Morphological Image Processing for Cocoa Bean Classification

Dr. Amitava Akuli[1], Samikshan Das[2], Dr. Anil Kumar Bag[3*], Suparna Parua[4], Alokesh Ghosh[5]

[1,2,4,5]Centre for Development of Advanced Computing, Kolkata 700 091, India.
[2]Dept. of Information Technology, Maulana Abul Kalam Azad University of Technology, Nadia 741 249, India.
[3]Heritage Institute of Technology, Kolkata, India.
[1]amitava.akuli@gmail.com, [2]samikshandas5@gmail.com, [3]anilkumar.bag@heritageit.edu, [4]suparna.parua.sp@gmail.com, [5]alokesh.ghosh@cdac.in

Abstract

The purpose of this paper is to offer a machine vision approach for classifying cocoa beans based on their morphological properties. Using traditional machine learning approaches, the shape and size of cocoa beans were retrieved from photographs. A series of image processing techniques are used to extract the features from the photos. Finally, typical machine learning approaches such as KNN, SVM, Decision Tree, and Random Forest are used to divide the cocoa beans into four groups: large, medium, small, and rejected. A comparison of different methodologies is also carried out. Two optimization strategies, Univariate Selection and Feature Importance, are used to maximize retrieved features prior to training the model. For performance analysis, trained models are evaluated using stratified K-fold cross validations and the mean cross validation score is produced. The Random Forest Classifier has the greatest accuracy score of 0.75, according to the results of the experiments.

Keywords: *Cocoa beans, Classification, Image processing, Machine Learning, Feature Optimization.*

1. Introduction

Chocolate and chocolate confections rely heavily on cocoa. Chocolate is highly popular and is one of the most widely consumed foods on the planet. Depending on the desired cocoa percentage, 300 to 600 cocoa beans are processed to generate 1 kg (2.2 pounds) of chocolate. Cocoa nibs, cocoa paste (mass or liquor), butter, powder, and coverture are examples of semi-finished cocoa goods made from roasted cocoa beans. These items are mostly used in the production of chocolate and other food items. They are rarely sold directly to the general public. Soaps and cosmetics are also made from cocoa beans. Cocoa beans are grown in tropical areas around the Equator, where the environment is ideal for cultivating cocoa trees. Ivory Coast, Ghana, Indonesia, Nigeria, and Cameroon produce over 70% of the world's cocoa beans. Figure 1. reflects the percentage of cocoa production by continent. After the cocoa beans are plucked from the tree, they are subjected to a post-harvesting treatment. Fermentation is the earliest chemical reaction [1]. It is one of the most crucial procedures since it increases the product's

ultimate quality, and processors constantly require well-fermented cocoa beans because it ensures the production of aroma precursors and cocoa flavour. Fermentation degree [2] is strongly linked to cocoa quality parameters such reducing sugars, free amino acids, and bean pH. Similarly, good fermentation helps to reduce the bitterness and astringency of cocoa. The next step in the process is drying, which decreases the acidity of the cocoa beans, resulting in cocoa. After that, the beans are roasted. Roasting aids in the management of the beans' final flavour. As a result, quality parameters such as imaging and sensory evaluation are available.

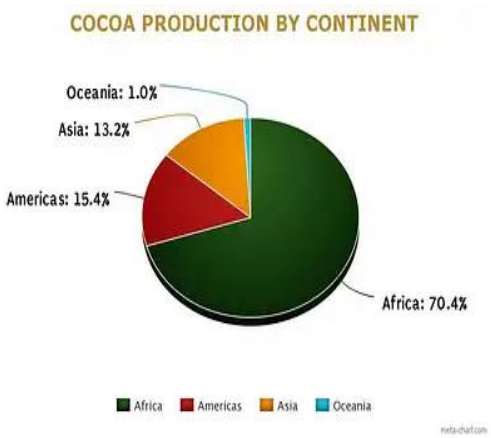


Figure 1. Cocoa Production by Continent

We performed an imagery evaluation as part of our scope [3]. Cocoa bean quality is assessed by its size, shape, colour, and texture. Large beans, medium beans, small beans, and rejected beans are the different types of beans. Traditional and manual processes are used to inspect cocoa beans for quality, such as using the visual approach on cocoa beans and selecting them one by one. The thing on the surface of the cocoa beans must be seen clearly by human vision. They are usually only armed with prior experience and information. Manual inspections include drawbacks, such as fatigued eyes and differing analytical results from examiner to examiner. This method is qualitative, time-consuming, and highly subjective. Furthermore, human errors might make it inefficient for big amounts of cocoa beans.

The manual analysis method is time-consuming, tedious, and costly, and it necessitates a higher level of skill. As a result, manual analysis may not be appropriate for routine quality inspections on industrial cocoa beans. As a result, a quick and effective method for classifying cocoa beans for quality control is important. In recent years, machine vision has become popular for assessing the quality of agricultural and food commodities. Automation technology, aided by artificial intelligence, can be deployed to overcome the drawbacks of manual inspection. To enable automated inspection, computer vision [4], which combines image analysis and machine learning [5] techniques, is used. Images can be analysed and processed here in order to provide helpful information to the user. The image's morphological features, such as size, shape, and texture, are extracted. Features are optimised utilising two feature optimization techniques, namely Univariate selection and Feature Importance, to reduce redundant and irrelevant features. The major goal of this research is to see if size, shape, and texture characteristics may be used to assess cocoa bean quality. To determine the cocoa bean quality, four classifications and four machine learning algorithms are used. After testing

on the cocoa bean test dataset, a comparison study was conducted based on the performance of the four algorithms: KNN [6], Support Vector Machine [7], Decision Tree [8], and Random Forest [9]. The findings of the experiments are presented accordingly. Figure 2. contains the images of raw cocoa beans and the process flow for post-harvesting process.

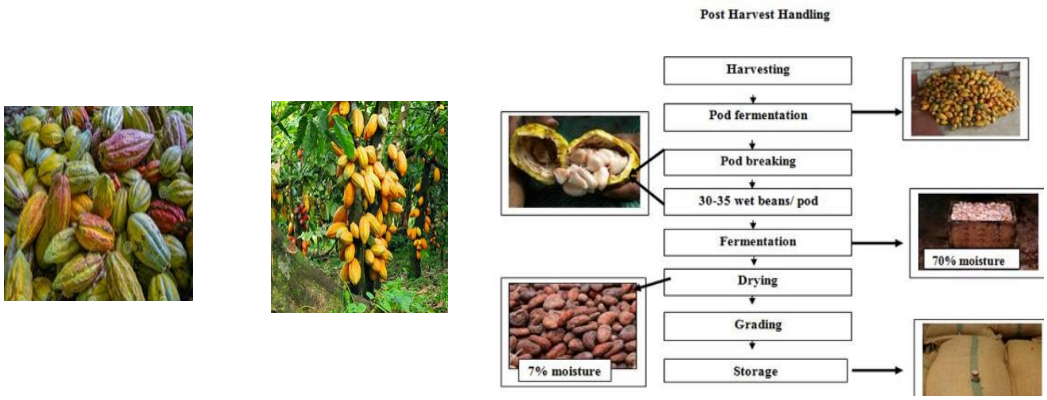
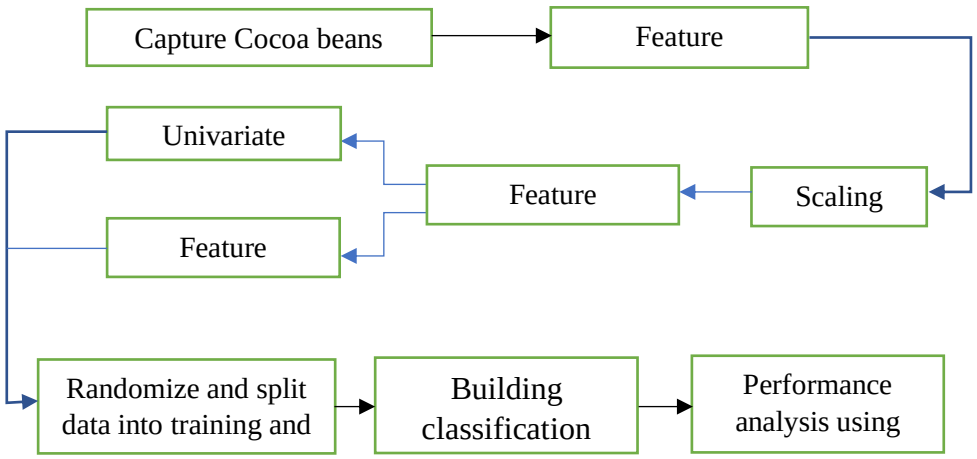


Figure 2. Cocoa Beans Process Flow for Harvesting and Post-Harvesting

2. Materials and Methods

2.1. Workflow Diagram:

The proposed workflow diagram is shown below



2.2. Data Collection & Preparation:

The data collection is done in the form of digital data or photographs of cocoa beans, and the samples are obtained from the market in India. Beans are placed on a white background before the photographs are taken. It is preferable to use 25 beans each image. With the use of a digital camera tool, data is collected by snapping images of objects.

2.2.1. Image capturing setup:

The e-COCOA Vision system involves capturing images from an input device, analysing them, and finally rating cocoa samples using predefined criteria. The block diagram of the system is shown in Figure 3.

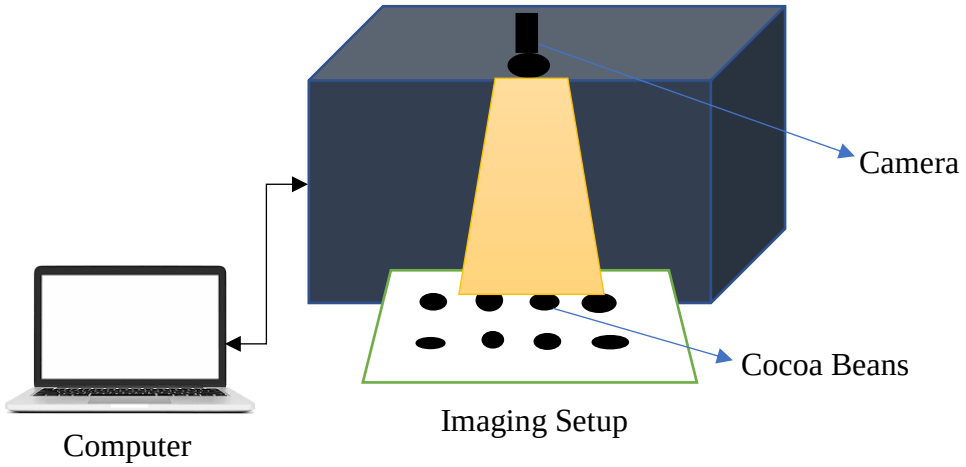


Figure 3. System Setup for Image Capturing

A portable image capture setup has been designed and is described in detail below.

- A total of 20 LEDs are evenly spaced around the cabinet's ceiling.
- The image is captured with a Logitech C920 webcam.
- To eliminate excessive reflections, the cabinet is built of aluminium sheet and painted black.

The digital images of cocoa beans are divided into four classes (Figure 4.), three of which are whole beans and are categorised as (1) large bean (2) medium bean (3) small bean, and the rest are rejected beans that are fragmented. 220 beans were photographed for experimentation. 70% of the 220 images were taken for model training, while 30% were used for testing.

Figure 4. Cocoa Beans on White Paper.

2.2.2. Data Pre-processing:

Following data collection, data must be processed to improve image quality and remove background noise. The following steps are taken.

- **Gray Image Conversion:** We're using a 24 bit RGB image. The RGB image was converted to an 8-bit gray scale image. Image analysis in gray scale allows us to remove the white background.
- **Image Segmentation:** For image thresholding, a global thresholding technique based on OTSU was used. The thresholding technique produces a binary image as the output image.
- **Smoothing with Gaussian Filter:** A Gaussian smoothing filter with a kernel size of 3 was used for smoothing. This aids in the removal of high frequency noise in the image.
- **Object Identification:** The erosion technique is used to identify and remove small particles that are close to the image boundaries. Finally, the area of the particles in the image is used to identify objects.

2.3. Feature Extraction:

From the images, a set of 23 image features are extracted. They are Perimeter, Convex Hull Perimeter, Max Feret Diameter, Equivalent Ellipse Major Axis, Equivalent Ellipse Minor axis, Equivalent Rectangle Long Side, Equivalent Rectangle Short Side, Equivalent Rectangle Diagonal, Hydraulic Radius, area, Convex Hull Area, Ratio of Equivalent Ellipse Axis, Ratio of Equivalent Rectangle sides, Elongation Factor, Compactness Factor, Heywood Circularity Factor, and 7 HU Moment features. Python 3.9.7 was used to create the classification model, and the following Python libraries were imported: matplotlib, pandas, numpy, seaborn, sklearn, pydotplus, and six.

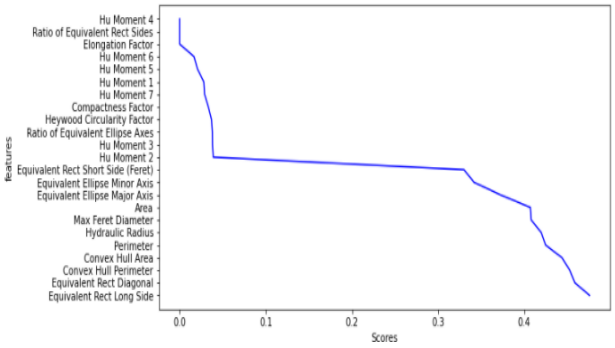
2.4. Feature Optimization:

In general, all of the independent features in the dataset have no effect on the dependent feature in the same measure for a machine learning model. Some features may have a negligible impact. Feature optimization [10] is used to improve machine learning models by removing redundant features. It reduces model training time and complexity without sacrificing accuracy. Two feature optimization techniques, namely Univariate feature selection and Feature Importance, were used in this study.

Scikitlearn provides the SelectKBest class for **Univariate feature selection** [11], which works with a suite of different statistical tests and measures the correlation between each feature and the target label based on the results of these tests. The classification statistical tests available in this class are 'f_classif' for ANOVA F-value [12] between features, 'mutual_info_classif' for mutual information [13] for a discrete target label, and 'chi2' for Chi-squared stats [14] of non-negative features. The mutual information test is chosen with the understanding that the independent features are continuous and the dependent features are categorical. Mutual information between two random variables is a non-negative value that measures the variables' dependence. It is equal to zero when two random variables are independent, and higher values indicate greater dependency.

Feature significance [15] is a class of methods that assigns a score to each independent feature in a prediction model based on the feature's relevance in producing accurate predictions. The higher the score, the more the trait is related to the goal label. The 'Extra Tree Classifier'

method from the 'scikit learn' package is utilised to measure feature importance in this study. It uses a variety of randomised decision trees as estimators to calculate the importance of each feature and identify the top relevant characteristics using different subsets of the main dataset.

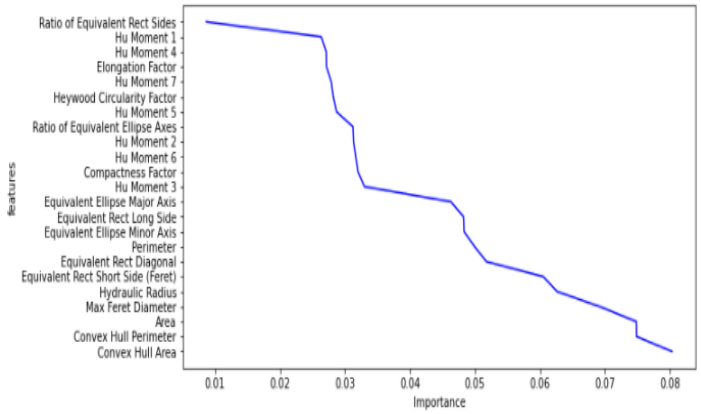


	features	Score
0	Perimeter	0.424857
1	Convex Hull Perimeter	0.452859
2	Max Feret Diameter	0.408114
3	Equivalent Ellipse Major Axis	0.372409
4	Equivalent Ellipse Minor Axis	0.341688
5	Equivalent Rect Long Side	0.475832
6	Equivalent Rect Short Side (Feret)	0.330108
7	Equivalent Rect Diagonal	0.458822
8	Hydraulic Radius	0.419752
9	Area	0.407097
10	Convex Hull Area	0.443806
11	Ratio of Equivalent Ellipse Axes	0.037955
12	Ratio of Equivalent Rect Sides	0.000000
13	Elongation Factor	0.000000
14	Compactness Factor	0.033345
15	Heywood Circularity Factor	0.036971
16	Hu Moment 1	0.028022
17	Hu Moment 2	0.039065
18	Hu Moment 3	0.038141
19	Hu Moment 4	0.000000
20	Hu Moment 5	0.020831
21	Hu Moment 6	0.016707
22	Hu Moment 7	0.028995

Figure 5. Line Chart and Feature Scores for Univariate Feature Selection

The line charts in Figure 5. and Figure 6. show the relationship between each of the individual features and the target dependent feature after applying univariate selection and feature importance optimization procedures. In both cases, the curve abruptly changes after the first 11 characteristics with the highest scores, indicating that these 11 elements have the greatest impact and the remaining features are significantly less relevant in predicting the target class labels. As a result, the categorization models are built around these 11 most important traits.

These features are- (1) ‘Convex Hull Area', (2) 'Convex Hull Perimeter', (3) 'Equivalent Rectangle Short Side (Feret)', (4) 'Max Feret Diameter', (5) 'Area', (6) 'Perimeter', (7) 'Equivalent Ellipse Major Axis', (8) 'Equivalent Rectangle Diagonal', (9) 'Equivalent Rectangle Long Side', (10) 'Hydraulic Radius', and (11) 'Equivalent Ellipse Minor Axis’.



	Features	Importance
0	Perimeter	0.049933
1	Convex Hull Perimeter	0.074856
2	Max Feret Diameter	0.069188
3	Equivalent Ellipse Major Axis	0.046265
4	Equivalent Ellipse Minor Axis	0.048323
5	Equivalent Rect Long Side	0.048236
6	Equivalent Rect Short Side (Feret)	0.060495
7	Equivalent Rect Diagonal	0.051756
8	Hydraulic Radius	0.062639
9	Area	0.074827
10	Convex Hull Area	0.080332
11	Ratio of Equivalent Ellipse Axes	0.031214
12	Ratio of Equivalent Rect Sides	0.008642
13	Elongation Factor	0.027147
14	Compactness Factor	0.031998
15	Heywood Circularity Factor	0.028187
16	Hu Moment 1	0.026344
17	Hu Moment 2	0.031306
18	Hu Moment 3	0.032991
19	Hu Moment 4	0.027105
20	Hu Moment 5	0.028685
21	Hu Moment 6	0.031656
22	Hu Moment 7	0.027874

Figure 6. Line Chart and Feature Importance Score

2.6. Data Analysis:

2.5.1. Training and Testing dataset:

The dataset containing 220 samples is split into training and testing datasets using the scikitlearn library's 'train test split' method. The training dataset has 70% of the original dataset, 154 beans, and the testing dataset contains the remaining 30%, 66 beans.

2.5.2. Classification Algorithms:

In terms of algorithms, two major types of supervised approaches [16] for classification have been chosen depending on the classification problem: distance based algorithms and tree based algorithms.

Two traditional classification methods, K-Nearest Neighbour (KNN) and Support Vector Machine (SVM), are utilised for distance-based approaches. KNN categorises cases based on their similarity, which is quantified using a distance matrix such as the Euclidean distance [17], Manhattan distance [18], Minkowski distance [19], or Hamming distance [20]. 'Neighboring' cases are those that are close to each other. When predicting classes for unknown data points, the most common class label or the class label with the majority value from its neighbours is used. SVM, on the other hand, is effective at dealing with dataset non-linearity by translating the data to a higher dimensional space and then classifying the data by identifying the optimum hyperplane that effectively distinguishes the classes. Although SVM is memory efficient because it only uses a subset of the training data in the decision function, it takes longer to train than KNN because KNN does not derive any discriminative function from the training data;

instead, it stores the training dataset and learns from it only when making real-time predictions, whereas SVM learns throughout the training period.

Two common algorithms for tree-based algorithms are Decision Tree Classifier and Random Forest Classifier. The Decision Tree Classifier is a tree-structured classifier with branches that reflect decision rules, internal nodes that contain sample dataset attributes, and leaf nodes that represent the final output or class labels. It splits the training records into parts using Recursive Partitioning [20], which minimises impurity at each stage. However, when a decision tree is developed to its full depth, it has low bias, implying that the model is overfitted to the training dataset, and high variance, implying that the model is prone to large amounts of errors while working with new test data. Instead of employing a single decision tree, the Random Forest Classifier considers numerous decision trees with high variance constructed from subsets of the primary dataset, and the high variance is reduced to low variance by mixing the trees with respect to a majority vote. Furthermore, if we alter or add new data to our model, it will have little impact because the changes will be dispersed throughout all decision trees while we sample the rows and columns randomly.

2.5.3. Feature scaling:

The range of characteristics has an impact on algorithms like KNN and SVM, which use the distance between samples to assess how similar they are. As a result, before training KNN and SVM models, the independent feature set is rescaled using Min-Max Normalization to range between 0 and 1.

$$X' = (X - X_{min})/(X_{max} - X_{min})$$

Tree-based classifiers, on the other hand, are not sensitive to feature scale; hence these classification models can function effectively without rescaling the feature set.

3. Results and Discussion

3.1. Training Classification Models:

3.1.1. KNN model:

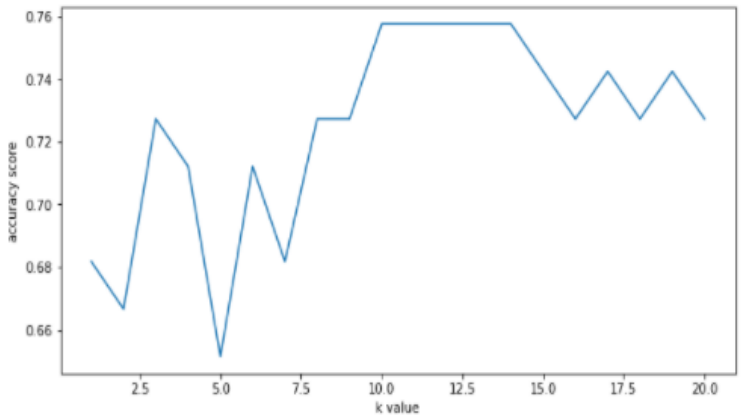
The training dataset for the KNN classification model has a K value in the range of 1 to 20. The best minimum value for K is discovered to be 10 using the testing dataset, for which the model predicts with the highest accuracy. The Value of K Vs. Accuracy Score and Value of K and corresponding accuracy score are shown in Figure 7.

3.1.2. SVM model:

The SVM kernel function translates a low-dimensional input space into a higher-dimensional one. The SVM model is trained in this work utilising four different kernel [21] functions: linear, RBF, polynomial, and sigmoid.

3.1.3. Decision Tree model:

Entropy [22] with information gain, Gini index [23], Gain Ratio [24], Reduction in Variance, and Chi-Square are some criteria provided by researchers before for selecting the root node or internal nodes at different levels for the Decision Tree Classifier. The attributes in this study were chosen based on Entropy. The amount of randomness or disorder in the data is measured by entropy. The objective is to locate the tree with the least amount of entropy in its nodes. So, the criterion function for splitting is 'entropy' with the 'best' splitter technique, and the maximum depth for the decision tree is 4 after trying a range of values from 1 to 10 as maximum depth to reach the highest accuracy score.



K	Accuracy Score
1	0.681818
2	0.666667
3	0.727273
4	0.712121
5	0.651515
6	0.712121
7	0.681818
8	0.727273
9	0.727273
10	0.757576
11	0.757576
12	0.757576
13	0.757576
14	0.757576
15	0.742424
16	0.727273
17	0.742424
18	0.727273
19	0.742424
20	0.727273

Figure 7: Value of K Vs Accuracy Score and Value of K and Corresponding Accuracy Score

3.1.4. Random Forest model:

The criterion function chosen for splitting the decision trees in the Random forest classifier is 'entropy' for 150 decision trees with a maximum depth of 4 to attain the highest accuracy score.

3.2. Testing Classification Models:

Because the distributions of different class labels are not homogeneous, the dataset employed in this study is not a balanced dataset. As a result, after testing the classification models using the testing dataset, F1 score [25] and Accuracy score were employed as assessment matrices.

3.2.1. Evaluation Matrix:

Accuracy score is the sum of True Positive (TP) and True Negative (TN) divided by the sum of True Positive, True Negative, False Positive (FP) and False Negative (FN).

$$Accuracy\ score = (TP + TN) / (TP + TN + FP + FN);$$

$$F\beta\ score = (1 + \beta^2) * (Precision * Recall) / (\beta^2 * Precision + Recall);$$

Precision is the ratio of correctly predicted positive observation to the total predicted positive observation. *Recall* is the ratio of correctly predicted positive observation to the all observations in actual positive class. F1 score is the Fβ score where β=1.

$$Precision = TP / (TP + FP);$$ $$Recall = TP / (TP + FN);$$

The accuracy score and F1_score for all four classification models are calculated using the Scikitlearn library's 'accuracy score' and 'f1 score' methods.

3.2.1.1. KNN model:

The accuracy score for the KNN classification model is 0.76, while the F1 score is 0.71. Classification report of KNN model is shown in Table 1.

Table 1. KNN Classification Report

	Precision	Recall	F1-score	Support
Large	0.71	0.56	0.63	9
Medium	0.71	0.86	0.78	35
Rejected	1.00	0.29	0.44	7
Small	0.73	0.73	0.73	15
Accuracy			0.73	66
Macro avg	0.79	0.61	0.65	66
Weighted avg	0.75	0.73	0.71	66

3.2.1.2. SVM model:

The SVM model has a maximum accuracy score of 0.73 and an F1 score of 0.71 when using the Polynomial Kernel function. Classification report of SVM model is shown in Table 2. The accuracy and F1 scores for several kernel functions are listed in Table 3:

Table 2. SVM Classification Report

	Precision	Recall	F1-score	Support
Large	0.60	0.67	0.63	9
Medium	0.75	0.94	0.84	35
Rejected	0.00	0.00	0.00	7
Small	0.92	0.73	0.81	15
Accuracy			0.76	66
Macro avg	0.57	0.59	0.57	66
Weighted avg	0.69	0.76	0.71	66

Table 3. Accuracy Score and F1 Score for Different SVM Kernel Functions

SVM Kernel Function	Accuracy Score	F1 Score
Linear	0.681818	0.61
RBF	0.727273	0.68
Sigmoid	0.530303	0.38
Polynomial	0.727273	0.71

3.2.1.3. Decision Tree model:

The maximum accuracy score and F1 score obtained from the Decision Tree model are 0.74 and 0.73 respectively and is shown in Figure 8. Classification report of Decision Tree model is shown in Table 4.

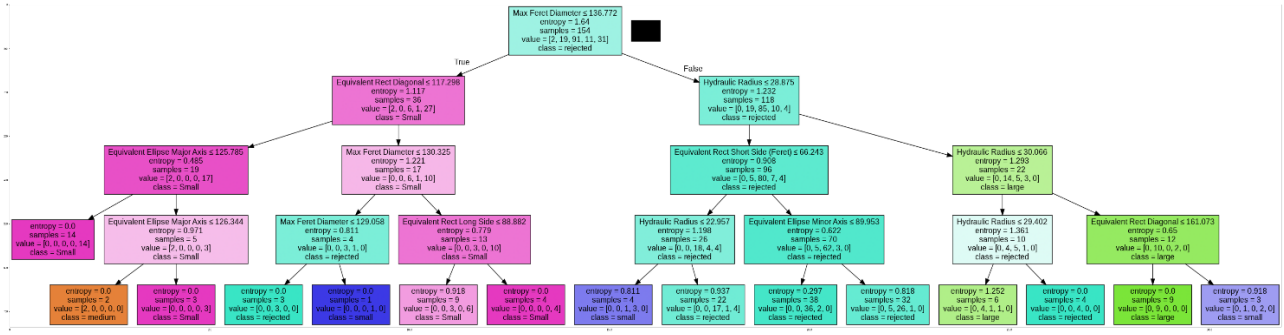


Figure 8. Decision Tree Obtained From Decision Tree Classifier

Table 4. Decision Tree Classification Report

	Precision	Recall	F1-score	Support
Large	1.00	0.56	0.71	9
Medium	0.73	0.94	0.83	35
Rejected	0.40	0.29	0.33	7
Small	0.82	0.60	0.69	15
Accuracy			0.74	66
Macro avg	0.74	0.60	0.64	66
Weighted avg	0.75	0.74	0.73	66

3.2.1.4. Random Forest model:

The Random Forest classification model results 0.74 accuracy score and F1 score is 0.71. Classification report of Random Forest model is shown in Table 5.

Table 5. Random Forest Classification Report

	Precision	Recall	F1-score	Support
Large	0.67	0.67	0.67	9
Medium	0.71	0.97	0.82	35
Rejected	1.00	0.14	0.25	7

Small	1.00	0.53	0.70	15
Accuracy			0.74	66
Macro avg	0.84	0.58	0.61	66
Weighted avg	0.80	0.74	0.71	66

3.2.2. Cross Validation:

Cross validation [26, 27] is a statistical method for evaluating and comparing machine learning algorithms that divides the dataset into two parts: one for training the model and the other for validating it. Four classification models are assessed using Stratified K fold cross validation with 10 folds because the dataset is not fully balanced at the end. It ensures that the proportion of target features in distinct classes is consistent throughout original, training, and testing data. Cross validation scores for the classifiers are shown in Table 6.

Table 6: Cross Validation Scores

Fold	**KNN** Cross Validation Scores	**SVM** Cross Validation Scores	**DT** Cross Validation Scores	**RF** Cross Validation Scores
1	0.7272	0.7272	0.6818	0.7727
2	0.5909	0.6818	0.7272	0.6818
3	0.7272	0.7727	0.7272	0.7727
4	0.6818	0.7727	0.6818	0.7272
5	0.6363	0.5909	0.6363	0.6818
6	0.7727	0.7727	0.7727	0.8181
7	0.7727	0.7727	0.7272	0.7727
8	0.8636	0.8636	0.8181	0.8181
9	0.5909	0.5909	0.6818	0.6363
10	0.7272	0.7272	0.7727	0.7727
Max	0.86	0.86	0.82	0.82
Min	0.59	0.59	0.64	0.64
Mean	0.71	0.73	0.72	0.75

4. Conclusion

The paper been emphasized on the classification of cocoa beans based on their morphological properties using machine vision technique. KNN, SVM, Decision Tree, and Random Forest machine learning algorithms were applied to categorize the cocoa beans into four classes such as large, medium, small, and rejected. It is seen that the resultant accuracy scores and F1 scores achieved by these models are in the range of 0.72 to 0.75 and 0.68 to 0.73, respectively, by training and testing the classification models using morphological feature set. According to the findings, Random Forest Classifier is the most accurate of the four algorithms in classifying cocoa beans.

References

[1] *L. De Vuyst,S. Weckx.(2016). Review Article: The cocoa bean fermentation process: from ecosystem analysis to starter culture development, Journal of Applied Microbiology ISSN 1364-5072, https://doi.org/10.1111/jam.13045.*

[2] *K. Sánchez; J. Bacca; L. Arévalo-Sánchez; H. Arguello; S. Castillo, "Classification of Cocoa Beans Based on their Level of Fermentation using Spectral information," TecnoLógicas, vol. 24, nro. 50, e1654, 2021. https://doi.org/10.22430/22565337.1654.*

[3] *Relf, Christopher. (2003). Image Acquisition and Processing with LabVIEW. 10.1201/9780203487303.*

[4] *Marciano M. Oliveira, Breno V. Cerqueira, Sylvio Barbon, Douglas F. Barbin,Classification of fermented cocoa beans (cut test) using computer vision,Journal of Food Composition and Analysis, Volume 97,2021, 103771,ISSN 0889-1575,https://doi.org/10.1016/j.jfca.2020.103771.*

[5] *Asharul Islam Khan, Salim Al-Habsi,Machine Learning in Computer Vision,Procedia Computer Science,Volume 167,2020, Pages 1444-1451, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2020.03.355.*

[6] *Guo, Gongde & Wang, Hui & Bell, David & Bi, Yaxin. (2004). KNN Model-Based Approach in Classification.*

[7] *Evgeniou, Theodoros & Pontil, Massimiliano. (2001). Support Vector Machines: Theory and Applications. 2049. 249-257. 10.1007/3-540-44673-7_12.*

[8] *Patel, Harsh & Prajapati, Purvi. (2018). Study and Analysis of Decision Tree Based Classification Algorithms. International Journal of Computer Sciences and Engineering. 6. 74-78. 10.26438/ijcse/v6i10.7478.*

[9] *Ali, Jehad & Khan, Rehanullah & Ahmad, Nasir & Maqsood, Imran. (2012). Random Forests and Decision Trees. International Journal of Computer Science Issues (IJCSI). 9.*

[10] *Abualigah, Laith & Aldulaimi, Akram & Al Shinwan, Mohammad & Khasawneh, Ahmad & Alabool, Hamzeh & Diabat, Mofleh & Shehab, Mohammad. (2020). Optimization Algorithms to Solve Feature Selection Problem: A Review. International Journal of Science and Applied Information Technology. 8. 10.30534/ijsait/2019/098620198.*

[11] *Subho, Razaul & Chowdhury, Md & Chaki, Dipankar & Islam, Samiul & Rahman, Md. (2019). A Univariate Feature Selection Approach for Finding Key Factors of Restaurant Business. 605-610. 10.1109/TENSYMP46218.2019.8971127.*

[12] *Ostertagova, Eva & Ostertag, Oskar. (2013). Methodology and Application of One-way ANOVA. American Journal of Mechanical Engineering. 1. 256-261. 10.12691/ajme-1-7-21.*

[13] *N. Hoque, D.K. Bhattacharyya, J.K. Kalita,MIFS-ND: A mutual information-based feature selection method,Expert Systems with Applications,Volume 41, Issue 14, 2014,Pages 6371-6385,ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2014.04.019.*

[14] *Singhal, Richa & Rana, Rakesh. (2015). Chi-square test and its application in hypothesis testing. Journal of the Practice of Cardiovascular Sciences. 1. 10.4103/2395-5414.157577.*

[15] *Saarela, M., Jauhiainen, S. Comparison of feature importance measures as explanations for classification models. SN Appl. Sci. 3, 272 (2021). https://doi.org/10.1007/s42452-021-04148-9.*

[16] Akinsola, J E T. (2017). Supervised Machine Learning Algorithms: Classification and Comparison. International Journal of Computer Trends and Technology (IJCTT). 48. 128 - 138. 10.14445/22312803/IJCTT-V48P126.

[17] I. Dokmanic, R. Parhizkar, J. Ranieri and M. Vetterli, "Euclidean Distance Matrices: Essential theory, algorithms, and applications," in IEEE Signal Processing Magazine, vol. 32, no. 6, pp. 12-30, Nov. 2015, doi: 10.1109/MSP.2015.2398954.

[18] M. D. Malkauthekar, "Analysis of euclidean distance and Manhattan Distance measure in face recognition," Third International Conference on Computational Intelligence and Information Technology (CIIT 2013), 2013, pp. 503-507, doi: 10.1049/cp.2013.2636.

[19] Bookstein, Abraham & Kulyukin, Vladimir & Raita, Timo. (2002). Generalized Hamming Distance. Information Retrieval. 5. 10.1023/A:1020499411651.

[20] Strobl, Carolin & Malley, James & Tutz, Gerhard. (2009). An Introduction to Recursive Partitioning: Rationale, Application, and Characteristics of Classification and Regression Trees, Bagging, and Random Forests. Psychological methods. 14. 323-48. 10.1037/a0016973.

[21] A. Patle and D. S. Chouhan, "SVM kernel functions for classification," 2013 International Conference on Advances in Technology and Engineering (ICATE), 2013, pp. 1-9, doi: 10.1109/ICAdTE.2013.6524743.

[22] Du, Ming & Wang, Shu & Gong, Gu. (2011). Research on Decision Tree Algorithm Based on Information Entropy. Advanced Materials Research. 267. 732-737. 10.4028/www.scientific.net/AMR.267.732.

[23] Suryakanthi, T.. (2020). Evaluating the Impact of GINI Index and Information Gain on Classification using Decision Tree Classifier Algorithm. International Journal of Advanced Computer Science and Applications. 11. 10.14569/IJACSA.2020.0110277.*

[24] A., Mabayoje & Akintola, Abimbola & Balogun, Abdullateef & Ayilara, Opeyemi. (2015). Gain Ratio and Decision Tree Classifier for Intrusion Detection. International Journal of Computer Applications. 126. 975-8887. 10.5120/ijca2015905983.

[25] Sokolova, Marina & Japkowicz, Nathalie & Szpakowicz, Stan. (2006). Beyond Accuracy, F-Score and ROC: A Family of Discriminant Measures for Performance Evaluation. AI 2006: Advances in Artificial Intelligence, Lecture Notes in Computer Science. Vol. 4304. 1015-1021. 10.1007/11941439_114.

[26] Berrar, Daniel. (2018). Cross-Validation. 10.1016/B978-0-12-809633-8.20349-X.

[27] Rodríguez, Juan & Lozano, Jose. (2007). Repeated stratified k-fold cross-validation on supervised classification with naive Bayes classifier: An empirical analysis.