Discrimination of cocoa beans using structural image features: An Experimental Analysis

Samikshan Das1,2\*, Amitava Akuli1, Suparna Parua Biswas1, Aniruddha Dey2, Ankush Ghosh3, Rabindra Nath Shaw4

1Centre for Development of Advanced Computing, Kolkata 700 091, India.

2Dept. of Information Technology, Maulana Abul Kalam Azad University of Technology, Nadia 741 249, India.

3School of Engineering and Applied Science, Neotia University, Kolkata, India

4Bharath Institute of Higher Education And Research, Chennai, india

[{samikshandas5@gmail.com, amitava.akuli@gmail.com, suparna.parua.sp@gmail.com, anidey007@gmail.com](mailto:%7bsamikshandas5@gmail.com,%20amitava.akuli@gmail.com,%20suparna.parua.sp@gmail.com,%20anidey007@gmail.com), <ankushghosh@gmail.com>, r.n.s@ieee.org}

\* Corresponding Author

*Abstract—* This paper proposes a machine vision solution for classification of cocoa beans using structural characteristics like shape and size of the cocoa beans extracted from images using classic machine learning techniques. The features are extracted from images through a chain of image processing techniques. Finally traditional machine learning techniques (KNN, SVM, Decision Tree and Random Forest) are employed to classify the cocoa beans in four classes e.g. large, medium, small and rejected. Comparative studies among different techniques are also performed. Prior to train the model, extracted features are optimized by two optimization techniques- Univariate Selection and Feature Importance. Trained models are evaluated using stratified K-fold cross validation and finally mean cross validation scores are calculated for performance analysis. The experimental results show that the Random Forest Classifier provides highest accuracy score of 0.75.

Keywords— Cocoa beans, KNN, SVM, Decision tree, Random Forest, Feature Optimization, Classification.

# Introduction

Cocoa is the key ingredient in chocolate and chocolate confections. Chocolate is incredibly popular, and it is one of the common foods in the world. About 300 to 600 cocoa beans are required to make 1 kg of chocolate, but this may vary depending on the cocoa content in the beans. The cocoa nibs, cocoa paste (mass or liquor), butter, powder and couverture etc. are the byproducts of the cocoa beans. These are the main ingredients to make chocolate and other food products. The beans are also used to make cosmetics and soaps. Cocoa beans are produced in tropical zones around the Equator. About 70 percent of the world's cocoa beans come from West African countries like Ghana, Nigeria Ivory Coast, Indonesia and Cameroon.

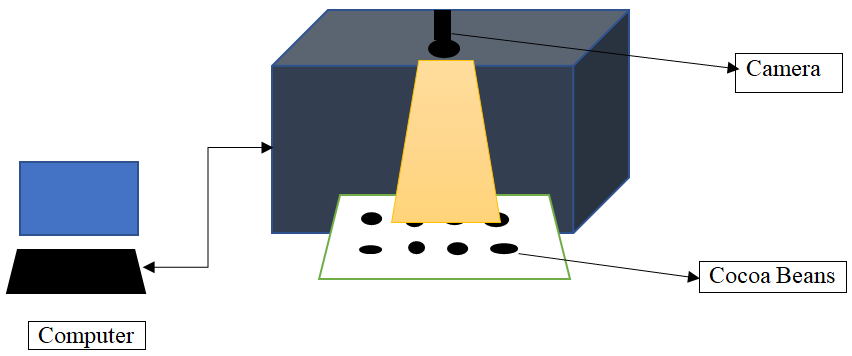
After plucking of cocoa beans from the tree, it undergoes through a post-harvesting treatment. The first chemical process is fermentation [1]. It is one of the most important operations because it improves the final quality and assuring the development of the cocoa flavor with sweet aroma. The degree of fermentation [2] is strongly correlated with cocoa quality attributes such as reducing sugars, free amino acids and bean PH. Likewise, a good fermentation contributes to the reduction of astringency and bitterness of cocoa. The next stage in the process is drying, which diminishes acidity levels in the cocoa beans. Beans are then roasted. Roasting helps to control the final flavour of the beans. So, the possible quality parameters are there like imagery evaluation and sensory evaluation. In our scope we have done the Imagery Evaluation [3].

The quality of cocoa beans is determined by its size, shape and texture. The beans are categorized as large beans, medium beans, small beans and rejected beans. Quality examination of cocoa beans is generally done using manual procedures; i.e., using the visual inspection. Accuracy of human inspection armed with experience depends on individual adaption, choice, mental state etc. Manual inspection of individual cocoa beans is qualitative, tedious and subjective. Hence, manual analysis may not suitable for routine checks on the quality of commercial cocoa beans. Therefore, a quick and reliable method is desired to classify cocoa beans for quality control. Machine vision techniques have been employed in the assessment of the quality of agricultural and food products in recent years. Automation technology backed up by artificial intelligence can be gainfully used to eliminate the limitations of manual inspection. Computer vision [4] which combines image analysis and machine learning [5] techniques are implemented to provide automated inspection. Here, images can be analyzed and processed to get useful information to the user. Structural features like size, shape, texture features are extracted from the image. In order to remove the redundant and unimportant features, features are optimized using two feature optimization techniques i.e. Univariate Selection and Feature Importance. The key objective of this study is to determine the possibility of using size, shape and texture features to determine cocoa bean quality. For classification four machine learning algorithms are adopted to determine the cocoa bean quality. And a comparison study has been made on the basis of the performances of these four algorithms ie. KNN [6], Support Vector Machine [7], Decision tree [8], and Random Forest [9] after testing on the cocoa bean test dataset. Experimental results are presented accordingly.

The rest part of the manuscript is planned as given below. Section II explains the proposed technique. The tentative outcomes on cocoa beans databases are available in Section III. Finally, Section IV includes the conclusion.

# PROPOSED METHOD

In this manuscript, our algorithm is based on discrimination of cocoa beans using structural image features. Indian samples are collected from the market and the data collection is done in the form of digital data or images of cocoa beans. Before taking the images, beans are placed on white background. 25 beans per image are preferred. Images of the cocoa beans are captured with the help of a digital image capturing setup. The image capturing setup comprises of a digital colour camera and a controlled illumination system placed inside an enclosed cabinet. The e-COCOA Vision system comprises of image acquisition from an input device to analyse and finally grading cocoa sample based on predefined criteria. Fig. 1 describes the system setup for cocoa beans image capturing. A portable image capturing setup has been made using 20 LEDs fitted throughout the roof of the cabinet equidistant from each other. A Logitech C920 webcam is there to capture the image. Color of the cabinet is made of aluminium sheet and painted with black colour to avoid the unnecessary reflection.



**Fig 1**: System setup for image capturing



**Fig. 2** cocoa beans on white paper

Capture Cocoa beans images

(

Preparing the Dataset

Scaling Data

Feature Optimization

Univariate Selection

Feature Importance

Randomize and split data into training and testing dataset

Building Classification models

Performance analysis using testing data.

Feature Extraction

**Fig. 3.** Block diagram of the proposed method

The digital images of cocoa beans comprises of 4 classes of cocoa beans (Fig. 2) where 3 classes are of whole beans and are categorized as (1) large bean (2) medium bean (3) small bean and the rest are categorized as (4) rejected beans which are fragmented. Images of 220 beans were taken for experimentation. Among 220 images, 70% were taken for model training and 30% were used for testing. The workflow diagram of the system is shown in Fig. 3.

1. Data Pre-processing:

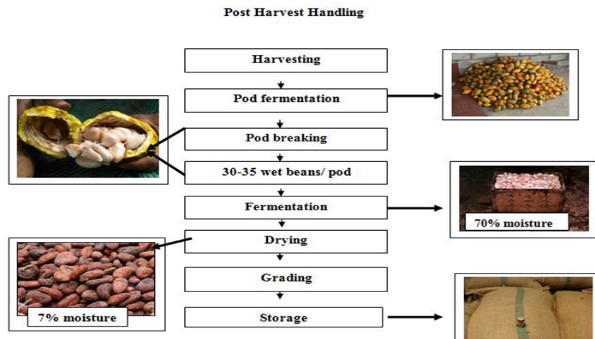
After data collection, data needs to be processed to enhance the quality of the image eliminating the background. The following steps are followed.

* *Gray image conversion:* We are working with 24 bit RGB image. We have converted the RGB image to 8 bit gray scale image. Image analysis using gray scale helps us to eliminate white background.
* *Image Segmentation:* A global thresholding technique using OTSU has been used for image thresholding. Output image after applying the thresholding technique yields with a binary image.
* *Smoothing with Gaussian filter:* Smoothing technique has been applied using a Gaussian smoothing filter with kernel size 3. This helps to eliminate the high frequency noise in the image.
* *Object Identification:* Erosion technique is applied for identifying and removing small particles which are adjacent to the image boundaries. Finally objects are identified based on the area of the particles in the image.





(a)



(b)

**Fig 4:** Images sample of Cocoa beans (a); Cocoa beans process flow for harvesting and post-harvesting (b).

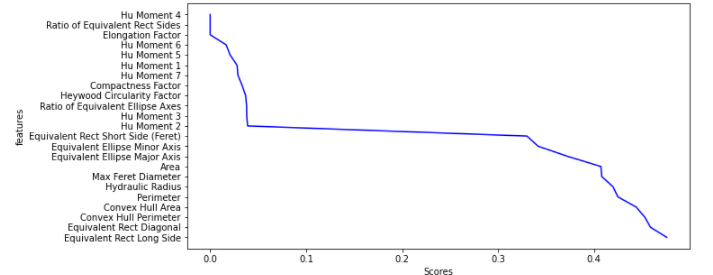
B. Feature Extraction:

A set of 23 image features are extracted from the images which 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.

C. Feature Optimization:

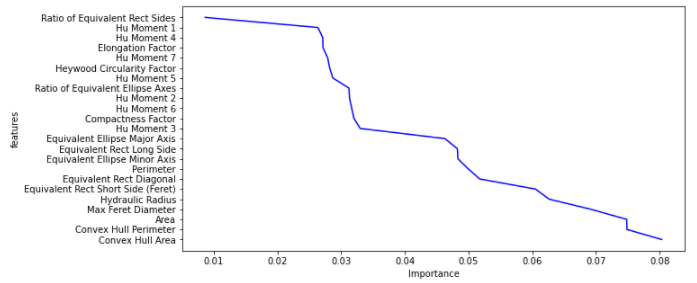
Generally for a machine learning model, all the independent features in the dataset do not impact the dependent feature in same measure. Some features may have very less impact. Feature optimization [10] is done to eliminate the redundant features to improve the machine learning models. It reduces the training time and complexity of the models without compromising the accuracy. In this study, two feature optimization techniques have been used Univariate analysis and feature iportance.

For Univariate selection[11], Scikitlearn provides the SelectKBest class that works with a suite of different statistical tests and based on the scores of these tests the correlation between each feature with the target label is measured. . The statistical tests available in this class for classification 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. Keeping in mind that the independent features here are continuous and dependent features are categorical the mutual information test is selected.



**Fig.5** Line chart for univariate feature selection

Feature importance [15] refers to a class of methods that provides scores for each independent feature in a prediction model based on the importance of the feature in making accurate predictions. The higher the score the more relevant the feature towards the target label. In this study ‘ExtraTreeClassifier’ method available in scikitlearn library is used for measuring feature importance. It implements a number of randomized decision trees as estimators using different subsets of main dataset to calculate the importance of each feature and select the top relevant features.



**Fig.6** Line chart of feature importance

After applying univariate selection and feature importance optimization techniques the relation between each of the individual features and the target dependent feature can be visualized by the line charts in Fig. 5 and Fig. 6. In both cases after the first eleven features with largest scores there is a sharp change in the curve that suggests that these eleven features have the most impact and the rest of the features are comparatively less relevant in terms of predicting the target class labels. Therefore these eleven most relevant features are selected to build the classification models.

D. Feature scaling:

Algorithms like KNN and SVM are affected by the range of features because they use the distance between the samples to determine the similarity between them. For this reason the independent feature set is rescaled ranging them in between 0 and 1 by Min-Max Normalization before training KNN and SVM models. Here, the formula denoted as follows:

On the other hand tree based classifiers are not sensitive towards the scale of the features, so this classification models can perform well without rescaling the feature set.

# Experimental Results

The performance of the proposed method has been tested on the cocoa beans database. Based on the classification problem in terms of algorithms two different type of supervised approach [16] for classification has been taken- Distance and Tree based algorithms. The programming language used to develop the classification model is python 3.9.7 and the necessary python libraries which were imported are matplotlib, pandas, numpy, seaborn, sklearn, pydotplus and six.

Accuracy score is the sum of True Negative and True Positive divided by the sum of True Negative, True Positive, False Positive and False Negative. Here, formula defined as follows:

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.

Using the ‘accuracy\_score’ and ‘f1\_score’ method available in Scikitlearn library the accuracy score and F1 score is obtained for all the four classification models.

For Distance based algorithms, two traditional classification algorithm K-Nearest Neighbour (KNN) and Support Vector Machine (SVM) are used. KNN classifies cases based on the similarity and this similarity is measured by a distance matrix such as Euclidean Distance [17], Manhattan Distance[18], Minkowski Distance or Hamming Distance[19]. Cases those are near to each other are said to be ‘Neighbours’. While predicting classes for unknown data point the most popular class label or class label with the majority value from its neighbours is considered as the class label for the unknown data point. On the other hand SVM is efficient in handling the non-linearity of dataset by transforming the data to a higher dimensional space and then classification is performed by finding the best hyperplane that differentiates the classes very well. Although SVM is also memory efficient as it uses a subset of the training data in the decision function but the training time is higher than KNN as KNN does not derive any discriminative function from the training data, it stores the training dataset and learns from it only while making real time predictions but SVM learns during training period.

The KNN classification model is trained with the training dataset having K value in a range from 1 to 20. And using the testing dataset the best minimum value for K is determined to be 10 for which the model predicts with maximum accuracy score 0.73 and F1 score 0.68. While Using the Polynomial Kernel function the SVM model gives maximum accuracy score 0.73 and F1 score 0.71. Accuracy score and F1 Score for different kernel functions are mentioned in Table 1.

Table 1: Performance evaluation of different kernel function.

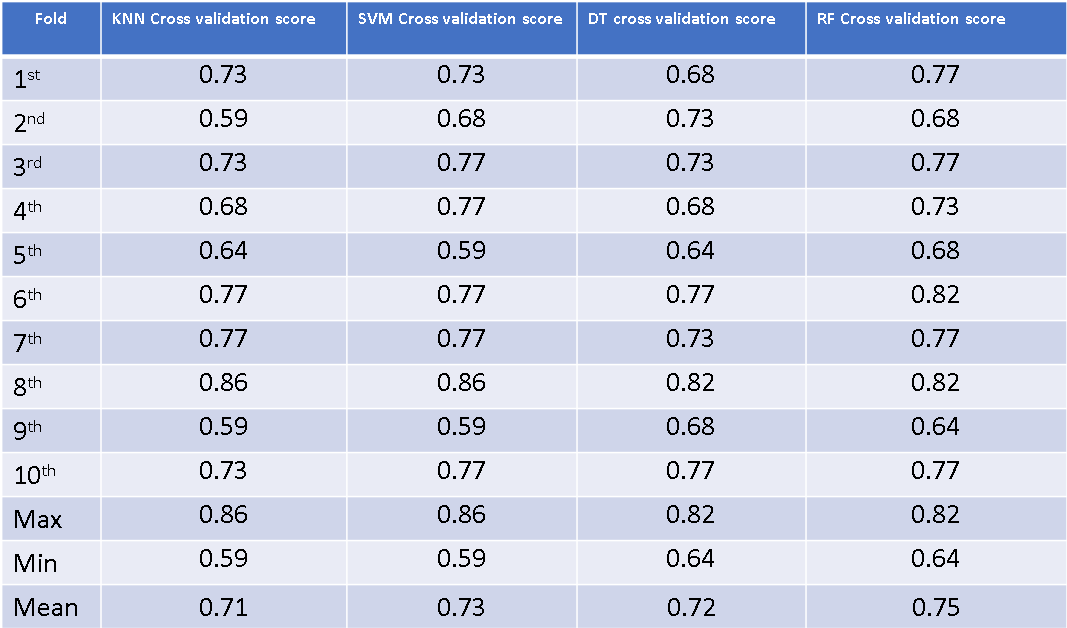
|  |  |  |
| --- | --- | --- |
| **SVM Kernel Function** | **Accuracy Score** | **F1 Score** |
| Linear | 0.68 | 0.61 |
| RBF | 0.73 | 0.68 |
| Sigmoid | 0.53 | 0.38 |
| Polynomial | 0.72 | 0.71 |

For Tree based algorithms, two popular algorithms Decision Tree Classifier and Random Forest classifier are used. Decision Tree classifier is a tree structured classifier where the branches represent the decision rules, the internal nodes represent the features of the sample dataset and the leaf nodes represents the final output or class labels. It uses Recursive Partitioning [20] to split the training records into segments by minimizing the impurity at each step. But whenever decision tree is built to its complete depth it comes with low bias that means the model gets overfitted to the training dataset and high variance which suggests that the model is prone to give large amount of errors while working with new test data. In Random Forest Classifier instead of using a single decision tree multiple decision trees with high variance generated from the subsets of the main dataset are considered and by combining the trees with respect to a majority vote the high variance gets converted into low variance. One more thing is if we change some data or add some new data to our model it would not affect much because the changes will be distributed to all the decision trees while we are doing random sampling of the rows and columns.

For both of these Tree Based algorithms the criterion function selected is Entropy [21] for selecting the root node or internal nodes at different level of the decision trees. The goal is to find the tree with smallest entropy in its nodes.So for Decision Tree model criterion function selected for splitting is ‘entropy’ with ‘best’ splitter strategy and maximum depth for the decision tree is determined to be 4 after trying a range of values from 1 to 10 to achieve the maximum accuracy score 0.74 and F1 score 0.73. While the random forest classifier is trained with same criterion function for 150 decision trees with maximum depth 4 to achieve the maximum accuracy score 0.73 and F1 score 0.71.

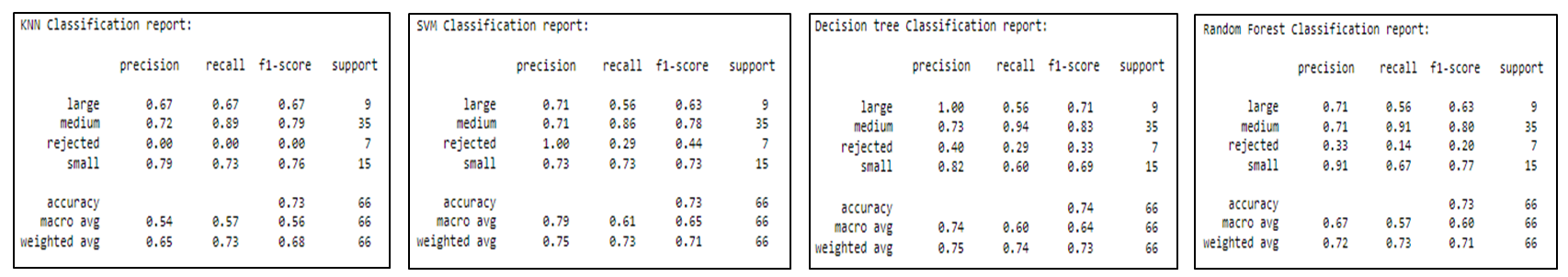
As the dataset is not perfectly balanced, at the end the four classification models are evaluated using stratified K-fold cross validation [22] with 10 folds. It ensures the proportion of target features of different classes is same across the original data, training data and testing data. The performance evaluation of the four algorithms using Stratified K-fold cross validation is mentioned in Table 2.

Table 2: Performance evaluation of cross validation using KNN, Decision Tree, SVM, Random Forest.



# CONCLUSION

By training and testing the classification models using structural feature set with four traditional classification algorithms KNN, SVM, Decision Tree and Random Forest it can be concluded that the resultant accuracy scores and F1 scores achieved by these models are in range between 0.72 to 0.75 and 0.68 to 0.73 respectively. The cross validation score reflects that out of these four algorithms Random Forest Classifier classifies the Cocoa beans with maximum accuracy.

****

**Fig. 7**. KNN, SVM, Decision Tree, Random Forest Classification Result

##### 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*,

[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.

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

[4] M. M. Oliveira, B. V. Cerqueira, S. B. Douglas, F. Barbin, Classification of fermented cocoa beans (cut test) using computer vision, Journal of Food Composition and Analysis, Volume 97,2021, 103771.

[5] A. I. Khan, S. Al-Habsi, Machine Learning in Computer Vision, Procedia Computer Science,Volume 167,2020,Pages 1444-1451.

[6] G. Guo, H. Wang, D. Bell, Y. Bi,. (2004). KNN Model-Based Approach in Classification.

[7] T. Evgeniou, M. Pontil, (2001). Support Vector Machines: Theory and Applications. 2049. 249-257.

[8] H. Patel, P. Prajapati, (2018). Study and Analysis of Decision Tree Based Classification Algorithms. International Journal of Computer Sciences and Engineering. 6. 74-78.

[9] J. Ali, R. Khan, N. Ahmad, I. Maqsood, (2012). Random Forests and Decision Trees. International Journal of Computer Science Issues(IJCSI). 9.

[10] L. Abualigah, A. Aldulaimi, M. Al Shinwan, A. Khasawneh, H. Alabool, M. Diabat, M Shehab, (2020). Optimization Algorithms to Solve Feature Selection Problem: A Review. International Journal of Science and Applied Information Technology. 8. 10.30534/.

[11] R. Subho, Md. Chowdhury, D. Chaki, S. Islam, Md. Rahman, (2019). A Univariate Feature Selection Approach for Finding Key Factors of Restaurant Business. 605-610.

[12] E. Ostertagova, O. Ostertag, (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.

[14] R. Singhal, R. Richa, R. 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).

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

[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.

[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.

[19] B.Abraham, K Vladimir, R. Timo. (2002). Generalized Hamming Distance. Information Retrieval. 5. 10.

[20] C. Strobl, J. Malley, G. Tutz, (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] 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.

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