A COMPARATIVE CLASSIFICATION ANALYSIS OF ROASTED COFFEE BEAN IMAGES

Lex-Anthony T. Bungabong
University of Southeastern Philippines
ltbungabong@usep.edu.ph

Cyrelle John A. Domingo
University of Southeastern Philippines
cjadomingo@usep.edu.ph

John Mark B. Diacono
University of Southeastern Philippines
jmbdiacono@usep.edu.ph

ABSTRACT

Coffee has been one of the most important beverages in the Philippines. It's considered as a fuel for the Filipinos to kick start their day with coffee. The demand for coffee is expected to surge in the Philippines in the coming years. And It is an arduous task for many to classify which coffee bean variety is which. Hence, it is high time to conduct a study regarding coffee bean classification.

This resulted in the development of a digital image analysis technique based on morphological and textural characteristics to distinguish three distinct varieties of coffee beans, namely Espresso, Kenya, and Starbucks Pike Place. The dataset consists of 1,554 photos and was obtained from Springer (with over 500 images each variety).

From the dataset of roasted coffee bean images, twelve morphological and four textural features were extracted for classification analysis. The morphological features are area, perimeter, length, width, main axis length, minor axis length, feret diameter, roundness, rectangular aspect ratio, and elongation. The textural characteristic consists of energy, entropy contrast, and homogeneity.

We have compared three classical algorithms of classification approaches of SVM, Naive Bayes and K-NN classifiers on each classification parameters of morphology and texture.

It was found that the classification performance of SVM with sixty four percent accuracy was the best among the three classical algorithms compared.

Keywords: Roasted Coffee Bean Varieties, Classification, Image Analysis, Pre-processing, Image Segmentation, Classical Algorithms

1. INTRODUCTION

During the 16th and 17th centuries, coffee was brought to a number of nations in Europe for the very first time. According to a number of accounts, its application as a religious practice, a political strategy, or a medical therapy has either been prohibited or approved [1]. Coffee has a significant impact on how people go about their lives on a day-to-day basis in the current day. Coffee is enjoyed by a wide variety of people, not only the Filipinos. In addition to its use as a stimulant, it has evolved into a beverage that is enjoyed for its recreational value by a growing number of individuals.

While the majority of coffee consumption takes place in affluent nations, over 90% of coffee is produced in developing nations, mainly in South America. The Philippine Coffee Board states that in 1880, it was thought the country was the fourth-largest exporter of coffee beans. In fact, The Philippines was the only country that exported coffee beans when the coffee rust epidemic began in Ceylon and spread throughout the region. However, coffee rust also struck the Philippines in 1890, wiping off almost all of the country's coffee trees in the province of Batangas[1]. As a result, affluent landowners and coffee producers were forced to alter their farms from cultivating coffee to growing sugarcane. Even if there is some coffee being grown now, the Philippines has never again produced as much coffee as it did in the past.

The four commercially viable Arabica, Robusta, Liberica, and Excelsa kinds are now only produced in a small number of nations, including the Philippines [1]. Locally referred to as Barako/Baraco, Philippine Liberica is a coffee bean that makes an incredibly robust and potent cup of coffee. Primarily farmed in the provinces of Batangas and Cavite, the larger cherries and beans are a mainstay in the Philippine coffee industry[2].

One of the most crucial elements that affect how coffee tastes in the cup is how much the coffee beans have been roasted. Green coffee beans are seen to be soft and have a "grassy" fragrance before roasting. By using extremely high temperatures, roasting caused chemical changes. The roasting process is stopped when the green coffee bean reaches its maximum temperature [3]. The most accurate way to describe roasted coffee beans is by observing their color. Coffee beans with dark roasts have a flat, bitter flavor with a strong smoky flavor. The medium-roasted coffee beans have a deeper brown color. And lastly, lightly roasted coffee beans have the most acidity and caffeine. Some roasted coffee bean oil may be present. The color of a roasted coffee bean is an accurate predictor of its flavor, aroma, and acidity.

The Filipino population consumes as well as produces. Asia's second-largest consumer of coffee is the Philippines. An estimated 3.78 kilograms of coffee per person will be consumed by 2025, up from 3.05 kilograms per person in 2021 (Cigaral, 2021).

With all the data presented, it clearly shows that coffee in the Philippines is widely recognized as one of the most consumed, produced, and one of the most popular beverages in the country.

1.2 PROBLEM

In the year 2020, the coffee industry in the Philippines suffered greatly from the pandemic, but with the restrictions being lifted from the establishments, and the rise of new coffee shops being built in the country, especially in Davao, the demand for coffee is expected to surge in the coming years.

Most of the coffee shops do not have Roasters - an expert in making roasted beans, instead, they buy local or international beans from the market. Hence, comes the problem, as they bought the roasted beans from the coffee market, the only thing they know is the roasting date and the number of grams of the coffee bean. Little do they know that the beans that they have may or may not be of good quality and most of them are mixed. Also, few Roasters, coffee enthusiasts, and individuals cannot really determine the quality or the distinctiveness of roasted coffee beans.

Depending on where it is cultivated, the coffee's form or shape might change. Hence, the customer finds it arduous to distinguish between different types of coffee owing to the similarities in taste, color, and form[4].

Coffee beans are usually classified according to expert opinion and tradition. Processing and sorting the characteristics of the coffee beans, however, takes time. and it is also incredibly costly and inefficient[4].

1.3 PURPOSE

The purpose of the study is to determine which classification model is more accurate as well as understanding the differences between coffee varieties with images of coffee beans, which in return, would help countless coffee shop owners in their businesses.

2. LITERATURE REVIEW

There are a handful of studies regarding the classification of roasted beans, but there are just a few specifically for the classification of defects of green beans using machine learning algorithms.

A. *Application of Pre-trained Deep Convolutional Neural Networks for Coffee Beans Species Detection*

This study's goal was to categorize three different coffee beans using their photos, using the transfer learning method and four different Convolutional Neural Networks-based models: SqueezeNet, Inception V3, VGG16, and VGG19. The dataset used to train the models was produced specifically for the study. The method acquired a total of 1,554 coffee bean photos of Espresso, Kenya, and Starbucks Pike Place coffee varieties, The collected photos were used in the model training and model testing processes. The cross-validation approach was employed to test the models[7].

B. *Grading Ethiopian Coffee Raw Quality Using Image Processing Techniques*

Using the HSB color characteristic, this study aims to use image processing techniques to sample coffee raw quality value grading.

A total of 145 image files and 10,000 coffee beans of varying grades were utilized. The Gaussian filter was applied to minimize noise, and contrast enhancement was employed to improve the picture quality of coffee beans.

In the image segmentation procedure, normalization and binarization were employed by thresholding the 8-bit images, which is a technique for dividing the image into areas[6].

C. *Image Analysis for Ethiopian Coffee Classification*

A computerized image analysis approach based on morphological and color features was developed in order to identify diverse varieties of Ethiopian coffee based on their growing area. For classification analysis, ten morphological and six color characteristics were extracted from each image of a coffee bean[7].

D. *Classification of Coffee Bean Degree of Roast Using Image Processing and Neural Network*

This research classifies the degree of roasting utilizing image processing and artificial neural networks. One coffee type, Excelsa, and one coffee origin were utilized to produce the method (Indang, Cavite). Using a smartphone, a snapshot of roasted Excelsa coffee beans was captured. In addition, using RGB values as input, a neural network was utilized to categorize the degree of roasting of coffee beans as light roast, medium roast, and really dark roast[8].

3. DATASET

This study is concerned with constructing a classification model for roasted coffee bean seed varieties. The dataset utilized in the models' training is taken from Springer and it composes a total of 1,554 roasted coffee bean images of Espresso, Kenya, and Starbucks Pike Place with over 500 images each. The dataset utilized in the study contains three varieties of coffee beans cultivated in various locales. Espresso roasted bean is from Ethiopia, Kenya roasted coffee bean is from Kenya, and Starbucks Pike Place roasted coffee bean is from Costa Rica, Mexico, and Columbia (Seninde and Chambers 2020).

4. METHODOLOGIES

4.1 Image Preprocessing

For us to classify the roasted coffee bean images, we need to first convert these images into numerical values. We have chosen to use the ImageJ, now also known as *Fiji*, a Java-based image processing application for the pre-processing technique.

4.2 Feature Extraction

Extraction of meaningful information from an image is referred to as Feature Extraction. A procedure described in the preceding section was used to preprocess the photos gathered for the investigation. Morphological and textural characteristics are the primary features taken from the images. Comprising a total of 16 features, including 12 morphological and 4 textural ones.

Morphological Features

Morphological characteristics, which determine the shape and size of the beans, are the most significant part for categorizing the varieties of coffee beans. The classification of coffee beans based on their quality relied heavily on morphological trait(9)s. The following features are as follows:

Area: This feature counted the number of pixels that were filled by the coffee bean in order to arrive at an estimate of its size.

Perimeter: This was the total quantity of pixels that the border or boundary of the item took up in its entirety.

Length: It is the length of the smallest rectangle that can enclose a coffee bean in its whole.
Width: It is equal to the width of the smallest rectangle that can completely enclose a coffee bean.

Major Axis Length: It is the distance between the two end points of the line that can be traced through the coffee bean that is the longest possible line. The end points of the main axis may be determined by computing the pixel distance between each possible combination of border pixels in the coffee bean boundary and locating the pair of pixels that have the longest distance between them.

Minor Axis Length: It is the distance between the two end points of the longest line that could be traced through the coffee bean while staying perpendicular to the primary axis of the bean. The longest line that could be traced through the bean is called the longest line that could be traced through the bean.

Roundness: This feature measures the degree of the roundness or the circularity of the shape of the coffee bean. It is formulated as $R = 4P\pi A$, where: A = area and P = perimeter. The value of R is somewhere in the range of 1 to 0. If the value of R is equal to 1, then the region in question is a perfect circle; on the other hand, if the value of R is equal to 0, then the region that it corresponds to, does not have any area.

Feret Diameter: It is the distance around the circumference of a circle that has the same surface area as a coffee bean. And it is also formulated as $D = 4A/\pi$, where A is the area of the coffee bean region.

Rectangular Aspect Ratio: This feature is the ratio of length to width of the coffee bean.

Elongation/Eccentricity: This characteristic may be defined as the ratio of the distance between the ellipse's major axis length and its foci.

Textural Features

The Coffee Bean picture has texture attributes that represent the structural organization of surfaces and the intensity variation of objects. They make it possible for us to visually examine the patterns and surface characteristics of coffee beans(8). The textural features are as follows:

Energy: This feature determines the proportion of intensity pairs that occur together in the co-occurrence matrix.

Entropy: This feature determines the degree to which the intensity distribution is influenced by random factors.

Contrast: This feature compares the strengths of the image's intensity and measures the difference between them.

4.3 SVM

To maximize resilience with regard to isotropic uncertainty tolerance, a support vector machine (SVM) is developed to fit a linear boundary between two binary problem samples. Linear, polynomial, and radial basis functions (RBFs) are only a few examples of the many types of functions used to make the transformation from input space to target function.

4.4 Naive Bayes

The Naive Bayes Classifier is one of the most straightforward and efficient classification algorithms based on the Bayes theorem. It aids in the development of quick machine learning models capable of making accurate predictions. It is a probabilistic classifier, which means that it makes predictions based on the likelihood that an object will appear. The Bayes theorem, commonly referred to as Bayes' Rule or Bayes' law, is used to calculate the likelihood of a hypothesis given some prior information[12]. The conditional probability determines this. The formula shows as follows:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

4.4 KNN

One of the most prominent data mining approaches is the k-nearest-neighbors (KNN) method, which attempts to categorize an unknown sample based on the known classification of its neighbors.

KNN works by selecting the number K of the neighbors and then calculating the Euclidean distance of K number of neighbors, taking K nearest neighbors as per the calculated Euclidean distance, then among the K neighbors, we can count the number of the data points in each category and then assigning the new data points to that category for which the number of the neighbor is maximum.

The K-NN Algorithm is one of the simplest to implement, it is also robust to noisy training data and lastly, it can be more effective if the training data is large[11].

4. RESULTS AND DISCUSSION

As the dataset photos were transformed into data numerical values through preprocessing and the use of various classification algorithms, the following images show the results of applying hyperparameter tuning of each algorithm along with the accuracy model.

Support Vector Machine (SVM

The best estimator across ALL searched parameters:

```
The best estimator across ALL searched params:
 SVC(C=33.75012868869448, degree=5, kernel='linear', random_state=7)

 The best score across ALL searched params:
 0.6513506625891947

 The best parameters across ALL searched params:
 {'C': 33.75012868869448, 'degree': 5, 'gamma': 'scale', 'kernel': 'linear'}
CPU times: total: 1.42 s
Wall time: 16.3 s
```

The best estimate for all parameters searched in SVM is shown in the above image with a result of 0.65 or 65%.

Accuracy model:

Classification Report - SVM

```
In [7]: from sklearn.metrics import classification_report
        best_model = grid_search

        y_pred = best_model.predict(X_test)
        report = classification_report(y_test, y_pred)
        print(report)
```

```
              precision    recall  f1-score   support

           0       0.70      0.71      0.71       114
           1       0.76      0.78      0.77        97
           2       0.50      0.48      0.49       100

    accuracy                           0.66       311
   macro avg       0.65      0.66      0.66       311
weighted avg       0.66      0.66      0.66       311
```

Best Accuracy: 64.73 %

Best Parameters: {'C': 20, 'break_ties': True, 'class_weight': 'balanced', 'coef0': 9, 'gamma': 'scale', 'kernel': 'linear', 'probability': True}

CPU times: total: 6.03 s

Wall time: 8min 14s

Naive Bayes

The best estimator across ALL searched parameters:

```
The best estimator across ALL searched params:
 GaussianNB(var_smoothing=12.241014555905716)

 The best score across ALL searched params:
 0.4131583418280666

 The best parameters across ALL searched params:
 {'var_smoothing': 12.241014555905716}
CPU times: total: 219 ms
Wall time: 248 ms
```

The best estimate for all parameters searched in Naive Bayes is shown in the above image with a result of 0.41 or 41%.

Accuracy model:

Classification Report - GaussianNB

```
In [9]: from sklearn.metrics import classification_report
        best_model = grid_search

        y_pred = best_model.predict(X_test)
        report = classification_report(y_test, y_pred)
        print(report)

              precision    recall  f1-score   support

           0       0.34      0.10      0.15       114
           1       0.45      0.71      0.55        97
           2       0.35      0.44      0.39       100

    accuracy                           0.40       311
   macro avg       0.38      0.42      0.36       311
weighted avg       0.38      0.40      0.35       311
```

Best Accuracy: 43.80 %

Best Parameters: {'var_smoothing': 1e-09}

CPU times: total: 125 ms

Wall time: 215 ms**

K-nearest Neighbors (KNN)

The best estimator across ALL searched parameters:

```
The best estimator across ALL searched params:
 KNeighborsClassifier(algorithm='kd_tree', leaf_size=51, n_neighbors=9, p=9,
                     weights='distance')

 The best score across ALL searched params:
 0.3973921168752973

 The best parameters across ALL searched params:
 {'algorithm': 'kd_tree', 'leaf_size': 51, 'n_neighbors': 9, 'p': 9, 'weights':
'distance'}
CPU times: total: 156 ms
Wall time: 2.49 s
```

The best estimate for all parameters searched in K-NN is shown in the above image with a result of 0.397 or 40%.

Accuracy model:

Classification Report - KNN

```
In [38]: from sklearn.metrics import classification_report
         best_model = grid_search

         y_pred = best_model.predict(X_test)
         report = classification_report(y_test, y_pred)
         print(report)

              precision    recall  f1-score   support

           0       0.42      0.46      0.44       114
           1       0.44      0.48      0.46        97
           2       0.30      0.25      0.27       100

    accuracy                           0.40       311
   macro avg       0.39      0.40      0.39       311
weighted avg       0.39      0.40      0.39       311
```

Best Accuracy: 39.94 %

Best Parameters: {'algorithm': 'auto', 'leaf_size': 1, 'n_neighbors': 10, 'p': 5, 'weights': 'uniform'}

5. CONCLUSION

In conclusion, with the image acquisition system developed especially for the investigation, a total of 1554 coffee bean images were employed in the context of this study. Espresso, Kenya, and Starbucks Pike Place coffee beans were depicted in the photographs. Three different algorithms, including SVM, K-NN, and Naive Bayes, were employed on photos of coffee beans to compare how accurate they were.

With an accuracy of sixty-four percent, it was revealed that the SVM approach had the greatest classification performance out of the three classical algorithms that were examined.

6. REFERENCES

[1]
Philippine Coffee - Philippine Coffee Board. *Phil Coffee Board*. Retrieved December 18, 2022 from https://philcoffeeboard.com/philippine-coffee/

[2]
Edwin R. Arboleda, Arnel C. Fajardo, and Ruji P. Medina. 2018. Classification of coffee bean species using image processing, artificial neural network and K nearest neighbors. *2018 IEEE International Conference on Innovative Research and Development (ICIRD)* (2018). DOI:https://doi.org/10.1109/icird.2018.8376326

[3]
T H Nasution and U Andayani. 2017. Recognition of Roasted Coffee Bean Levels using Image Processing and Neural Network. *IOP Conference Series: Materials Science and Engineering* 180, (2017), 012059. DOI:https://doi.org/10.1088/1757-899x/180/1/012059

[4]
Brandon Damitz. 2021. How Do Coffee Beans Get Their Flavor? – Bean & Bean Coffee . *https://beannbeancoffee.com/blogs/beansider/how-do-coffee-beans-get-their-flavor*. Retrieved December 18, 2022 from https://beannbeancoffee.com/blogs/beansider/how-do-coffee-beans-get-their-flavor

[5]
Peaberry and normal coffee bean classification using CNN, SVM. Retrieved December 18, 2022 from https://www.aimspress.com/aimspress-data/aimsagri/2022/1/PDF/agrfood-07-01-010.pdf

[6]
Muktar Bedaso, Million Meshesha, and Chala Diriba. 2022. Grading Ethiopian Coffee Raw Quality Using Image Processing Techniques. (2022). DOI:https://doi.org/10.21203/rs.3.rs-1980632/v1

[7]

Habtamu Minassie Aycheh. Image Analysis for Ethiopian Coffee Classification and Grading. *Shawndra*. Retrieved October 22, 2022 from http://shawndra.pbworks.com/f/IMAGE+ANALYSIS+FOR+ETHIOPIAN+COFFEE+CLASSIFICATION.pdf

[8]
Karpaga Se;vo Subramanian and Tsadkan Gebremichael. 2019. Features Extraction and Dataset Preparation for Grading of Ethiopian Coffee Beans Using Image Analysis Techniques. (January 2019). Retrieved 2022 from https://www.researchgate.net/publication/330004670_Features_Extraction_and_Dataset_Preparation_for_Grading_of_Ethiopian_Coffee_Beans_Using_Image_Analysis_Techniques_Proceedings_of_ICTIS_2018_Volume_1

[9]

Zach Daggett. 2015. How does altitude affect coffee and its taste in the Cup? (June 2015). Retrieved December 22, 2022 from https://perfectdailygrind.com/2015/06/how-does-altitude-affect-coffee-and-its-taste-in-the-cup/

[10]

Bruno Stecanella. Support Vector Machines (SVM) Algorithm Explained. *Monkey Learn*. Retrieved November 22, 2022 from https://monkeylearn.com/blog/introduction-to-support-vector-machines-svm/

[11]

Yavuz Unal, Yavuz Selim Taspinar, Ilkay Cinar, Ramazan Kursun, and Murat Koklu. 2022. Application of Pre-Trained Deep Convolutional Neural Networks for Coffee Beans Species Detection. *Food Analytical Methods* 15, 12 (2022), 3232-3243. DOI:https://doi.org/10.1007/s12161-022-02362-8

[12]
Prashant Banerjee. Naive Bayes Classifier in Python | Kaggle. Retrieved October 24, 2022 from https://www.kaggle.com/code/prashant111/naive-bayes-classifier-in-python