classification algorithms in machine learning has been implemented in many applications including agricultural problems, please see [23], [24], [25], [26], [27] for more examples.

This study presents a simple but very efficient tool derived from machine learning techniques for Cavendish banana classification. This tool can effectively control fruit quality at each farm before sending the banana fruit to the packing house. This Cavendish banana grade classification focused on the physical data of banana fruit. The standard for commercially exported banana grade comprises the criteria considered here. Therefore, we designed a tool for banana grading classification using Random Forest, a popular machine learning algorithm. The RF and a Synthetic Minority Oversampling Technique (SMOTE) were combined to classify banana grading based on fruit length and fruit diameter parameters. Furthermore, the web application implemented from the proposed model can be used in real-time to effectively assist in banana classification.

This paper is structured as follows. Section II presents the Cavendish banana dataset prepared and used in this study, while the proposed method is outlined in section III. Classification results and a detailed discussion are provided in section IV, with conclusions drawn and recommendations for further research suggested in section V.

II. Cavendish Banana Dataset

A total of 208 bunches of Cavendish banana (Musa (AAA group) 'Kluai Hom Khieo') were harvested at around 11 weeks after inflorescent emergence at the plantation site in Phaya Meng Rai District, Chiang Rai Province, Thailand (19°49′14.0”N 100°10′56.5”E) during March 2018 to August 2018. The third and fifth hands were selected for quality evaluation and cleaned with chlorinated water (150 ppm). Fruit diameter was measured using a Digitronic Vernier caliper (Moore & Wright MC MW110-15DBL 0-150 mm, UK) across two positions at the middle point, and averaged for the diameter of each finger (cm). Fruit length (cm) was measured along both sides of each finger using a measuring tape [28], [29]. The dataset contained 415 records, each of which was classified into one of five categories based on diameter and length: L1, L2, L3, L4, and Rejection by specialist collection. Moreover, some parts of the dataset included rejected classes as sunburn, insect infection, oversize, and over mature. Examples of these rejections are shown in Fig. 1.

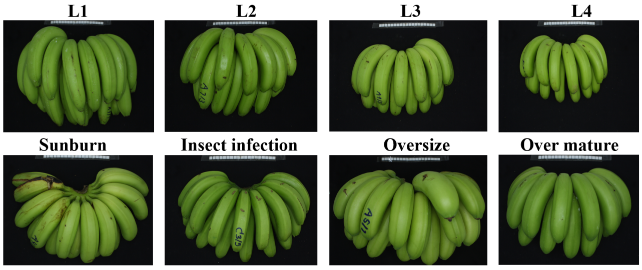


Fig. 1: Samples of the banana dataset.

III. Methodology

The proposed approach involved six crucial steps. First, the collected raw dataset of banana grading was loaded from the database. Second, since some class labels were imbalanced sets in the acquired data, we remedied this problem by implementing the SMOTE algorithm. To do this, minority class oversampling was improved by increasing the amount of data in this set. Details of SMOTE are presented in the following subsection. Third, the dataset was split into 70% as the training dataset and 30% as the testing dataset. Fourth, the number of trees and maximum depth of trees were tuned according to hyperparameter optimization, leading to an optimal Random Forest. Fifth, the prediction model was performed and then the proposed method was evaluated using 10-fold cross-validation. Finally, we applied the learned model to the testing dataset using the RapidMiner data mining tool. A block diagram of the proposed approach for Cavendish banana grade classification and prediction is illustrated in Fig. 2.

A. Synthetic Minority Over-sampling Technique (SMOTE)

Data mining process with imbalance dataset is one of the most difficult problems in machine learning in many applications. It is difficult to achieve overall accuracy where imbalanced data classification using machine learning methods [30], [31], [32]. Recently, several researchers publish their work on several techniques for classifying imbalanced datasets[33], [34], [35], [36], [37].

Data used in most practical classification problems are imbalanced. Here, the dataset also contained imbalanced data. The SMOTE algorithm was utilized to reduce the imbalanced data by synthesizing the minority class to produce new minority sample data. Specifically, SMOTE provides a new set of the data from the original oversampling data

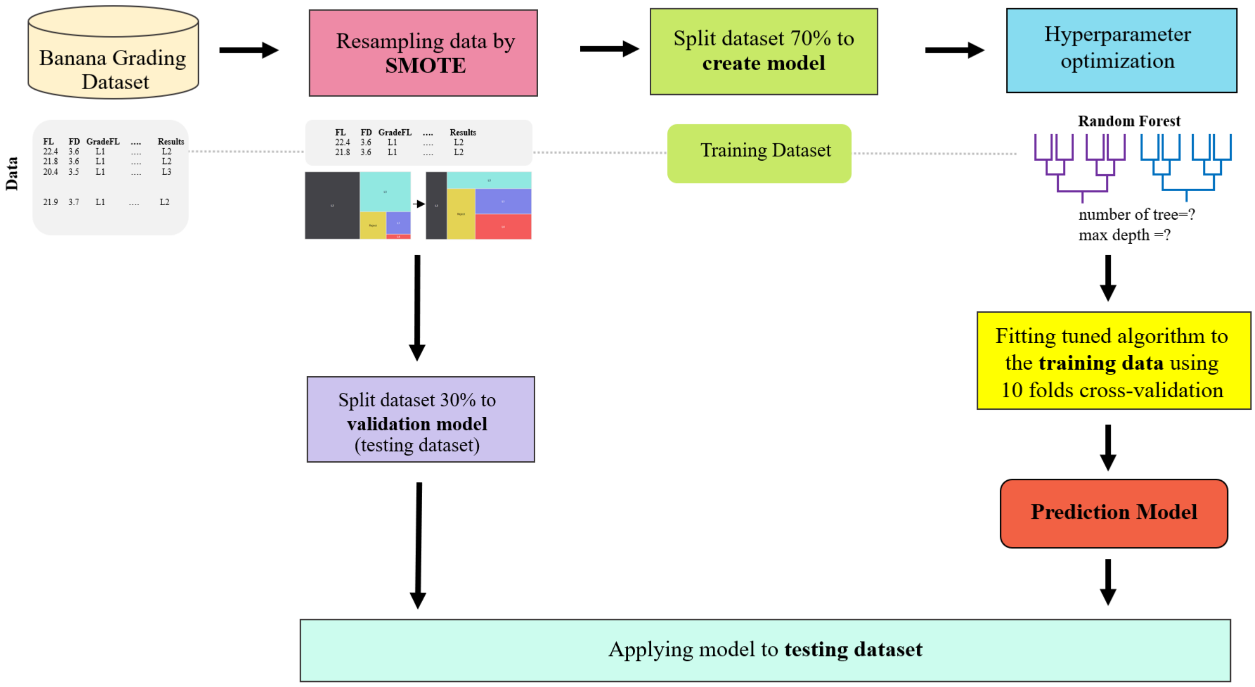

Banana Grading Dataset
Resampling data by SMOTE
Split dataset 70% to create model
Hyperparameter optimization
Data
Training Dataset
Random Forest
number of tree=?
max depth =?
Split dataset 30% to validation model (testing dataset)
Fitting tuned algorithm to the training data using 10 folds cross-validation
Prediction Model
Applying model to testing dataset

Fig. 2: Diagram for proposed methodology.

by searching for the k-nearest neighboring samples in each minority class and randomly selecting numbers of nearest neighbors. Therefore, each instance of the original minority samples is allowed to create new samples, and then integration between the original minority class samples and its neighboring samples provide a new balanced dataset [38].

B. Random Forest

Random Forest (RF) is a classifier that uses various models of many Decision Trees (DT) to achieve better prediction performance. The key idea of the RF algorithm is explained as follows: starting from the jth tree, a random vector $\boldsymbol{\theta}_j$ is created independently from the past random vectors $\boldsymbol{\theta}_1$, ...,$\boldsymbol{\theta}_{j-1}$ with the identical distribution; a tree at this stage is then grown using the training set and θ_j, resulting in a classifier $f(\mathbf{x}, \boldsymbol{\theta}_j)$, where $\mathbf{x}$ represents an input vector. To further illustrate the concept; in bagging, the random vector $\boldsymbol{\theta}$ is created as the counts in N boxes obtained from N darts thrown randomly at the boxes. Here, N is the number of samples in the training set, while for random split selection, the quantity $\boldsymbol{\theta}$ contains a number of independent random integers between 1 and J, where J is number of best splits. Tree construction dictates the nature and dimensionality of $\boldsymbol{\theta}$. Moreover, this model classifies new instances. Each decision tree provides a classification for input data; RF collects the classification result of each decision tree and then chooses the most voted prediction [39], [40].

The RF algorithm can be summarized as follows:

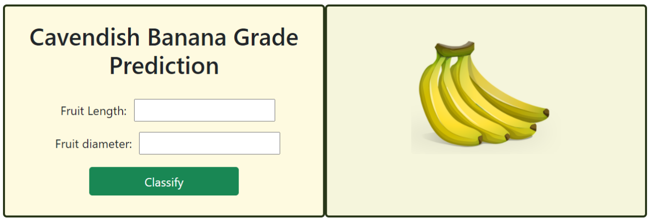
1) Initially, labels are assigned to all root nodes.
2) Assign N as a root node that searches a feature (F) from a random subset of the feature. Additionally, search a threshold value (T) that divides the samples assigned to N into two subsets on the left and right to highest direct label within these subsets.
3) Assign (F, T) to N
   - If subsets on the left and right are too small to be divided
     - Let child leaf nodes develop on the left and right of N
     - Assign leaves with the latest label in subset left and subset right, respectively.
   - Otherwise,
     - Let child nodes develop to the left and right of N
     - Set left subsets and right subsets to these nodes, respectively.
     - Repeat the process for N = root nodes on the left, and N = root nodes on the right for step 2.

C. Hyperparameter Optimization

The study by [41] suggested differentiation of the RF algorithm by tuning parameters to improve the accuracy and performance. Here, we optimized the model by correcting the hyperparameters to obtain the best classification performance. This was done by varying the number of decision trees, maximum depth of trees, confidence, majority voted prediction and the Gini Index criteria parameter for achieving the highest performance and accuracy. A set of parameters that provided the best results was used as the model for further evaluation.

D. Performance Evaluation Matric

The k-fold cross-validation technique was applied to estimate the skill of the model in assessing new data. Cross-validation is a well-known technique for evaluating the performance of classifiers and is typically used in machine



Cavendish Banana Grade Prediction

Fruit Length:

Fruit diameter:

Fig. 3: Screenshot samples of Cavendish banana grade prediction application.

learning performance evaluation. In this procedure, the data are separated into k-sub-datasets, one sub-dataset is stored as validation data for testing the model, while the remaining $k-1$ sub-datasets are used as training data. The inspection process is repeated k times (number of folds) [42], [43], i.e., the training and testing processes are repeated k times, and the average accuracy of these k repeated processes is considered as the accuracy of the classifier. Here, we used 10-fold cross-validation ($k = 10$) since this is the most common setting used to provide classification accuracy.

In predictive analytic, we present a confusion matrix of classification results from four different outcomes containing True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). The actual values are recorded in the columns and the predicted value labels are presented in the rows. The confusion matrix provides information to compute the recall, precision, and overall class prediction accuracy that evaluates the performance of the classifier [44], [45]. Definitions of these quantities are given as follows:

$$Precision = \frac{TP}{TP + FN} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

E. Web Application

After obtaining the best model described earlier, we developed a web application using HTML and Python language to grade Cavendish banana in real time. The system was then deployed on https://cavendish-banana-prediction.herokuapp.com. To use this system, users fill in fruit length and fruit diameter values as the classification inputs. The information is then processed and the prediction result is displayed. The proposed web application is demonstrated in Fig. 3.

IV. Experiments and Results

A. Data Characterization

Table I shows the banana classification results based on length (criterion I) and diameter (criterion II). The dataset's banana grading included 415 records. Fruit length and fruit diameter criteria were used to classify this dataset. The experts divided the banana fruit into five classes: class L1 (7.7%), class L2 (52.0%), class L3 (28.4%), class L4 (1.7%), and class Rejection (10.2%). The grading results showed the same trend with the diameter criterion since fruit diameter was the critical grading decision. Banana fruits with a diameter of less than 3.3 cm were rejected due to their insufficient size and also most of the rejected samples were sunburn, insect infection, oversize, and over mature.

B. Performance Comparison Among Classifiers

After collecting the data explained above, we applied the SMOTE algorithm for each minority class to oversampling in the dataset. After collecting the data explained above, we applied the SMOTE algorithm for each minority class to oversampling in the dataset. The majority class as L2 contained 216 records, and this algorithm was implemented to upsize the classes L1, L3, L4 and Rejection to obtain the same size which was 216 records. Furthermore, we split the data into 70% for the training dataset and 30% for the testing dataset to validate the classification model. For the classification model, we implemented the proposed method on Rapidminer GUI with several classifiers including our proposed method Random Forest (RF), Naive Bayes, k-Nearest Neighbor (kNN), Neural networks (NN) and Decision Trees (DT) to compare the performance of our proposed method against the others.

Illustrated in Figure 4 the results from an experiment for performance evaluation of all classifiers. Figure 4b show the results for Naive Bayes model, with the outcome of the confusion matrix indicating that the accuracy is 91.39%. Furthermore, Fig. 4c shows an experimental results form k-Nearest Neighbor model, with the confusion matrix suggesting an accuracy of 94.17%. In addition, Fig. 4d shows an experiment results using Neural Networks model that presents 94.57% accuracy in the confusion matrix. In addition, Decision Tree model delivers the results in the confusion matrix presented in Fig. 4e, the accuracy of 96.83% was achieved. Finally, experimental results showed that Random Forest delivered superior performance to classify the Cavendish banana grade dataset based on fruit length and fruit diameter. As previously mentioned, model validation was performed using 10-fold cross-validation, this can be seen in Fig. 4a. In Table II, it summarizes the accuracy, precision, and recall of all classifiers, random forest offers the best performance among the others. From these results, we clearly see that random forest is able to perform Cavendish banana grade classification at this stage.

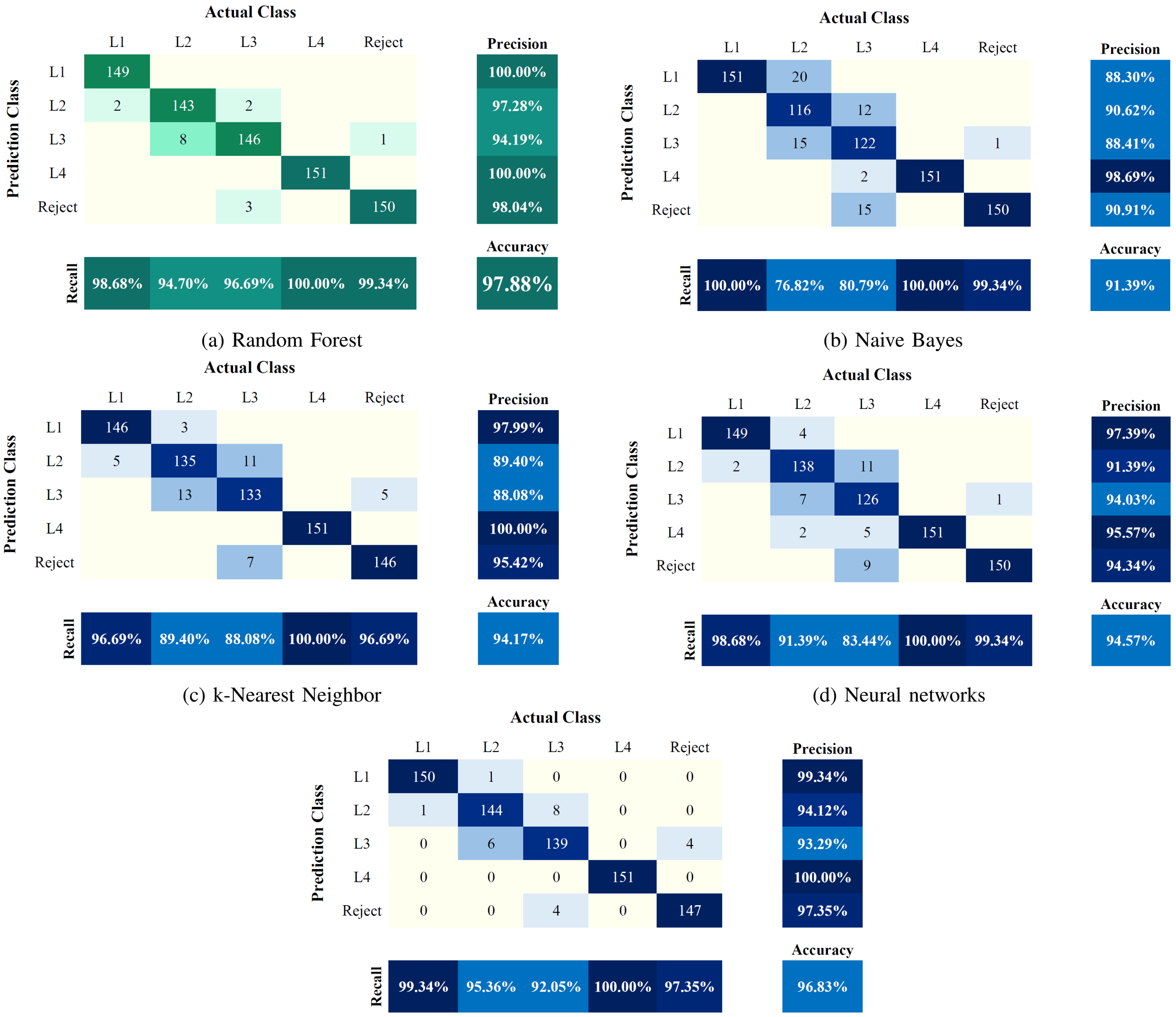
In the experiment, we fine-tuned the number of decision trees for the RF model set to 20, 40, 60, 80 and 100. A maximum depth of trees was also selected at 2, 4, 6 and 8. Accuracy of the RF classifier model for each case is shown in Table III. Moreover, our RF algorithm sought for the optimal parameters to predict class labels. The majority voted prediction and the Gini Index criteria were used to classify labels and achieved higher accuracy.

C. Model Evaluation With Unseen Data

To further investigate the performance of the method, the proposed prediction model was utilized for the testing dataset. Figure 5 showed the result of the confusion matrix of the validation models applying the prediction model to the test set in different classification models. Results showed that

TABLE I: Banana classification standard and classification result based on length and diameter criteria.

Class	Criteria		Amount	Percentage of banana based on criteria
	I (cm)	II (cm)		
L1	20.0	4.0	32	7.7 %
L2	18.1-20.0	3.6-4.0	216	52.0 %
L3	16.1-18.0	3.3-3.5	118	28.4 %
L4	14.1-16.0	2.9-3.2	7	1.7 %
Rejection	oversize	<3.3	42	10.1 %



Prediction Class	Actual Class					
	L1	L2	L3	L4	Reject	Precision
L1	149					100.00%
L2	2	143	2			97.28%
L3		8	146		1	94.19%
L4				151		100.00%
Reject			3		150	98.04%
Recall	98.68%	94.70%	96.69%	100.00%	99.34%	Accuracy 97.88%

(a) Random Forest

Prediction Class	Actual Class					
	L1	L2	L3	L4	Reject	Precision
L1	151	20				88.30%
L2		116	12			90.62%
L3		15	122		1	88.41%
L4			2	151		98.69%
Reject			15		150	90.91%
Recall	100.00%	76.82%	80.79%	100.00%	99.34%	Accuracy 91.39%

(b) Naive Bayes

Prediction Class	Actual Class					
	L1	L2	L3	L4	Reject	Precision
L1	146	3				97.99%
L2	5	135	11			89.40%
L3		13	133		5	88.08%
L4				151		100.00%
Reject			7		146	95.42%
Recall	96.69%	89.40%	88.08%	100.00%	96.69%	Accuracy 94.17%

(c) k-Nearest Neighbor

Prediction Class	Actual Class					
	L1	L2	L3	L4	Reject	Precision
L1	149	4				97.39%
L2	2	138	11			91.39%
L3		7	126		1	94.03%
L4		2	5	151		95.57%
Reject			9		150	94.34%
Recall	98.68%	91.39%	83.44%	100.00%	99.34%	Accuracy 94.57%

(d) Neural networks

Prediction Class	Actual Class					
	L1	L2	L3	L4	Reject	Precision
L1	150	1	0	0	0	99.34%
L2	1	144	8	0	0	94.12%
L3	0	6	139	0	4	93.29%
L4	0	0	0	151	0	100.00%
Reject	0	0	4	0	147	97.35%
Recall	99.34%	95.36%	92.05%	100.00%	97.35%	Accuracy 96.83%

(e) Decision Trees

Fig. 4: Summary of the confusion matrix prediction in different classification models.

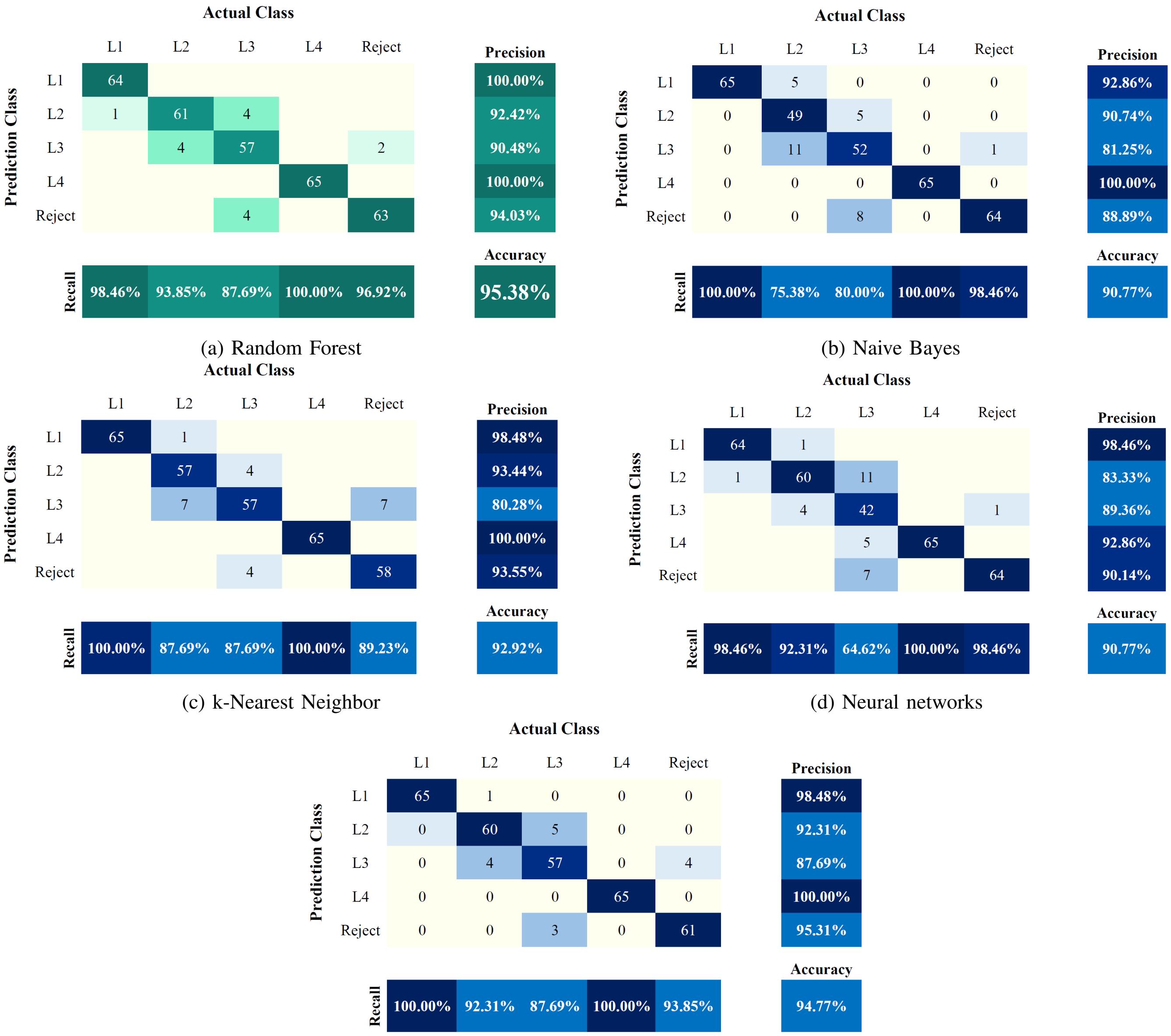
TABLE II: Comparison of prediction performance of different classifiers (%).

Classifier	Accuracy	Precision	Recall
Random Forest	97.88	98.04	97.88
Navie Bayes	91.39	92.06	91.38
k-Nearest Neighbor	94.17	94.48	94.15
Neural Networks	94.57	94.91	94.56
Decision Tree	96.83	97.01	96.79

the proposed method produced excellent classification accuracy of 95.38%. The confusion matrix for the testing stage is presented in Fig. 5a. Results confirmed good performance and reliability of the proposed method for Cavendish banana size classification. According to the accuracy of the classifiers, Decision Trees, k-Nearest Neighbor, Neural networks and Naive Bayes offers lower performance, respectively.

D. Web Application

After obtaining the best model, we finally implemented the model on a web application system. The experiment classified Cavendish banana grade in real time using the new fruit length and diameter features. Classification results



Actual Class / Prediction Class

	L1	L2	L3	L4	Reject	Precision
L1	64					100.00%
L2	1	61	4			92.42%
L3		4	57		2	90.48%
L4				65		100.00%
Reject			4		63	94.03%
						Accuracy
Recall	98.46%	93.85%	87.69%	100.00%	96.92%	95.38%

(a) Random Forest

Actual Class / Prediction Class

	L1	L2	L3	L4	Reject	Precision
L1	65	5	0	0	0	92.86%
L2	0	49	5	0	0	90.74%
L3	0	11	52	0	1	81.25%
L4	0	0	0	65	0	100.00%
Reject	0	0	8	0	64	88.89%
						Accuracy
Recall	100.00%	75.38%	80.00%	100.00%	98.46%	90.77%

(b) Naive Bayes

Actual Class / Prediction Class

	L1	L2	L3	L4	Reject	Precision
L1	65	1				98.48%
L2		57	4			93.44%
L3		7	57		7	80.28%
L4				65		100.00%
Reject			4		58	93.55%
						Accuracy
Recall	100.00%	87.69%	87.69%	100.00%	89.23%	92.92%

(c) k-Nearest Neighbor

Actual Class / Prediction Class

	L1	L2	L3	L4	Reject	Precision
L1	64	1				98.46%
L2	1	60	11			83.33%
L3		4	42		1	89.36%
L4			5	65		92.86%
Reject			7		64	90.14%
						Accuracy
Recall	98.46%	92.31%	64.62%	100.00%	98.46%	90.77%

(d) Neural networks

Actual Class / Prediction Class

	L1	L2	L3	L4	Reject	Precision
L1	65	1	0	0	0	98.48%
L2	0	60	5	0	0	92.31%
L3	0	4	57	0	4	87.69%
L4	0	0	0	65	0	100.00%
Reject	0	0	3	0	61	95.31%
						Accuracy
Recall	100.00%	92.31%	87.69%	100.00%	93.85%	94.77%

(e) Decision Trees

Fig. 5: Summary of confusion matrix of the validation model applying the prediction model to the test set.

TABLE III: Optimized parameters with the 10-fold cross-validation accuracy (%).

Iteration	Maximal depth	Number of trees	Accuracy (%)
1	2	20	62.81
2	4	20	82.24
3	6	20	97.75
4	8	20	97.09
5	2	40	64.28
6	4	40	79.61
7	6	40	97.75
8	8	40	96.82
9	2	60	63.30
10	4	60	79.74
11	6	60	97.88
12	8	60	96.96
13	2	80	65.84
14	4	80	81.73
15	6	80	97.88
16	8	80	97.35
17	2	100	58.04
18	4	100	79.74
19	6	100	97.88
20	8	100	97.22

showed the detail and picture of the predicted banana grade. Figure 6 presents a sample from a web application.

E. User Satisfaction

Web application was evaluated by 20 agriculturists. Figure 7 shows the results of the assessment in three areas of the question. Firstly, the question focuses on ”Is the web application easy to use?”. Four users scored Undecided, six users scored Agree, and ten users scored Strongly Agree. Secondly, the question focuses on the usefulness of the web application in helping users to classify the Cavendish banana grade, two users rated Disagree, three users rated Undecided, nine users rated Agree, and six users rated Strongly Agree. Finally, the users were asked whether or not recommending this application to their friends. The recommendation resulted that two users Disagreed, four users Undecided, ten users Agreed, and four users Strongly Agreed.

Table IV summarizes the statistics including means and standard deviations of the above three satisfactory results.

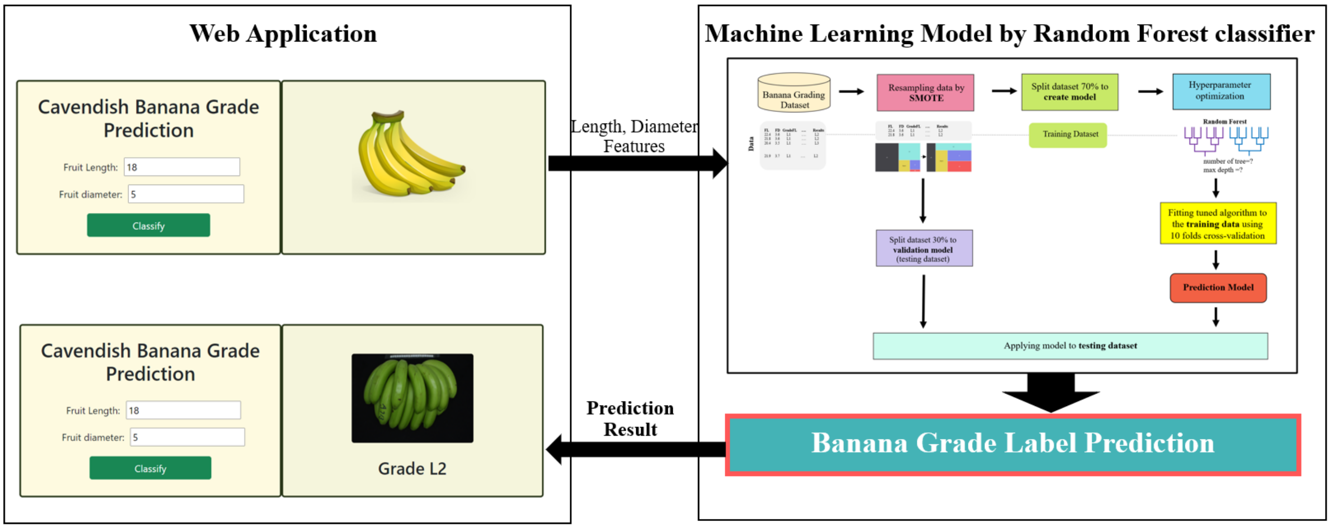

Web Application
Cavendish Banana Grade Prediction
Fruit Length: 18
Fruit diameter: 5
Classify
Cavendish Banana Grade Prediction
Fruit Length: 18
Fruit diameter: 5
Classify
Grade L2
Length, Diameter Features
Prediction Result
Machine Learning Model by Random Forest classifier
Banana Grading Dataset
Data
Resampling data by SMOTE
Split dataset 70% to create model
Training Dataset
Hyperparameter optimization
Random Forest
number of tree=?
max depth =?
Split dataset 30% to validation model (testing dataset)
Fitting tuned algorithm to the training data using 10 folds cross-validation
Prediction Model
Applying model to testing dataset
Banana Grade Label Prediction

Fig. 6: Proposed Cavendish banana prediction web application system.

The first row shows that the simplicity and ease of use score obtained from the users has a strongly satisfied level with a mean of 4.3 out of 5.0 and standard deviation of 0.78. Following the ”Usefulness” satisfaction, paying attention to the program can make it easier to grade Cavendish bananas, the responders were rather satisfied with a mean of 3.95 out of 5.0, and standard deviation of 0.92. Finally, ”Recommendation to others” satisfaction has a mean of 3.98 out of 5.0 with standard deviation of 0.87. The users would recommend the program to their friends. From these statistics, we observe that our proposed web application is preferable for the users to utilize it in their work. Some qualitative feedbacks and suggestions for improvement will be discussed as the followings.

After asking the users to use the web application for a period of time, we then collected the comments from all the users for further improvement. The summary of the suggestions from agriculturist retrieving from the survey are shown in Table V. They mentioned that the user interface was easy to use, but the design was unattractive. Therefore, the application should contain more colorful to the typography to help the application standing out and be remembered. In Content, they recommended expanding on the background of the Cavendish bananas’ characteristics as well as instructing the user on how to use the application and how to classify Cavendish bananas for each grade, such as L1, L2, L3, L4, and Rejection. Furthermore, the application should display a clear unit of measurement in centimeters (cm), and the application should inform the minimum and maximum values for the data to be filled in. Lastly, the agriculturists acknowledged that while this program makes it easier to classify the Cavendish banana grade. However, as a result of this, the program display time is extremely slow in terms of image performance. Therefore, for the recommendation, there should be more explanations of the text material on the analyzed level to make it easier to understand the results of the Cavendish banana grade classification, and the image result should be clearer and diversified.

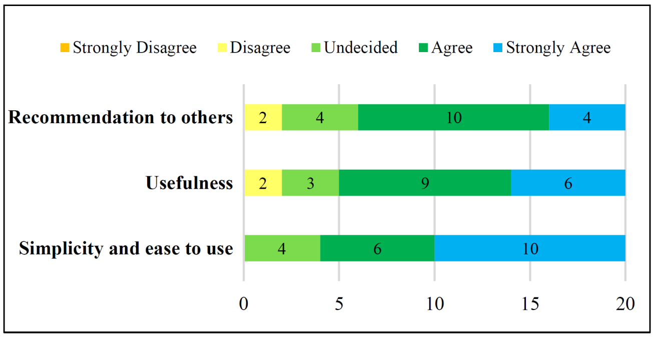


Fig. 7: The result of the user satisfaction survey.

TABLE IV: User satisfaction level.

Satisfaction	Mean	Standard Deviation
Recommendation to others	3.8	0.87
Usefulness	3.95	0.92
Simplicity and ease to use	4.3	0.78

V. Conclusion and Future work

A new model was presented for classifying Cavendish banana size into five classes as L1, L2, L3, L4, and Rejection. Our model used fruit length and fruit diameter to formulate features in the RF classifier with Synthetic Minority Oversampling Technique. The SMOTE algorithm was used for upsampling data size of the minority class and then fine-tuning the hyperparameter for number of trees at 60, 80 and 100. Maximum depth of trees was 6, and majority voted prediction was employed using the Gini Index criteria and the confidence parameter of 0.1. These hyperparameters provided the highest accuracy. A 10-fold cross-validation was performed to evaluate and validate the model. Experimental results, with 10-fold cross-validation, revealed the highest prediction accuracy of 97.88%, with 95.38% for the unseen testing dataset, portraying excellent classification capability of the proposed method for Cavendish banana size classification.

Finally, we showed a prototype web application that can be used in real time for Cavendish banana grading. This web application only requires the input of two features as fruit

TABLE V: The summary of the agriculturist suggestions from the survey.

Category of feedback	**Feedback / Suggestions** **(+ for positive feedback, - for negative feedback, and * for suggestion)**
User Interface	+ The web application is simple to use. - The web application design is not particularly attractive. * Should add more colors or fonts to help the application stand out and be remembered.
Content of Application	* Should include a unit of measurement in centimeter (cm). * Should explain the background of Cavendish bananas characteristics a little more. * Should inform the information on how Cavendish bananas are classified for each grade, such as L1, L2, L3, L4, and Rejection. * The minimum and maximum values for the data to be filled in should be notified. * It is more convenient to take a photo with a mobile phone and then classify it by image.
Result of Classification	+ This program makes it easier to classify Cavendish bananas. - Too slow to display the Cavendish image results. * There should be more explanations of the text information on the analyzed level to make it easier to understand the results of the Cavendish banana grade classification. * The image result of Cavendish classification should be more clear and diversified.

length and fruit diameter that have already been collected in the workplace. In future work, we will further investigate Cavendish banana grading by using an image classification model, which will be used in mobile applications. It is more practical than using web applications, allowing larger utilization of this work in community.

ACKNOWLEDGEMENTS

The authors acknowledge the facility support from Phaya Meng Rai Agriculture Limited Partnership for providing the experimental site and banana trees.

REFERENCES

[1] Food and Agriculture Organization of the United Nation, *Banana*, 2020. [Online]. Available: http://www.fao.org/faostat/en/#data/QC

[2] Department of Agriculture Thailand, *Information centre. Banana market review*, 2014.

[3] Philippine National Standard, *PNS/BAFPS 64:2008 Fresh Fruits-Banana*, 2008. [Online]. Available: http://www.bafs.da.gov.ph/bafs_admin/admin_page/pns_file/2021-02-24-PNS-BAFPS%2064-2008-Banana.pdf

[4] S. Hunta, T. Yooyativong, and N. Aunsri, "A novel integrated action crossing method for drug-drug interaction prediction in non-communicable diseases," *Computer Methods and Programs in Biomedicine*, vol. 163, pp. 183 – 193, 2018.

[5] N. Chumrit, C. Weangwan, and N. Aunsri, "Ecg-based arrhythmia detection using average energy and zero-crossing features with support vector machine," in *2020 - 5th International Conference on Information Technology (InCIT)*, 2020, pp. 282–287.

[6] J. de la Torre, J. Marin, S. Ilarri, and J. J. Marin, "Applying machine learning for healthcare: A case study on cervical pain assessment with motion capture," *Applied Sciences*, vol. 10, no. 17, 2020.

[7] S. Hunta, N. Aunsri, and T. Yooyativong, "Drug-drug interactions prediction from enzyme action crossing through machine learning approaches," in *Proceedings of Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), 2015 12th International Conference*, June 2015.

[8] C.-C. Le, P. W. C. Prasad, A. Alsadoon, L. Pham, and A. Elchouemi, "Text classification: Naïve bayes classifier with sentiment lexicon," *IAENG International Journal of Computer Science*, vol. 46, no. 2, pp. 141–148, 2019.

[9] I. K. Izzah and A. S. Girsang, "Modified tf-assoc term weighting method for text classification on news dataset from twitter," *IAENG International Journal of Computer Science*, vol. 48, no. 1, pp. 142–151, 2021.

[10] Y.-C. Liang, Y. Maimury, A. H.-L. Chen, and J. R. C. Juarez, "Machine learning-based prediction of air quality," *Applied Sciences*, vol. 10, no. 24, 2020. [Online]. Available: https://www.mdpi.com/2076-3417/10/24/9151

[11] S. Saenmuang and N. Aunsri, "A new spinach respiratory prediction method using particle filtering approach," *IEEE Access*, vol. 7, pp. 131 559–131 566, 2019.

[12] L. Jiang, B. Qiu, X. Liu, C. Huang, and K. Lin, "Deepfood: Food image analysis and dietary assessment via deep model," *IEEE Access*, vol. 8, pp. 47 477–47 489, 2020.

[13] S. W. Sidehabi, A. Suyuti, I. S. Areni, and I. Nurtanio, "Classification on passion fruit's ripeness using k-means clustering and artificial neural network," in *2018 International Conference on Information and Communications Technology (ICOIACT)*, 2018, pp. 304–309.

[14] H. Kinjo, N. Oshiro, and S. C. Duong, "Fruit maturity detection using neural network and an odor sensor: Toward a quick detection," in *2015 10th Asian Control Conference (ASCC)*, 2015, pp. 1–4.

[15] J. Naranjo-Torres, M. Mora, R. Hernández-García, R. J. Barrientos, C. Fredes, and A. Valenzuela, "A review of convolutional neural network applied to fruit image processing," *Applied Sciences*, vol. 10, no. 10, 2020.

[16] A. Clark and J. McKechnie, "Detecting banana plantations in the wet tropics, australia, using aerial photography and u-net," *Applied Sciences*, vol. 10, no. 6, 2020.

[17] H. Saad, Ahmad Puad Ismail, Noriza Othman, Mohamad Huzaimy Jusoh, Nani fadzlina Naim, and Nur Azam Ahmad, "Recognizing the ripeness of bananas using artificial neural network based on histogram approach," in *2009 IEEE International Conference on Signal and Image Processing Applications*, 2009, pp. 536–541.

[18] E. Olaniyi, O. Oyedotun, and A. Khashman, "Intelligent grading system for banana fruit using neural network arbitration," *Journal of Food Process Engineering*, vol. 40, pp. n/a–n/a, 01 2016.

[19] K. Dittakan, N. Theera-Ampornpunt, W. Witthayarat, S. Hinnoy, S. Klaiwan, and T. Pratheep, "Banana cultivar classification using scale invariant shape analysis," in *2017 2nd International Conference on Information Technology (INCIT)*, 2017, pp. 1–6.

[20] S. Marimuthu and S. M. M. Roomi, "Particle swarm optimized fuzzy model for the classification of banana ripeness," *IEEE Sensors Journal*, vol. 17, no. 15, pp. 4903–4915, 2017.

[21] F. M. A. Mazen and A. A. Nashat, "Ripeness classification of bananas using an artificial neural network," *Arabian Journal for Science and Engineering*, vol. 44, pp. 6901–6910, 2019.

[22] B. J. Samajpati and S. D. Degadwala, "Hybrid approach for apple fruit diseases detection and classification using random forest classifier," in *2016 International Conference on Communication and Signal Processing (ICCSP)*, 2016, pp. 1015–1019.

[23] L. F. Santos Pereira, S. Barbon, N. A. Valous, and D. F. Barbin, "Predicting the ripening of papaya fruit with digital imaging and random forests," *Computers and Electronics in Agriculture*, vol. 145, pp. 76–82, 2018.

[24] A. Jahanbakhshi, M. Momeny, M. Mahmoudi, and Y.-D. Zhang, "Classification of sour lemons based on apparent defects using stochastic pooling mechanism in deep convolutional neural networks," *Scientia Horticulturae*, vol. 263, p. 109133, 2020.

[25] H. Azarmdel, A. Jahanbakhshi, S. S. Mohtasebi, and A. R. Muñoz, "Evaluation of image processing technique as an expert system in mulberry fruit grading based on ripeness level using artificial neural networks (anns) and support vector machine (svm)," *Postharvest Biology and Technology*, vol. 166, p. 111201, 2020.

[26] M. Momeny, A. Jahanbakhshi, K. Jafarnezhad, and Y.-D. Zhang, "Accurate classification of cherry fruit using deep cnn based on hybrid pooling approach," *Postharvest Biology and Technology*, vol. 166, p. 111204, 2020.

[27] M. N. Ribeiro, I. A. Carvalho, G. A. Fonseca, R. C. Lago, L. C. Rocha, D. D. Ferreira, E. V. Vilas Boas, and A. C. Pinheiro, "Quality control of fresh strawberries by a random forest model," *Journal of the Science of Food and Agriculture*, vol. n/a, no. n/a.

[28] S. Chaiwong, R. Saengrayap, and C. Prahsarn, "Effects of different materials for banana bunch covers during winter in thailand," *Acta horticulturae*, vol. 1245, pp. 21–28, 2019.

[29] R. Saengrayap, S. Chaiwong, and J. Rattanakaran, "Application of artificial neural network for'cavendish'banana maturity and chilling injury prediction," *Acta horticulturae*, vol. 1245, pp. 29–34, 2019.

[30] L. Gao, P. Lu, and Y. Ren, "A deep learning approach for imbalanced crash data in predicting highway-rail grade crossings accidents," *Reliability Engineering & System Safety*, vol. 216, p. 108019, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0951832021005287

[31] T.-T. Wong and H.-C. Tsai, "Multinomial naïve bayesian classifier with generalized dirichlet priors for high-dimensional imbalanced data," *Knowledge-Based Systems*, vol. 228, p. 107288, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0950705121005505

[32] J. Liu, "A minority oversampling approach for fault detection with heterogeneous imbalanced data," *Expert Systems with Applications*, vol. 184, p. 115492, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417421009027

[33] K. Savetratanakaree, K. Sookhanaphibarn, S. Intakosum, and R. Thawonmas, "Borderline over-sampling in feature space for learning algorithms in imbalanced data environments," *IAENG International Journal of Computer Science*, vol. 43, no. 3, p. 363–373, 2016.

[34] M. Y. Arafat, S. Hoque, S. Xu, and D. M. Farid, "Machine learning for mining imbalanced data," *IAENG International Journal of Computer Science*, vol. 46, no. 2, pp. 332–348, 2019.

[35] C. Teck Kai and C. S. Chin, "Health stages diagnostics of underwater thruster using sound features with imbalanced dataset," *Neural Computing and Applications*, vol. 31, 10 2019.

[36] S. Sarkar, A. Pramanik, J. Maiti, and G. Reniers, "Predicting and analyzing injury severity: A machine learning-based approach using class-imbalanced proactive and reactive data," *Safety Science*, vol. 125, p. 104616, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925753520300138

[37] S. Bej, K. Schulz, P. Srivastava, M. Wolfien, and O. Wolkenhauer, "A multi-schematic classifier-independent oversampling approach for imbalanced datasets," *IEEE Access*, vol. 9, pp. 123 358–123 374, 2021.

[38] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[39] Bauer, Kohavi, Breiman, T. K., Ho, Kleinberg, Freund, Bartlett, and Lee, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, 1997.

[40] RapidMiner, *Random Forest (Concurrency)*, 2020. [Online]. Available: https://docs.rapidminer.com/latest/studio/operators/modeling/predictive/trees/parallel_random_forest.html

[41] A. Cuzzocrea, S. L. Francis, and M. M. Gaber, "An information-theoretic approach for setting the optimal number of decision trees in random forests," in *2013 IEEE International Conference on Systems, Man, and Cybernetics*, 2013, pp. 1013–1019.

[42] T. Fushiki, "Estimation of prediction error by using k-fold cross-validation," *Statistics and Computing*, vol. 21, no. 2, pp. 137–146, 2011.

[43] J. D. Rodriguez, A. Perez, and J. A. Lozano, "Sensitivity analysis of k-fold cross validation in prediction error estimation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 3, pp. 569–575, 2009.

[44] D. Powers, "Evaluation: From precision, recall and f-factor to roc, informedness, markedness & correlation," *Mach. Learn. Technol.*, vol. 2, Jan 2008.

[45] A. Tharwat, "Classification assessment methods," *Applied Computing and Informatics*, vol. 17, no. 1, pp. 168–192, Jan 2020.

Sujitra Arwatchananukul received the B.S. and M.S. degrees in Computer Science from Chiang Mai University, Thailand, in 2004 and 2008 respectively and Ph.D. degree in Computer Engineering from Yunnan University, China.

In 2016, she joined the School of Information Technology, Mae Fah Luang University, as a lecturer. See is also with the Integrated AgriTech Ecosystem Research Group (IATE), Mae Fah Luang University, Chiang Rai, Thailand. Her current research interests include Data Science, Machine Learning, Data Analysis, Image Processing, Software engineering, Algorithms and Database Management system.

Rattapon Saengrayap obtained a B.Sc. degree in Fishery from Kasetsart University in 2006. He obtained his M.E. and D.Eng. degrees in Food Engineering from King Mongkut's University of Technology Thonburi - Thailand, in 2008 and 2014, respectively. He is currently an Assistant Professor of Food and Agricultural Engineering in the Postharvest Technology Programme, School of Agro-Industry, Mae Fah Luang University, Thailand. He is also with the Integrated AgriTech Ecosystem Research Group (IATE), Mae Fah Luang University, Chiang Rai, Thailand. He has also served as consultant in various government's entrepreneurial boost-up projects. His research of interests include non-destructive evaluation, computer vision, machine learning and coffee processing.

Saowapa Chaiwong obtained the B.Sc. degree and M.Sc. degree in Agriculture from Chiang Mai University in 2000 and Kasetsart University in 2004, Thailand. She received Ph.D. in Horticulture from Writtle College, University of Essex, The United Kingdom in 2016. She has been working as Assistant Professor of Postharvest Technology at School of Agro-Industry, Mae Fah Luang University, Thailand since May 2012. Her research interests involve pre-and postharvest technology, packaging distribution of fresh produce as well as cold chain management in fruit and vegetable supply chain management. Her research studies are focused on integrated knowledge among various fields such as horticulture, postharvest technology, material science, food engineering, computer science and information technology and logistics management. Now she is a leader of the Integrated AgriTech Ecosystem Research Group at Mae Fah Luang University.

Nattapol Aunsri received the B.Eng. degree and M.Eng. degree in Electrical Engineering from Khon Kaen University and Chulalongkorn University, Thailand in 1999 and 2003, respectively. He obtained M.Sc. degree in Applied Mathematics and Ph.D. degree in Mathematical Sciences from New Jersey Institute of Technology, Newark, NJ, in 2008 and 2014, respectively.

Since May 2017, he has been working as an Assistant Professor of Computer Engineering with the School of Information Technology, Mae Fah Luang University, Chiang Rai, Thailand. He is also with the Integrated AgriTech Ecosystem Research Group (IATE), Mae Fah Luang University, Chiang Rai, Thailand. His research interests include ocean acoustics, Bayesian estimation and filtering, signal processing, biomedical signal processing, drug-drug interactions, applications of machine learning techniques, and mathematical and statistical modeling.

Dr. Aunsri was a recipient of the Best Paper Awards in 21st International Symposium on Wireless Personal Multimedia Communications (WPMC) in 2018 and 2018 International Conference on Digital Arts, Media and Technology (ICDAMT). Dr. Aunsri is a member of APSIPA, ECTI, IEEE, and IAENG.