Machine Learning-Driven Classification of Thai Mango Groups using Clustering and Random Forest

Thananan Setajit[1], Chanathip Khamchan[1], Amornphong Naitip[1] ,Chuntawat Thongmee[1] ,
Kwanhatai Tanongjid[3], Sujitra Arwatchananukul[1,2]
[1]*School of Applied Digital Technology, Mae Fah Luang University, Chiang Rai, Thailand*
[2]*Integrated AgriTech Ecosystem Research Group, Mae Fah Luang University, Chiang Rai, Thailand*
[3]*Pakchong Research Station, Faculty of Agriculture, Kasetsart University, Pakchong, Nakhon Ratchasima, Thailand*
Email: 6431503028@lamduan.mfu.ac.th, 6431503131@lamduan.mfu.ac.th, 6431503071@lamduan.mfu.ac.th,
6431503009@lamduan.mfu.ac.th, kwanhatai.t@ku.ac.th, sujitra.arw@mfu.ac.th

***Abstract*—Classifying Thai mango varieties based on physical traits such as leaf shape and fruit size is challenging. This issue was addressed using K-means clustering of mango varieties followed by the Elbow, Silhouette, and Calinski-Harabasz methods to determine the optimal number of clusters. The clusters were validated and classified using a Random Forest model which achieved an accuracy of 91.26%. Results demonstrated significant improvements in the precision and efficiency of mango classification, benefiting both producers and consumers.**

***Keywords*—Mango Groups, Mango Physical Characteristics, K-Means, Silhouette Method, Hyperparameter**

I. Introduction

Mangoes were first brought to the country by Buddhist monks traveling from India in the 13th century and have long been a staple food in Thailand [1]. Mango trees flourish in tropical climates and Thailand is one of the largest global exporters. The significance of mangoes in Thai culture and economics remains substantial [2].

Mangoes are a popular and economically significant fruit in many parts of the world [3]. However, despite their widespread cultivation and consumption, differentiating between various mango groups based on physical characteristics remains challenging. Mangoes exhibit a variety of physical attributes such as the edge of the leaf, the shape of the fruit, the plate of the leaf, the base of the leaf, the tip of the leaf, and the shape of the leaf which vary between different groups. These variations often lead to confusion among farmers and consumers, making accurate classification crucial to ensure that the appropriate variety reaches the market and meets consumer expectations.

Accurate classification of different mango varieties is important in several areas including marketing, agricultural practices, consumption, and research. Misidentification leads to inefficiencies in the supply chain, negatively affecting both producers and consumers. Hence, establishing a reliable method for distinguishing mango varieties based on their physical attributes is essential.

The performances of the Decision Trees (DT) and Random Forest (RF) algorithms were previously analyzed and compared for nutrient classification using existing data [4]. Results showed that RF outperformed DT in handling complex, high-dimensional data for classification tasks. Reference [5] highlighted the widespread use of the RF algorithm in automatic fruit classification. By combining the outputs of multiple DT, RF improved accuracy and reduced bias, making it particularly effective for high-dimensional complex datasets. Previous studies demonstrated that RF provided high stability and precision in fruit classification. Recent research [6] underscored the effectiveness of integrating a hyperparameter-tuned RF with deep convoluted neural network (CNN) feature extraction. Deep CNNs are responsible for extracting specific features of plants or fruits, while hyperparameter tuning of the RF algorithm plays a crucial role in optimizing the classification process. By adjusting parameters such as the number of trees and model depth, Random Forest significantly enhanced the accuracy and reliability of fruit classification, making this combined approach especially powerful.Reference [7] applied machine learning techniques to classify mangoes into categories such as raw, cooked, and processed. This study used an RF model for mango classification due to its robustness, ability to handle large datasets, and high accuracy in classification tasks.

The potential of machine learning in agricultural applications was demonstrated, offering a practical solution to accurately classify mango groups. by first using clustering techniques to group mangoes based on their external physical characteristics, and then applying classification

979-8-3503-6630-3/24/$31.00 ©2024 IEEE

techniques to validate these groupings. Our results can be used to enhance the efficiency and quality of mango distribution.

II. Methodology

The research methodology used a dataset from the Pakchong Research Station of Kasetsart University [8], containing information on the physical characteristics of 92 mango varieties. The dataset included features such as species name, leaf edge, fruit shape, leaf plate, leaf base, leaf tip, leaf shape, and mango weight (in grams). As shown in Figure 1, the data preparation phase involved several techniques for cleansing the data including converting categorical columns to numerical values and grouping the weight data into four categories. Clustering was then performed using the K-means algorithm, with the optimal number of clusters determined using the Elbow, Calinski-Harabasz, and Silhouette Methods. The clustered mango groups were then classified using machine learning algorithms such as Decision Trees, Random Forest, Extra Trees, and Gradient Boosting. Following this, hyperparameter tuning was conducted to further enhance the model's performance. Finally, the model was evaluated using accuracy, precision, recall, and F1-score to assess the effectiveness of the classification approach

A. Dataset preparation

The dataset contained features such as species name, leaf edge, fruit shape, leaf plate, leaf base, leaf tip, leaf shape, and mango weight (grams). The data preparation process began by removing the 'species name' column, as it was not relevant to the classification task. The remaining columns, which provided information on the physical attributes of mangoes, were used for classification. Categorical data such as leaf edge, fruit shape, leaf plate, leaf base, leaf tip, and leaf shape were converted into numerical representations using the one-hot encoding technique.

The 'weight (grams)' column was transformed into a categorical variable by grouping the weights into four categories: 'Small fruit,' 'Medium fruit,' 'Large fruit,' and 'Extra large fruit.' The weight ranges for each category were defined as follows:

- Small fruit: 1-250 grams
- Medium fruit: 251-350 grams
- Large fruit: 351-450 grams
- Extra large fruit: more than 450 grams

The 'weight (grams)' column was then converted into numerical representations using one-hot encoding.

B. Clustering

This research focused on data clustering using three prominent techniques—Elbow Method, Silhouette Method, and Calinski-Harabasz Method—to identify the optimal number of clusters (K) within the dataset. Each method offers a unique approach to guide the selection of the most suitable cluster structure. K-means, the clustering algorithm used in this study, partitions the dataset into K distinct clusters by minimizing the variance within each cluster. The effectiveness of K-means largely depends on choosing the right value for K, which these methods helped to determine.

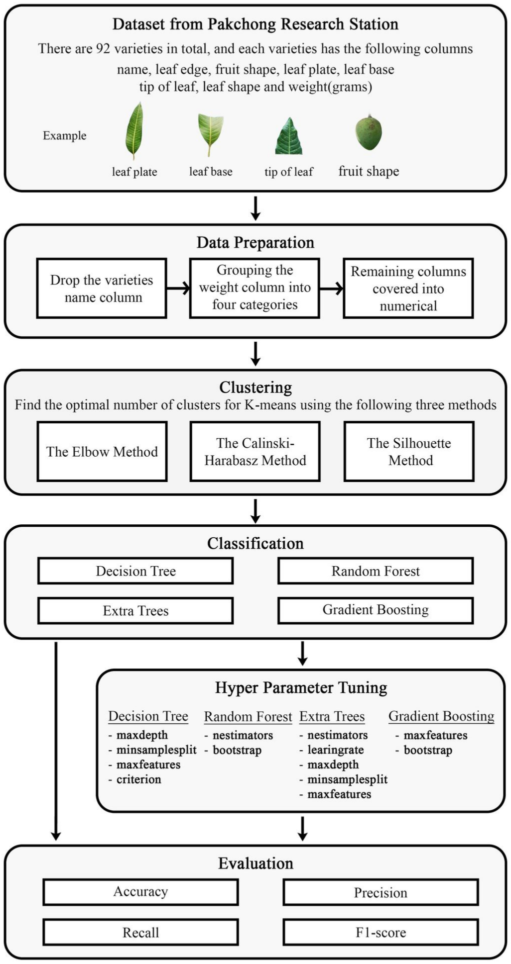

Fig. 1. An overview of the methodology block diagram

The Elbow Method [9], one of the most widely used techniques for selecting the optimal number of clusters, involves plotting the sum of squared distances (inertia) between each point and its assigned cluster center as a function of the number of clusters. The point where the curve begins to flatten, forming an "elbow," indicates the optimal number of clusters. The Elbow Method is simple and intuitive but the curve sometimes lacks a distinct elbow, requiring additional methods for confirmation.

The Silhouette Method [10] measures how similar an object is to its own cluster compared to the other clusters. It calculates the silhouette coefficient, which ranges from -1 to 1, with higher values indicating that an object is well-matched to its own cluster and poorly matched to neighboring clusters. The average silhouette coefficient across all samples is used to determine the optimal number of clusters. This method is particularly useful because it evaluates both the compactness within clusters and the separation between clusters, offering a more comprehensive assessment of clustering quality.

The Calinski-Harabasz Index [11], also known as the Variance Ratio Criterion, evaluates the ratio of between-cluster dispersion to within-cluster dispersion. A higher Calinski-Harabasz score indicates better-defined clusters. This method effectively balances cluster compactness with the distance between clusters, making it ideal for identifying well-separated and compact clusters. By combining these three methods, this research ensured a robust determination of the optimal number of clusters. Together, they provided a thorough analysis of both within-cluster compactness and between-cluster separation, ultimately guiding the selection of the most appropriate cluster structure for the dataset.

C. *Classification*

The classifiers selected for this study are widely used in classification tasks and were evaluated for their ability to enhance model accuracy and effectiveness in identifying mango groups. The performance of these models was tested using an 80/20 train-test split to ensure a thorough assessment of their classification capabilities. The four classifiers selected for experimentation were:

1) Decision Trees: A tree-like model [12] that classifies data points by following a series of decision rules, making it intuitive and easy to interpret.
2) Random Forest: An ensemble method [13] that aggregates the predictions of multiple decision trees, thereby improving classification accuracy and reducing the risk of overfitting by introducing randomness during tree construction.
3) Gradient Boosting: An advanced ensemble technique [14] that sequentially builds decision trees, each one correcting the errors of the previous trees, resulting in improved classification performance, particularly for complex datasets.
4) Extra Trees: A variant of the Random Forest algorithm [15] that selects features randomly at each node split, aiming to further reduce correlations between trees and potentially improve the model's ability to generalize unseen data.

Each of these classifiers was rigorously evaluated to determine which provided the best performance for mango group classification, to optimize both accuracy and reliability in the classification process.

D. *Hyperparameter Tuning*

Hyperparameter tuning is a critical process in optimizing machine learning classifiers by adjusting predefined settings that control the learning process. Unlike model parameters, which are learned from training data, hyperparameters significantly influence a model's ability to generalize and avoid overfitting. Proper tuning transforms an average classifier into a highly effective instrument. In this study, different hyperparameters were optimized for each classifier. For Decision Trees, the key hyperparameters included tree depth (maxdepth), minimum samples to split a node (minsamplessplit), number of features considered at each split (maxfeatures), and the splitting criterion (e.g., ”gini” or ”entropy”). For Random Forest, the number of trees (nestimators) and bootstrap sampling were tuned to reduce overfitting. Gradient Boosting involved tuning the number of trees, learning rate, tree depth, minimum samples per split, and the number of features. For Extra Trees, hyperparameter tuning focused on the number of features and whether bootstrap sampling was used. Finetuning these hyperparameters improved the classifiers' accuracy and generalization capabilities, leading to enhanced performance in mango classification.

E. *Evaluation*

The effectiveness of the machine learning classifiers was evaluated using key metrics including accuracy, precision, recall, and F1-score to provide a comprehensive view of model performance. Accuracy measures the overall correctness of predictions, precision assesses the correctness of positive predictions, recall evaluates how well the model identifies actual positives, and F1-score balances precision and recall. To ensure a robust evaluation, K-fold cross validation was applied. The dataset was split into K subsets, and the model was trained and tested K times, with each subset used as the test set once. This approach reduced overfitting and provided a more reliable estimate of model performance across different data samples.

III. RESULTS AND DISCUSSION

This section provides a detailed analysis of the findings from the clustering phase and the evaluation of the classifier models. The results are discussed for their effectiveness in categorizing the data and the performance of the classification algorithms.

A. *Clustering results*

The clustering results were evaluated using the Elbow Method, Silhouette Method, and Calinski-Harabasz Method.

The Elbow Method suggested K=1 as the optimal number of clusters, as shown in Figure 2, indicating significant inherent similarity within the data, though this oversimplified the diversity of the dataset.

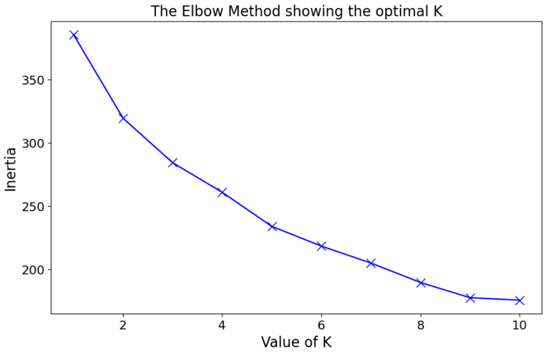

Fig. 2. The Elbow Method showing the optimal K

The Calinski-Harabasz Method identified K=9 as the optimal number of clusters, as shown in Figure 3, suggesting finer data partitioning. However, this could lead to overfitting by creating overly specific clusters.

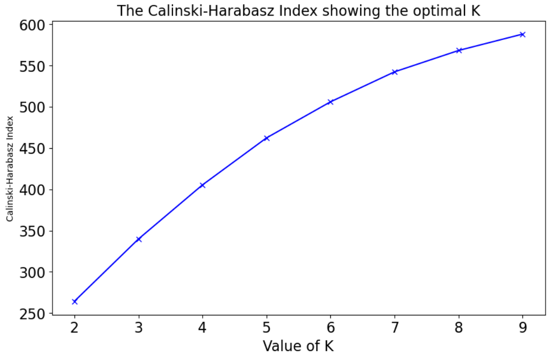

Fig. 3. The Calinski-Harabasz Method showing the optimal K

The Silhouette Method found K=8 to be the optimal number of clusters, providing a balance between compactness and separation, as shown in Figure 4. Among these methods, the Silhouette Method offered the best overall solution.

The eight clusters were further analyzed based on their dominant characteristics, as shown in Figure 5.

- Cluster 1 featured diverse fruit shapes and leaf margins, representing crossbred cultivars or local landraces.
- Cluster 2 contained mostly elliptical fruits and pointed leaf bases, representing cultivars with shared geographical origins.
- Cluster 3 was characterized by diverse fruit shapes and tapering leaf bases, indicating a genetically diverse group adapted to various environments.

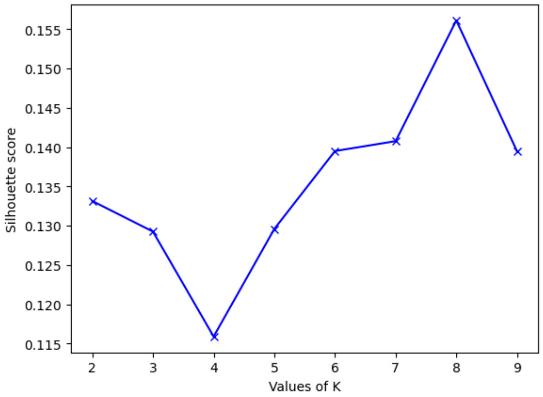

Fig. 4. The Silhouette Method showing the optimal K

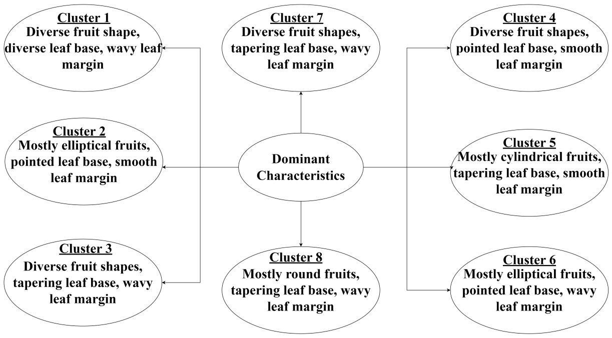
Cluster 1
Diverse fruit shape, diverse leaf base, wavy leaf margin
Cluster 7
Diverse fruit shapes, tapering leaf base, wavy leaf margin
Cluster 4
Diverse fruit shapes, pointed leaf base, smooth leaf margin
Cluster 2
Mostly elliptical fruits, pointed leaf base, smooth leaf margin
Dominant Characteristics
Cluster 5
Mostly cylindrical fruits, tapering leaf base, smooth leaf margin
Cluster 3
Diverse fruit shapes, tapering leaf base, wavy leaf margin
Cluster 8
Mostly round fruits, tapering leaf base, wavy leaf margin
Cluster 6
Mostly elliptical fruits, pointed leaf base, wavy leaf margin

Fig. 5. Clustering result

- Cluster 4 represented a commercially cultivated group with desirable traits for the market.
- Cluster 5 was a distinct group with specialized characteristics for breeding programs.
- Cluster 6 included mostly elliptical fruits, known for their sweetness or market recognition.
- Cluster 7 featured diverse fruit shapes, representing imported or newly developed cultivars.
- Cluster 8 comprised mostly round fruits, representing large, high-yield varieties.

B. *Results of classification performance*

This section presents the performance evaluation of the four machine learning classifiers: Decision Trees, Random Forest, Extra Trees, and Gradient Boosting. The models were assessed using key metrics, including accuracy, precision, recall, and F1-score to determine their effectiveness in classifying the dataset. As shown in Table I, the Random Forest model outperformed the others, achieving the highest accuracy of 84.21%, with a precision of 92.54% and an F1-score of 85.46%. The Extra Trees model gave a moderate accuracy of 78.95%, outperforming the Decision Trees and Gradient Boosting models, which recorded the lowest accuracy of 73.68%.

TABLE I
MODEL VALIDATION RESULTS: COMPARISON OF CLASSIFIER PERFORMANCE ON TEST DATA (%)

Model	Precision	Recall	F1-score	Accuracy
Decision Tree	87.72	73.68	76.29	73.68
Random Forest	**92.54**	**84.21**	**85.46**	**84.21**
Extra Trees	85.26	78.95	78.71	78.95
Gradient Boosting	89.47	73.68	78.40	73.68

TABLE II
HYPERPARAMETER TUNING RESULTS: CLASSIFIER PERFORMANCE ON TEST DATA (%)

Model	Precision	Recall	F1-score	Accuracy
Decision Tree	90.79	89.47	89.56	89.47
Random Forest	**97.06**	**96.78**	**96.07**	**96.78**
Extra Trees	96.49	89.47	91.23	89.47
Gradient Boosting	84.21	89.47	84.21	84.21

TABLE III
K-FOLD CROSS VALIDATION RESULTS: AFTER HYPERPARAMETER TUNING (%)

Model	Precision	Recall	F1-score	Accuracy
Decision Tree	87.47	88.07	86.70	88.07
Random Forest	**91.80**	**90.28**	**89.98**	**91.26**
Extra Trees	88.84	88.12	87.13	87.13
Gradient Boosting	69.47	94.74	65.82	69.47

Based on these results, hyperparameter tuning was applied to optimize the performance of each model and improve classification accuracy and overall effectiveness.

C. Optimization of hyperparameter tuning for classification

After the initial model evaluation, the focus shifted to identifying the optimal hyperparameters to improve key performance metrics such as accuracy, precision, recall, and F1 score. The goal was to enhance the overall performance of each classifier. Table II presents the optimized results for each model, along with their corresponding hyperparameter settings:

- Decision Trees: criterion: gini, max_depth: None
- Random Forest: criterion: gini, n_estimators: 10, max_depth: 5
- Extra Trees: criterion: gini, n_estimators: 30, max_depth: 15
- Gradient Boosting: criterion: mae, n_estimators: 1, max_depth: None

The hyperparameter tuning results showed that the Random Forest model achieved the highest performance across all metrics with a precision of 97.06%, recall of 96.78%, F1-score of 96.07%, and accuracy of 96.78%. The Extra Trees model followed with a strong precision of 96.49% but a lower recall and accuracy of 89.47%. The Decision Trees model performed moderately with an accuracy of 89.47%. The Gradient Boosting model had the lowest performance with accuracy and precision of 84.21%. Overall, Random Forest emerged as the most effective classifier after hyperparameter tuning.

D. Comparison with K-fold cross validation results

Despite the improvements achieved through hyperparameter tuning, the challenge of selecting the best-performing model persisted. To address this, K-fold cross validation was employed. This technique systematically divides the dataset into multiple folds and evaluates the model's performance on each fold to ensure a robust assessment. In this study, K was set to 5, providing a thorough evaluation of each model's generalization ability. Results in Table III show that the Random Forest model achieved the highest average accuracy of 91.26%, outperforming all the other models. This suggested that the Random Forest model was highly effective at capturing patterns and relationships within the data, making it the most reliable choice for accurately classifying mango groups.

IV. CONCLUSIONS

This research applied machine learning to classify mango groups based on physical characteristics using the Random Forest model. The model achieved a high accuracy of 91.26%, highlighting its effectiveness in distinguishing mango varieties. This advancement has important implications for the agricultural industry by improving the efficiency and accuracy of mango classification. The potential of machine learning to enhance sorting, grading, and marketing strategies for producers was demonstrated while also providing consumers with better insights into mango varieties.

ACKNOWLEDGMENT

The authors are grateful for the Mae Fah Luang University grant under the National Science, Research, and Innovation Fund (NSRF) 2024, grant no. 6721A03036. Mae Fah Luang University provided financial support and experimental sites for this research study.

REFERENCES

[1] W.P.Review, "https://worldpopulationreview.com/country-rankings/mango-production- by-country," in *"Mango production by country 2024,"*, 2024, pp. accessed: 2024–05–02.

[2] I. Mehta, "International journal of engineering science invention," in *"History of mango – 'king of fruits',,"*, vol. 6, no. 7, 2021, pp. 1–11.

[3] V. Wangnain, "Foreign mango varieties," in *In Proceedings of the seminar on the production of mangoes for export.*, vol. 2, 1990, pp. 41–45.

[4] L. Kamelia, "Analysis of decision tree and random forest algorithms for nutrient deficiency classification in citrus leaves through image processing," in *2023 17th International Conference on Telecommunication Systems, Services, and Applications (TSSA)*, 2023.

[5] H. Zawbaa, "Automatic fruit classification using random forest algorithm," in *2014 14th International Conference on Hybrid Intelligent Systems*, 2014.

[6] A. Pravin, "Classification using deep cnn feature extraction and hyper parameter tuned random forest in piperaceae plant type," in *2023 International Conference on New Frontiers in Communication, Automation, Management and Security (ICCAMS)*, 2023.

[7] U. Pise, "Grading of harvested mangoes quality and maturity based on machine learning techniques," in *2018 International Conference on Smart City and Emerging Technology (ICSCET)*, vol. 2, 2018, pp. 1–6.

[8] W. A. Thong., "The diversity of mango species and genetics in thailand." in *mangoes for export. Kasetsart University*, 2015, pp. 183–188.

[9] D. Marutho, "The determination of cluster number at k-mean using elbow method and purity evaluation on headline news," in *2018 International Seminar on Application for Technology of Information and Communication*, 2018.

[10] P. J., "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," in *Journal of Computational and Applied Mathematics*, 1987, pp. 53–65.

[11] S. Łukasik, "Clustering using flower pollination algorithm and calinski-harabasz index," in *2016 IEEE Congress on Evolutionary Computation (CEC)*, 2016.

[12] O. S. Friedman, "Classification and regression trees." in *Monterey, CA: Wadsworth and Brooks/Cole Advanced Books and Software.*, 1984, pp. 322–331.

[13] L. Breiman, "Random forests," in *Statistics Department, University of California, Berkeley, CA, 94720*, vol. 45, 2001, pp. 5–32.

[14] D. Arnold, "Gradient boosting feature selection with machine learning classifiers," in *Faculty of Computer Science, Dalhousie University, Halifax, Canada*, vol. 18, 2020, pp. 1104 – 1116.

[15] B. Merz, "Extra trees," in *Extremely randomized trees*, vol. 63, 2006, p. 3–42.