Preprints.org

Article

Roasted Arabic Coffee Classification Based on Color, Chemical Compositions and Antioxidants Using Machine Learning

Eman S. Alamri [*] , Ghada Altarawneh , Hala M. Bayomy , Ahmad B. Hassanat [*]

Article

Roasted Arabic Coffee Classification Based on Color, Chemical Compositions and Antioxidants Using Machine Learning

Eman S. Alamri [1,‡,*], Ghada A. Altarawneh [2,‡] (iD), Hala M. Bayomy [1,3,‡] and Ahmad B. Hassanat [4,‡,*] (iD)

[1]   Food Science and Nutrition Department, University of Tabuk, Tabuk 47512, KSA; ialamri@ut.edu.sa

[2]   Accounting Department, Mutah University, Karak 61711, Jordan; Ghadaa@mutah.edu.jo

[3]   Food Science and Technology Department, Damanhour University, Damanhour 22516, Egypt; hm.mohamed@ut.edu.sa

[4]   Computer Science Department, Mutah University, Karak 61711, Jordan; Ghadaa@mutah.edu.jo

*   Correspondence: Hasanat@mutah.edu.jo (H.A.), ialamri@ut.edu.sa (E.A.)

‡   These authors contributed equally to this work.

Abstract: This research assesses the classification of Arabic coffee into three primary variations (light, medium, and dark) using simulated data based on actual measurements of color information, antioxidant laboratory tests, and chemical composition tests. Two types of simulated data were generated, with the standard deviation of the measures varied. Multiple classifiers were used for classification. The findings demonstrate the use of color information in accurately distinguishing Arabic coffee types. The different CIE color values, as well as the excellent connection between color information and coffee classes, lead to flawless classification results. However, depending primarily on antioxidant information results in lower classification performance, with variability amongst classifiers due to increasing data complexity. Chemical composition information, on the other hand, possesses outstanding discriminatory strength, allowing faultless classification on its own. Certain characteristics, such as crude protein and crude fiber, show strong correlations and are critical in the classification of coffee types. Based on the findings, it is recommended to create a mobile application that uses image recognition to analyze coffee color while providing chemical composition information. End users, such as consumers, will be able to make informed judgments regarding their coffee preferences owing to this application.

Keywords: smart sustainability; smart food; Arabic coffee; classification; color information; antioxidant information; chemical composition; mobile application

---

1. Introduction

Arabic coffee in the Arab world is considered an important part of Arab culture and tradition, particularly, in Saudi Arabia, and surrounding Arab countries, where it is said to be the most popular hot beverage [1]. Arabica coffee beans are the primary ingredient [2]. An estimated 1.4 billion cups of coffee are consumed daily throughout the world. Consumption of coffee has significantly grown in Saudi Arabia, where 18,000 tons of coffee are imported each year at a cost of $54 million [3]. According to recent estimates, the typical Saudi adult consumes between 60 and 300 mL of Arabic coffee in a single sitting, amounting to 1.6 kg of coffee per person yearly [2].

Coffee drinking has certain beneficial effects on human health as a result of its biochemical features [4,5], especially non-communicable diseases [6]. There is proof that drinking coffee can improve several functions, such as memory, mood, and cognitive performance [7–9].

On the other side, coffee can raise LDL-C and the body's total cholesterol, which can increase the risk of cardiovascular disease [10,11]. Additionally, drinking Arabic coffee was linked to a significant osteoporosis increase in Saudi females over the age of 40 [12]. Regardless of whether drinking coffee is

harmful or beneficial, it is apparent that there are many people who do so worldwide, especially in Saudi Arabia when it comes to Arabic coffee.

Arabic coffee comes in several types with varying levels of roasting. The roasting temperature and roasting time affect the degree of roasting, and these two factors determine whether the coffee is a light, medium, or dark roast in terms of color [13]. The biological activity and chemical composition of the coffee can be altered significantly during the roasting process. According to Wang et al. [14], certain components, such as natural phenolic compounds, can be lost while other elements, such as antioxidants, including Maillard reaction products, can be generated. As a result, antioxidants can be either preserved or enhanced.

Hence, it is crucial to be able to categorize coffee into one of its three common roasting degrees (light, medium, or dark) for human consumption since each has a different effect on health. For instance, Alamri et al. [3] recommended medium Arabic coffee as having a high concentration of chemicals with antioxidants and other biologically advantageous effects.

The interesting research of Alamri et al. [3] examined the chemical composition of typically roasted Arabic coffee at three distinct roasting levels: light, medium, and dark. To analyze various facets of coffee, the researchers used a variety of methodologies. They measured the amount of caffeine using a UV-visible spectrophotometer, the amount of acrylamide using a gas chromatograph, and the amount of free radical scavenging power using the 1,1-Diphenyl-2-picryl-hydrazy (DPPH) technique. They also used Gas Chromatography-Mass Spectrometry (GC-MS) to separate and distinguish the volatile components contained in the coffee samples, and they estimated the browning index. A temperature of $180 \pm 10°C$ was held for around 6.0 ± 1.0 minutes for the light roast, 8.0 ± 1.0 minutes for the medium roast, and 10.0 ± 1.0 minutes for the dark roast. Their laboratory experiments were conducted on nine Arabic coffee samples, three for each type. According to their findings, roasting Arabic coffee alters its chemical compositions such as moisture content, ether extract, crude protein, crude fiber, ash content, nitrogen-free extract (NFE), caffeine content, acrylamide levels, and DPPH activity, and Tables 1 and 2 demonstrate that each roasting degree has unique chemical component value. in addition, they provided distinctive color information for each coffee type using the L*a*b* color system and browning index as shown in Table 3.

There are several studies in the literature designed to classify Arabic coffee and coffee, in general, using various sources of information and technologies, including coffee beans images [15–18], Infrared spectroscopy [19–23], and electronic nose [24–26]. To the best of our knowledge, none of them, however, addressed the chemical classification of Arabic coffee.

The aim of this paper is to address the classification of Arabic coffee using its chemical compounds. For this purpose we utilize the data obtained by [3] (shown in Tables 1, 2 and 3), to generate new samples for each type of Arabic coffee using the Monte Carlo method because we only have one mean value for each feature ± its standard deviation. Then, using the simulated data, we employ a variety of machine-learning approaches to classify Arabic coffee; feature importance is also explored in this study to determine which chemical components have the most influence on the classification process. It is worth noting that obtaining significant samples of real data for each chemical test involves laboratory testing, which comes at a very high cost that we cannot afford at present.

Table 1. Caffeine content, acrylamide, and free radical scavenging capacity (DPPH) in light, medium, and dark coffee [3].

Degree	Light coffee	Medium coffee	Dark coffee
Caffeine content %	1.13±0.02	1.17±0.07	1.08±0.06
Acrylamide(mg/100gm)	0.41±0.086	0.31±0.063	0.36±0.048
DPPH (mg TE/g)	88.72±2.91	84.61±1.76	78.76±2.49

Table 2. Proximate chemical composition (% in dry matter) content in light, medium, and dark coffee [3].

Degree	Light coffee	Medium coffee	Dark coffee
Moisture content %	4.80±0.24	4.30±0.17	3.89±0.28
Ether extract %	10.39±0.30	10.47±0.19	10.65±0.22
Crude protein %	13.05±0.14	12.36±0.24	11.10±0.06
Crude fiber %	24.24±0.47	28.31±0.31	28.40±0.42
Ash content %	3.95±0.26	3.89±0.08	4.10±0.17
Nitrogen free extract (NFE) %	48.37	44.97	45.76

Table 3. Color information in light, medium, and dark coffee [3].

Degree	Light coffee	Medium coffee	Dark coffee
Browning index	0.4540±0.13b	0.8600±0.13b	1.8400±0.24a
L *	58.62±2.73a	48.83±1.73b	41.04±3.06b
a *	9.75±0.44b	13.04±0.07a	13.93±0.62a
b *	31.74±1.21a	32.21±0.75a	29.80±1.52b

2. Related Work

Several studies have been conducted in the literature to classify Arabic coffee and coffee, in general, utilizing various sources methods, and technologies such as computer vision, Infrared spectroscopy, and electronic nose.

2.1. Computer vision

Because basic computer vision systems are less expensive, they consist of a digital camera used to acquire images, a standard lighting system, and software for image processing and analysis. This affordability has allowed many computer vision systems to emerge in the last decade, not only for coffee classification but also for various food issues.

For example, de Oliveira and coworkers [15] developed a basic computer vision system that produces color measurements of 120 samples of green Arabic coffee beans and classifies them into four groups using Artificial Neural Networks (ANN) and the Bayes classifier. Their system yielded 100% classification accuracy, although the conclusion is not significant owing to the small amount of data utilized.

Arboleda et al. [16] proposed a computer vision method to classify the species of coffee bean samples automatically. From a collection of 195 training images and 60 testing images, various morphological characteristics of the beans, including area, perimeter, equivalent diameter, and roundness percentage, were retrieved. The coffee beans were automatically classified using two classification algorithms: K Nearest Neighbor (KNN) and ANN. ANN achieved the best results, with a classification score of 96.66%.

Arboleda [17] employed image processing techniques and data mining algorithms to classify green coffee beans from three species: Liberica, Robusta, and Excelsa. Four features were retrieved from 255 photos of coffee beans; 85 samples were taken for each species, which were classified using 22 classifiers. The Coarse Tree Algorithm had the best accuracy of 94.1%.

Subjectivity and standards problems plague the manual classification process, resulting in potential inconsistencies. To overcome these challenges, Pizzaia et. al [18] proposed A computer vision approach using a Multilayer Perceptron (MLP) Neural Network for digital image processing of coffee bean samples. The network uses shape, size, and color as classification criteria. The use of an MLP Neural Network in coffee classification seeks to improve analysis speed while reducing the subjectivity inherent in the manual technique. This study emphasizes the benefits of automation in the coffee business, such as increased efficiency, lower costs, and standardized classification methods.

2.2. Infrared spectroscopy

Near and mid-infrared spectral studies have evolved as a valid and promising analytical method for objectively assessing coffee quality features over the last two decades [19–23]. These studies demonstrate that near and mid-infrared techniques have enormous potential for rapidly obtaining information about the chemical composition and related aspects of coffee.

In order to classify Arabica and Robusta coffee species, Calvinia et al. [20] compared two sparse classification approaches, sparse variants of Principal Component Analysis (sPCA) with KNN, and sparse versions of Partial Least Squares with Discriminant Analysis (sPLS-DA). Green coffee samples were analyzed using near-infrared hyperspectral imaging. The average spectra from each hyperspectral image were used to build training and testing sets. The reported results show that the sparse methods produced similar results to the conventional methods. Sparse techniques, on the other hand, produced more interpretable and parsimonious models. Notably, both sparse classification algorithms converged on using the same spectral regions for variable selection. This convergence shows that those locations are chemically relevant in distinguishing between Arabica and Robusta coffee species.

The goal of Link's et al. [21] study was to create a neural network using Radial Basis Function (RBF) to classify the geographic and genotypic origin of Arabica coffee. The spectra were collected using Fourier Transform Infrared (FTIR) technology and subsequently processed using RBFs. The results demonstrated that the modified RBF successfully classified Arabica coffee samples. Geographically, the classification accuracy was 100%, however, In terms of genotypic classification, the classification accuracy was 94.44%.

Soft independent modeling of class analogies (SIMCA) model was built by Mutz et al. [22] employing a portable near-infrared (NIR) spectrometer and a dataset of 182 coffee samples. The goal was to distinguish between distinct coffee qualities. Specialty coffees from two species, C. arabica and C. canephora, Arabica coffees from specific geographical indication regions (GI), and commodity coffee blends were among the samples. The proposed SIMCA model had a good classification accuracy for individual Arabica coffees from GI areas, ranging from 76% to 90%. This suggests that the model was able to distinguish various Arabica coffees depending on their geographical origins. Furthermore, for specialized Arabica coffees and Conilon coffees, the classification accuracy was 98% and 95% respectively.

Okubo and Kurata [23] used classification analysis and NIR to determine the production area of green coffee beans. SIMCA was used for the classification of Green coffee bean samples, which were collected from seven different places. The study achieved an overall correct classification rate of over 73% for different types of green coffee beans using NIR.

2.3. Electronic nose

An electronic nose is a type of electronic sensing equipment that detects scents or flavors by mimicking human senses with sensor arrays and pattern recognition systems. This technology has lately been employed for cuffee classification.[24–26].

Makimori and Bona [24] utilized An electronic nose (E-nose) outfitted with seven metal-oxide-semiconductor (MOS) sensors to examine 53 samples of six distinct commercial instant coffees produced by the same industry. For sample classification, they used chemometric methods such as common dimension analysis (ComDim) and linear discriminant analysis (LDA). ComDim, an unsupervised multiblock analysis method, was used to minimize the dimensions of the E-nose sensor data. The first derivative of the transitory signal was used to construct each sensor's block. In the E-nose data, four common dimensions (CDs) were identified, accounting for 99.86% of the total variation. Salience tables revealed links between sensors S1, S3, S5, S6, and S8, inside CD1, but sensors S7 and S9 had a higher influence on CD2. The scores from the first four CDs were used as input for building LDA classifiers. All generated models obtained 100% sensitivity and specificity using leave-one-out cross-validation. This suggests that the models accurately classified the coffee samples studied.

Bona et al. [25] developed an ANN to classify instant coffee using an E-nose fragrance profile. A hybrid algorithm with several components was developed. To begin, the dataset was subjected to a bootstrap resample process. The network parameters were then fine-tuned using a factorial design and sequential simplex optimization. Backpropagation was used to train MLP for coffee classification. Finally, knowledge was extracted from the trained ANN using a causal index approach. The proposed approach performed well by correctly identifying 100% of the coffee samples tested.

Tang et al. [26] constructed a classification system comprised of an environmental control system, an E-nose, and a data signal readout system. The system's goal was to distinguish between different degrees of mold on coffee beans by assessing the scent of the beans. A standard operating process was designed to collect gas samples from coffee beans in a controlled environment. The E-nose was utilized to capture the changes in the signals after the target gas was introduced. Dimensionality reduction techniques such as PCA and LDA were applied to the E-nose data in order to reduce data dimensionality and eliminate noise. The proposed system achieved a classification accuracy of 91.77%.

Other studies that addressed coffee classification issues, regardless of approach, include but are not limited to [27–34].

According to the literature, most classification methods proposed to classify coffee samples are based on shape, size, color, infrared spectroscopy, and/or aroma. To the best of our knowledge, none of them addressed the classification of Arabic coffee samples using their chemical components; therefore, this research seeks to fill that gap.

3. Materials, Data, and Methods

3.1. Materials

Nine kg of fermented and dried beans of a local Arabic coffee cultivar known as Kholani were acquired by [3] from supermarkets in Tabuk City, Saudi Arabia. Drum roasters, which are typically used for coffee roasting, were utilized to roast the coffee beans. The roasting temperature was established at $180 \pm 10°C$, which was held for around 6.0 ± 1.0 minutes to obtain the light roast, 8.0 ± 1.0 minutes to obtain the medium roast, and 10.0 ± 1.0 minutes to obtain the dark roast. After roasting, the coffee beans were ground and then stored in an airtight jar in the refrigerator until further analysis. Analytical-grade chemicals and indicators were used, which were obtained from Sigma-Aldrich, USA. The results of these laboratory analyses were reported by [3] and are shown in Tables 1, 2, and 3.

3.2. Data

Typically, having a sufficiently large dataset increases the performance and generalization of machine learning models. It enables them to capture a broader range of patterns, reduce overfitting, and better handle real-world conditions [35–39]. Therefore, in order to achieve the primary goal of this study, which is to classify Arabic coffee based on its chemical components, we need a sufficiently large dataset of coffee samples, which is a luxury we do not have because each chemical feature requires laboratory testing, which comes at a very high cost that we cannot afford at the moment.

As a result, we opt for simulation data, which is generated from real data to train our machine learning models. we utilize the data obtained by [3] (shown in Tables 1, 2 and 3), to generate new 1000 samples for each type of Arabic coffee (light, medium, and dark coffee) using the Monte Carlo method because we only have one mean value for each feature ± its standard deviation.

Using simulation data for machine learning is not new because it serves numerous goals such as data augmentation, imbalanced data, privacy and security, rare events, novel scenarios, and a lack of labeled data [40–43]. The latter is our case because actual laboratory testing for a big number of coffee samples is expensive to obtain.

Our simulated data is based on real-world chemical tests on three distinct coffee classes. It is critical to differentiate this simulation approach from oversampling approaches used in machine learning to handle class imbalance. Oversampling by creating new instances can produce unsatisfactory

results if the newly formed samples are wrongly thought to be part of the minority class based only on their closeness to existing minority examples. When dealing with imbalanced datasets, it is important to avoid conflating the usage of simulated data with oversampling class-imbalanced data[44,45].

The central limit theorem implies that the mean sampling distribution will approach a normal distribution, regardless of the population's original distribution [46]. It is easy to verify that the assumption of the central limit theorem holds for our specific dataset simply by looking at Figure 1. However, without obtaining real tests for a large number of coffee samples, it is difficult to ensure that the simulated data accurately reflects the underlying distribution of the coffee samples and their chemical features in the real population.

Table 4 shows the statistical characteristics of the simulated data and Figure 1 depicts the distribution of each simulated feature.
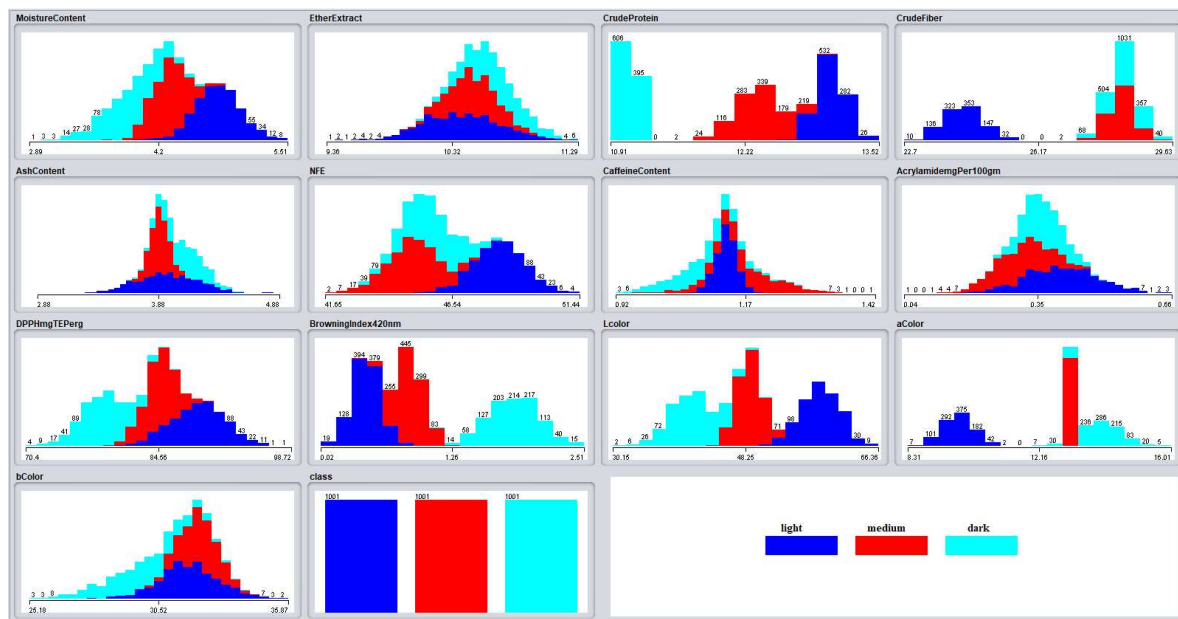
Figure 1. The distribution of each simulated feature. The visualization of the data is generated using Weka 3 [47].

Table 4. The statistical characteristics of the simulated data, 1000 samples for each class, in addition to the three actual samples.

| Class	Light coffee			Medium coffee			Dark coffee		
Features	Min	Max	Mean	Min	Max	Mean	Min	Max	Mean
MoistureContent	3.98	5.51	4.80	3.67	4.83	4.30	2.89	4.93	3.90
EtherExtract	9.36	11.29	10.39	9.90	11.04	10.48	10.04	11.26	10.65
CrudeProtein	12.58	13.52	13.05	11.64	13.13	12.36	10.91	11.28	11.10
CrudeFiber	22.70	25.64	24.23	27.30	29.55	28.31	27.02	29.63	28.39
AshContent	2.88	4.88	3.95	3.63	4.16	3.89	3.51	4.59	4.10
NFE	45.29	51.44	48.36	41.65	48.17	44.98	42.48	49.27	45.74
CaffeineContent	1.07	1.19	1.13	0.95	1.42	1.17	0.92	1.27	1.08
Acrylamide	0.12	0.66	0.40	0.04	0.52	0.31	0.20	0.50	0.36
DPPH	78.22	98.72	88.80	79.54	90.07	84.68	70.40	86.05	78.64
Browning Index	0.02	0.90	0.44	0.35	1.23	0.86	1.04	2.51	1.84
Lcolor	48.67	66.36	58.44	43.52	53.87	48.81	30.15	52.24	41.10
aColor	8.31	11.25	9.78	12.79	13.24	13.04	12.08	16.01	13.92
bColor	27.56	35.87	31.75	30.00	34.46	32.22	25.18	35.43	29.76

The simulated data, as shown in Table 4, has means that nearly match the actual means presented in Tables 1, 2, and 3. This observation shows that the Monte Carlo method was effective in generating samples centered on the true means, with the goal of approximating the features of the real population.

3.3. Feature importance

To pick a subset of relevant features from a number of available features, a feature selection technique is used. The major goal of this process is to improve the machine learning model's evaluation metrics by omitting irrelevant redundant data. Two feature selection techniques were used in our experiment to analyze and determine the most influential features in classifying coffee. These methods include the Correlation coefficient (shown in Figure 2) and feature importance (shown in Figure 3). The feature importance scores are calculated using the Mean Decrease Gini index with a random forest classifier.
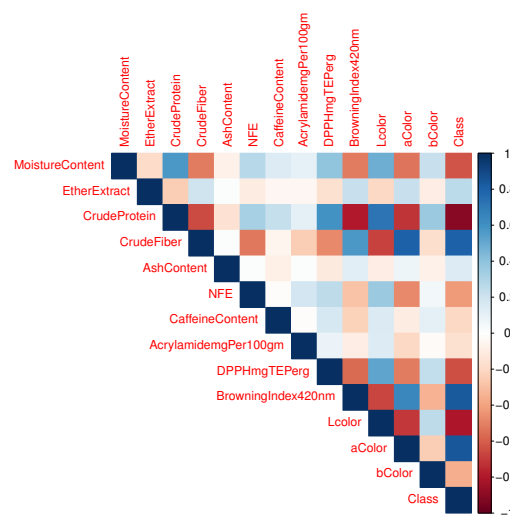
Figure 2. The results of the Pearson correlation analysis between all features and the Class variable, which represents the three coffee types.
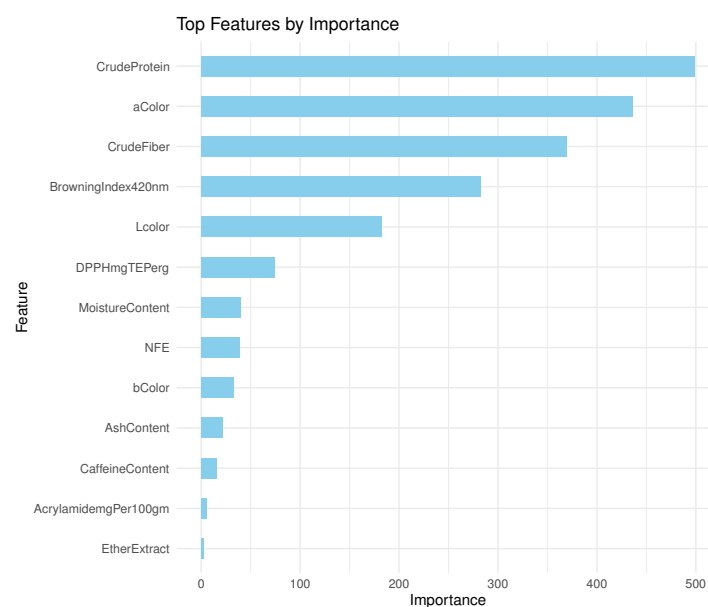
Figure 3. The importance scores of each feature were determined using a random forest algorithm to classify the three different coffee types.

According to Figure 2, the variables with the highest link to types of coffee (Class) include Crude Protein (-0.91), Crude Fiber (0.80), and color-related parameters like a* (0.85), Browning Index (0.84), and L* (-0.81). Similarly, Figure 3 shows that Crude Protein, a*, Crude Fiber, Browning Index, and L* are the most important features for classifying coffee types based on the simulated data.

Given that, each coffee type has distinct color properties when compared to the other two, a high link between color information and the three coffee types is reasonable. Furthermore, the high correlation of Crude Protein and Crude Fiber percentages with coffee classes can be related to their discriminative strength, as shown by the separability of their values in the distributions presented in the top-right corner of Figure 1.

3.4. Methods: Machine learning classifiers

Several widely used classifiers were studied to determine the best option for the proposed Arabic coffee classification system. The analysis focused on the Weka3 implementations of Random Forests (RF), Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Naive Bayes (NB), and the decision tree algorithm C4.5.

In this study, we used the default parameters for each classifier, for example, KNN (K=1, distance function = Euclidean), SVM (method=libsvm library, kernel=radial basis function), and RF (number of trees=100, number of features=$int(\log_2{(\#features + 1)})$).

It is worth noting that these default hyperparameters may not be suitable for every dataset or classification task. To discover the ideal configuration for a classification task, it is sometimes recommended to tweak the hyperparameters depending on the specific properties of the data source [48,49]. However, such an investigation is beyond the scope of this study.

4. Results

The chemical tests conducted in the laboratory for the three types of coffee by [3] can be classified into three categories of information as follows:

- Arabic Coffee color information: This category includes measurements such as the Browning index, which is determined using a UV-Vis spectrophotometer at a wavelength of 420 nm, in addition, it includes color information based on the International Commission of Illumination (CIE) color system, which involves the L*, a*, and b* channels.
- Arabic Coffee Antioxidant information: This category comprises the caffeine percentage, acrylamide concentration (measured in mg per 100g), and DPPH.
- Arabic Coffee chemical compositions: This category includes the percentage of moisture content, percentage of ether extract, percentage of crude protein, percentage of crude fiber, percentage of ash content, and percentage of NFE.

A series of experiments focusing on the aforementioned types of information were carried out to evaluate the proposed Arabic coffee classification system based on its chemical properties. It's worth noting that all of the tests used machine learning methods, and a 10-fold cross-validation approach was applied with each method's default parameters. This approach ensures that the evaluation was consistent and unbiased across all experiments.

Table 4 shows the classification results of Arabic coffee based on its color information: L*, a*, and b*. We did not add the browning index to this information category because it can be assessed from the L*, a*, and b* channels according to [50].

Table 5. Classification results of Arabic coffee based on its color information: L*, a* and b*.

Classifier	Precision	Recall	F-Measure
RF	1.000	1.000	1.000
NB	0.998	0.998	0.998
SVM	0.993	0.993	0.993
KNN	0.994	0.994	0.994
J48	0.993	0.993	0.993

According to the data presented in Table , it is evident that the color information alone was sufficient to achieve perfect classification of Arabic coffee into its three types: light, medium, and dark. This observation can be attributed to the fact that the CIE colors of these three classes are easily differentiated. Furthermore, Figure 2 depicts the high correlation between color information and coffee classes, underlining the association between color and coffee classification. Additionally, Figure 3 demonstrates the significant value of color information in discriminating across coffee types.

Table 6 presents the classification results of Arabic coffee based on its antioxidant information: caffeine percentage, acrylamide concentration, and DPPH.

Table 6. Classification results of Arabic coffee based on its antioxidant information: caffeine percentage, acrylamide concentration, and DPPH.

Classifier	Precision	Recall	F-Measure
RF	0.900	0.899	0.900
NB	0.912	0.912	0.912
SVM	0.866	0.856	0.858
KNN	0.868	0.868	0.868
J48	0.893	0.892	0.892

Results presented in Table 6 show that when both the color and chemical composition information were eliminated; leaving only the Antioxidant information, the classification results significantly decreased. The performance of the classifiers varied, with the best performer being NB, followed by RF. This means that depending exclusively on antioxidant information is insufficient for accurate Arabic coffee classification. The correlation analysis illustrated in Figures 2 and 3 supports this observation. The correlation values and importance rankings show that variables like caffeine percentage and acrylamide concentration are less important in the classification process and have weaker correlations with the dependent variable (coffee class).

A closer look at Figure 1 reveals that some of the curves depicting these features for the three types of Arabic coffee have significant overlap. This overlap implies that the values of these features for various types of coffee are not unique enough to create obvious borders or distinctions between the classes. As a result, depending entirely on Antioxidant features for classification would almost certainly result in misclassifications or poorer classification accuracy. The observed difficulties in effectively identifying Arabic coffee based on antioxidant information can be attributable mostly to the near proximity of these variables' actual values. As shown in Table 1, the caffeine percentages of light, medium, and dark coffee were reported as 1.13±0.02, 1.17±0.07, and 1.08±0.06, respectively.

Table 7 presents the classification results of Arabic coffee based on its chemical compositions information: Moisture content %, Ether extract %, Crude protein %, Crude fiber %, Ash content %, and NFE) %.

According to the data in Table 7, all classifiers' classification results are either perfect, with the exception of J48, which earned a classification accuracy of 0.999. However, this classifier achieved perfect training accuracy by obtaining the following rules:

1. if CrudeProtein <= 11.283023 then class= 3 (1001 out of 1001 examples).

2. if CrudeProtein > 11.283023 and CrudeFiber <= 25.642686 then class= 1 (1001 out of 1001 examples).
3. if CrudeProtein > 11.283023 and CrudeFiber > 25.642686 then class= 2 (1001 out of 1001 examples).

Table 7. Classification results of Arabic coffee based on its chemical compositions information: Moisture content %, Ether extract %, Crude protein %, Crude fiber %, Ash content %, and NFE) %.

Classifier	Precision	Recall	F-Measure
RF	1.000	1.000	1.000
NB	1.000	1.000	1.000
SVM	1.000	1.000	1.000
KNN	1.000	1.000	1.000
J48	0.999	0.999	0.999

This level of agreement and excellent accuracy across classifier families shows that the training data utilized in the studies was distinct and well-separated. Based on these outcomes, it is clear that chemical composition data alone was sufficient to achieve faultless classification of Arabic coffee into three types: light, medium, and dark. This great precision is most likely due to the distinctive characteristics of the chemical composition information for these classes.

Figure 2 also shows a substantial link between chemical composition information and coffee classes. The percentages of crude protein and crude fiber, in particular, have strong correlation values of -0.91 and 0.80, respectively. This demonstrates the connection and discriminatory power of these variables in classifying coffee types.

Furthermore, Figure 3 shows that when compared to other features, the percentage of crude protein ranks first and the percentage of crude fiber is ranked third. This highlights the importance and discriminative capacity of this type of information in the classification of coffee types, particularly, crude protein and crude fiber, which are extremely separable as shown in Figure 4 and can achieve perfect classification alone even without the need for the other variables.
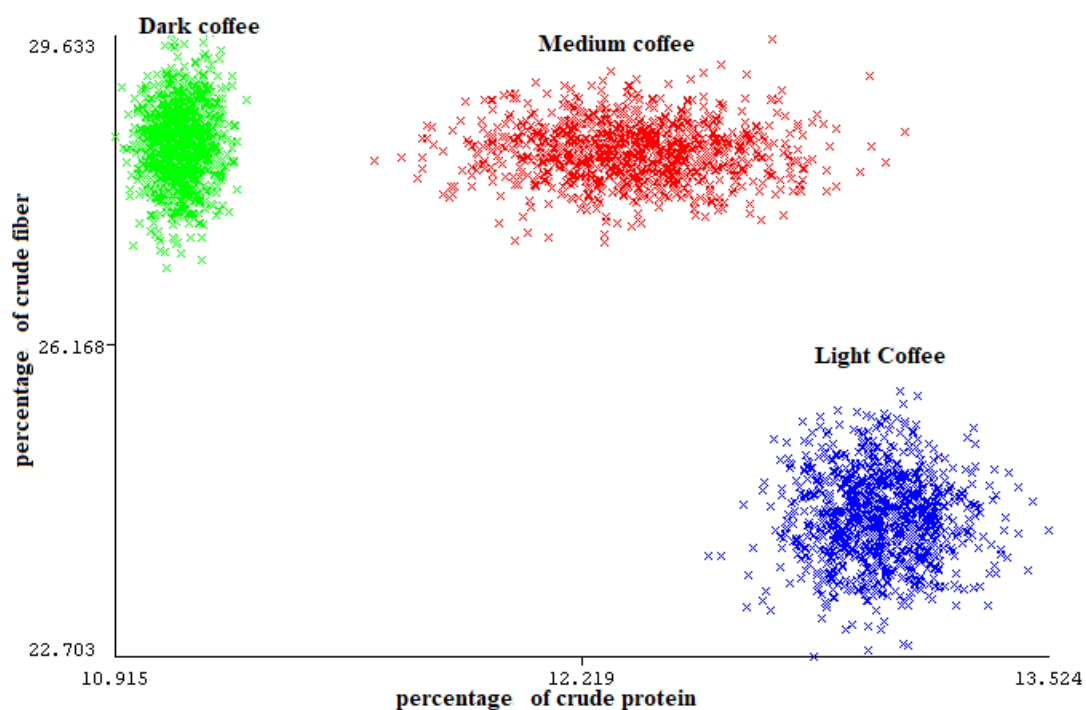
Figure 4. The crude fiber% as a function of crude protein for all simulated data points. The visualization of the data is generated using Weka 3 [47].

Based on the preceding results, it is evident that the simulated data used for classification was clean and well-suited for various types of classifiers, owing to the features' strong discriminatory power. This was especially true for the categories of color information and chemical composition information.

The standard deviation of the normal distribution for each feature was doubled (2STD) to bring more unpredictability and variability into the simulated data, resulting in a dataset with higher dispersion. By providing more variability in the feature values, this modification attempts to increase diversity and make the classification task more challenging. The classification results of the new 2STD simulated data are shown in Tables 8, 9, 10, and 11.

Table 8. Classification results of Arabic coffee based on its color information: L*, a*, and b*, obtained from 2STD simulated data.

Classifier	Precision	Recall	F-Measure
RF	0.968	0.968	0.968
NB	0.972	0.972	0.972
SVM	0.953	0.952	0.952
KNN	0.957	0.956	0.956
J48	0.964	0.964	0.964

Table 9. The confusion matrices of four classifiers of 2STD simulated color data.

	Light	Mediun	Dark	Light	Mediun	Dark
Light	1001	0	0	992	0	9
Mediun	0	980	21	0	968	33
Dark	12	53	936	9	94	898
		KNN			RF	
Light	991	0	10	993	1	7
Mediun	0	959	42	0	969	32
Dark	8	71	922	10	47	944

Table 10. Classification results of Arabic coffee based on its 2STD simulated antioxidant information: caffeine percentage, acrylamide concentration, and DPPH.

Classifier	Precision	Recall	F-Measure
RF	0.701	0.700	0.700
NB	0.740	0.740	0.740
SVM	0.678	0.660	0.662
KNN	0.640	0.641	0.640
J48	0.749	0.703	0.725

Table 11. Classification results of Arabic coffee based on its 2STD simulated chemical compositions information: Moisture content %, Ether extract %, Crude protein %, Crude fiber %, Ash content %, and NFE) %.

| Classifier	Precision	Recall	F-Measure
RF	0.991	0.991	0.991
NB	0.993	0.993	0.993
SVM	0.967	0.998	0.982
KNN	0.983	0.983	0.983
J48	0.985	0.985	0.985

According to the results obtained from the newly generated data with doubled standard deviation, the classification performance of Arabic coffee based on its 2STD simulated chemical composition information reduced by 1% to 3%. Similarly, depending on the classifier, the classification results based on its 2STD simulated color information are reduced by 2% to 5%. This decrease in performance might be due to the growing complexity of the data, which causes classifiers to differ in their performance. It is worth mentioning that clean data often yields superior classification accuracy regardless of the classifier used.

In the case of the 2STD simulated antioxidant information, The classification results showed a drop of 16% to 20%, depending on the classifier utilized. The KNN classifier, in particular, performed poorly in this scenario due to its reliance on similarity across samples for classification, which becomes less accurate when dealing with complicated data.

This drop in performance can be traced primarily to the nature of the original data shown in Table 1. The antioxidant characteristics in the table demonstrate no significant difference between the three coffee classes (light, medium, and dark). As a result, increasing the standard deviation to generate the 2STD simulated data reduced the distinctiveness and separability of the antioxidant characteristics, resulting in an overall decrease in classification accuracy.

When analyzing the antioxidant classification results, it is critical to recognize these constraints as well as the impact of data characteristics. In such circumstances, additional characteristics or alternative methodologies may be necessary to improve classification performance.

The observed decline in classification performance, as well as the variation in performance among classifiers, emphasizes the need for classifier selection when dealing with complicated data. Figure 5 (2STD simulated data) and Figure 4 (1STD data) show how the data became increasingly complex and overlapping. This intricacy is illustrated further in Figure 6 by the decision tree of J48, where the rules have become more intricate. However, it is worth mentioning that the increase in data complexity had no significant effect on classification accuracy for color information and chemical composition.
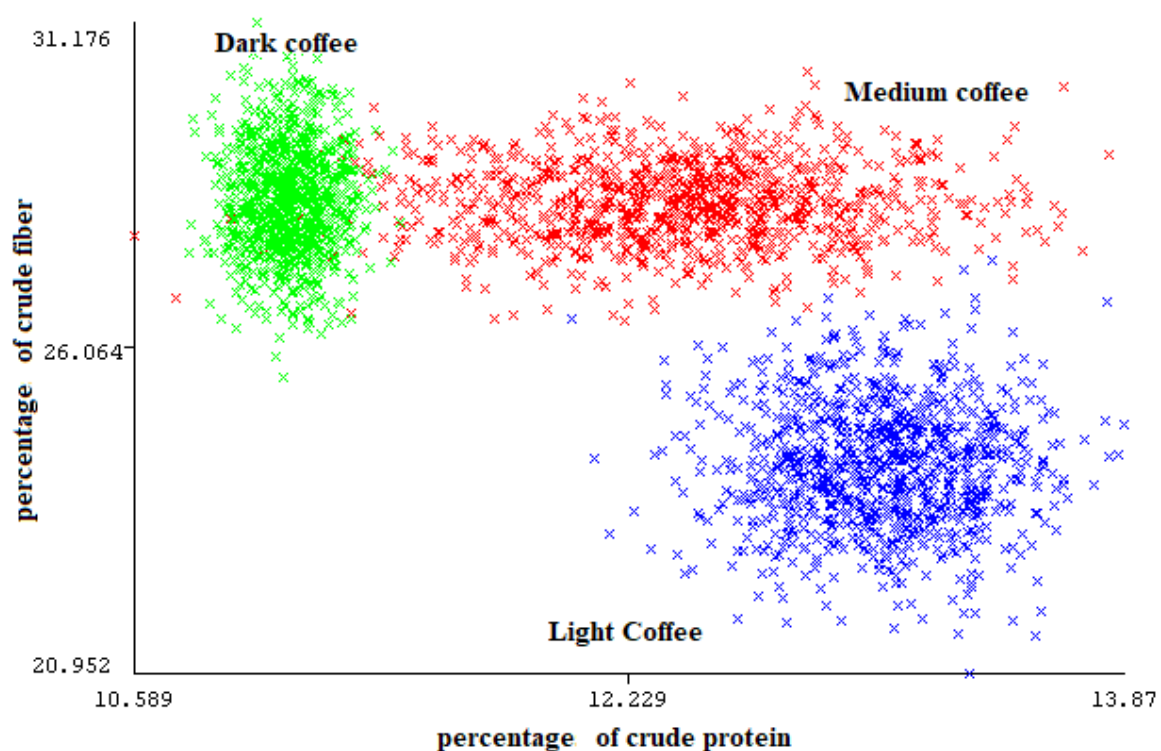
Figure 5. The crude fiber% as a function of crude protein for all 2STD simulated data points. The visualization of the data is generated using Weka 3 [47].
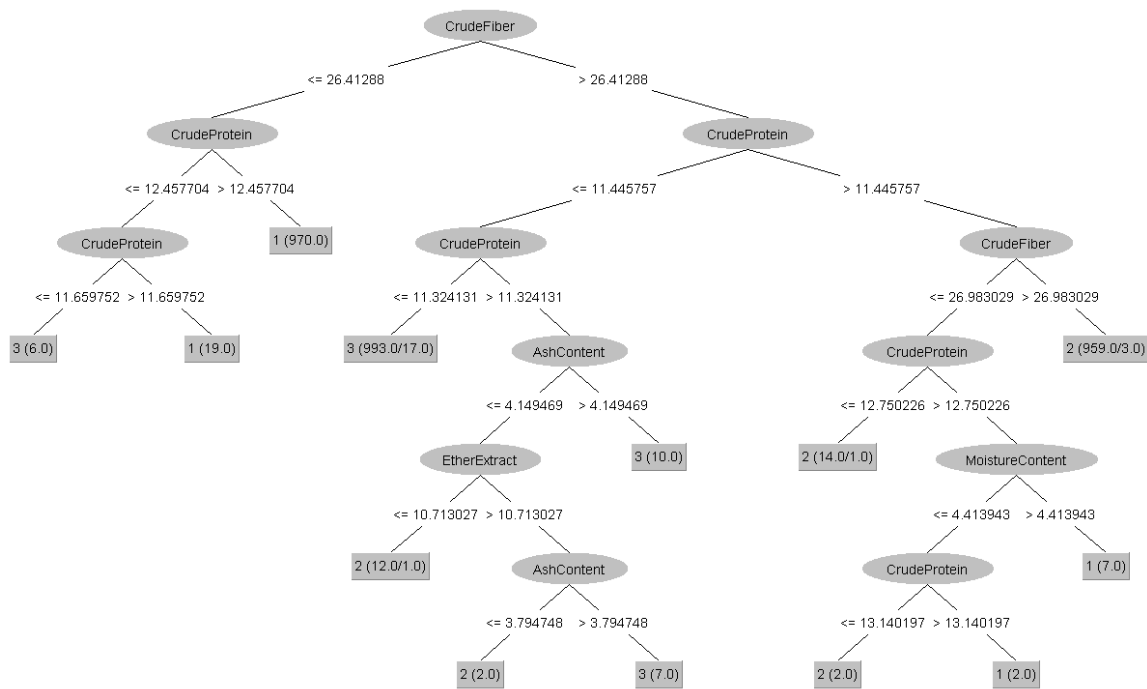
Figure 6. J48 decision tree for classifying coffee types based on 2STD simulated chemical compositions information. The visualization of the data is generated using Weka 3 [47].

5. Conclusions

In this paper, we evaluated the applicability of classifying Arabic coffee into three major varieties, light, medium, and dark, using simulated data based on Iman's actual data of color information, antioxidant laboratory tests, and chemical composition tests. We created two types of simulated data using the Monte Carlo approach: the first used the standard deviation of each of the actual measures, and the second used a double standard deviation of each of the actual measures. We classified both of the simulated data using a number of commonly used classifiers.

Based on our large set of experiments, several key findings have emerged:

- The color information alone proved to be quite successful in accurately classifying Arabic coffee into three types: light, medium, and dark. The different CIE color values, as well as the excellent correlation between color information and coffee classes, were critical in achieving faultless classification outcomes.
- When the antioxidant information was considered alone, it became clear that this type of information was insufficient for correct coffee classification. The results showed that relying simply on antioxidant information reduced classification performance significantly, and different classifiers performed differently. The KNN classifier in particular struggled with the additional data complexity (2STD simulated data).
- In identifying Arabic coffee types, the chemical composition information revealed significant discriminatory strength. This information alone produced flawless classification results, demonstrating its significance and separability among the coffee groups. The substantial correlation between certain chemical composition features, such as crude protein and crude fiber, underscored their importance in the classification process even more.
- The classification performance of the antioxidant information and the choice of classifier grew more relevant as the data complexity rose. The overlapping curves and increasingly elaborate decision tree rules demonstrated the difficulties encountered when dealing with complex data and the importance of proper classifier selection.

Indeed, based on the data stated, color information alone has proved the ability to reliably identify the type of coffee, showing its practical application potential. Using this knowledge, it is possible to

create a smartphone application that analyzes the color of a coffee sample using image recognition technology such as [51–54]. Such an application could allow end users, such as consumers, to utilize their smartphone's camera to snap an image of their coffee. The application would next transform the image's RGB pixel values to the CIE color space, extracting the color information required for classification. After recognizing the coffee type, the application might offer the consumer with useful information on the chemical components of their coffee. Details such as caffeine content, antioxidant levels, moisture content, and other pertinent chemical compositions could be included. Our future mobile application would provide users with a handy and user-friendly tool for understanding the qualities of their coffee and making informed decisions depending on their tastes by delivering such information. Furthermore, it has the potential to increase transparency in the coffee sector, allowing customers to make more educated purchasing decisions.

There are various limitations that could be addressed in future coffee classification studies:

- The paper largely concentrated on color, antioxidant, and chemical composition information. Exploring additional features that could improve classification accuracy, such as scent profiles, geographical origin, or processing processes, would be advantageous. Incorporating a larger variety of characteristics may result in a more comprehensive understanding of the factors determining coffee classification.
- According to the findings, the distinctness and separability of certain variables, particularly antioxidant information, were limited, resulting in lower classification accuracy. Introducing more diverse and representative datasets with a broader range of actual (not simulated) antioxidant feature values could help alleviate this restriction and provide a more genuine portrayal of coffee variations.
- The study focused on several off-the-shelf classifiers, although different algorithms or ensemble approaches could produce better results [55–58]. Exploring and comparing different classifier alternatives could provide useful insights into the most successful ways for coffee classification.
- This paper focused on Arabic coffee varieties and classification. It would be beneficial to replicate the study using different coffee types from various places throughout the world to improve the generalizability of the findings. This would assist test the categorization algorithms' efficiency across different coffee varieties and broaden the research's practicality.
-

Our future studies will focus on these limitations in order to progress the field of coffee classification and contribute to more accurate and comprehensive classification models. Furthermore, additional research and development would be required to ensure the reliability and practicality of our future mobile application.

References

1. Al-Mssallem, M.Q.; Brown, J.E. Arabic coffee increases the glycemic index but not insulinemic index of dates. *Saudi Med J* **2013**, *34*, 923–928.

2. Butt, M.S.; Sultan, M.T. Coffee and its consumption: benefits and risks. *Critical reviews in food science and nutrition* **2011**, *51*, 363–373.

3. Alamri, E.; Rozan, M.; Bayomy, H. A study of chemical Composition, Antioxidants, and volatile compounds in roasted Arabic coffee. *Saudi Journal of Biological Sciences* **2022**, *29*, 3133–3139.

4. Ciaramelli, C.; Palmioli, A.; Airoldi, C. Coffee variety, origin and extraction procedure: Implications for coffee beneficial effects on human health. *Food chemistry* **2019**, *278*, 47–55.

5. Messina, G.; Zannella, C.; Monda, V.; Dato, A.; Liccardo, D.; De Blasio, S.; Valenzano, A.; Moscatelli, F.; Messina, A.; Cibelli, G.; others. The beneficial effects of coffee in human nutrition. *Biology and Medicine* **2015**, *7*, 1.

6. Poole, R.; Kennedy, O.J.; Roderick, P.; Fallowfield, J.A.; Hayes, P.C.; Parkes, J. Coffee consumption and health: umbrella review of meta-analyses of multiple health outcomes. *bmj* **2017**, *359*.

7. Borota, D.; Murray, E.; Keceli, G.; Chang, A.; Watabe, J.M.; Ly, M.; Toscano, J.P.; Yassa, M.A. Post-study caffeine administration enhances memory consolidation in humans. *Nature neuroscience* **2014**, *17*, 201–203.

8. Olson, C.A.; Thornton, J.A.; Adam, G.E.; Lieberman, H.R. Effects of 2 adenosine antagonists, quercetin and caffeine, on vigilance and mood. *Journal of clinical psychopharmacology* **2010**, *30*, 573–578.

9. Nehlig, A. Is caffeine a cognitive enhancer? *Journal of Alzheimer's Disease* **2010**, *20*, S85–S94.

10. Cai, L.; Ma, D.; Zhang, Y.; Liu, Z.; Wang, P. The effect of coffee consumption on serum lipids: a meta-analysis of randomized controlled trials. *European journal of clinical nutrition* **2012**, *66*, 872–877.

11. Zhou, A.; Hyppönen, E. Habitual coffee intake and plasma lipid profile: Evidence from UK Biobank. *Clinical nutrition* **2021**, *40*, 4404–4413.

12. AlQuaiz, A.M.; Kazi, A.; Tayel, S.; Shaikh, S.A.; Al-Sharif, A.; Othman, S.; Habib, F.; Fouda, M.; Sulaimani, R. Prevalence and factors associated with low bone mineral density in Saudi women: a community based survey. *BMC musculoskeletal disorders* **2014**, *15*, 1–10.

13. Somporn, C.; Kamtuo, A.; Theerakulpisut, P.; Siriamornpun, S. Effects of roasting degree on radical scavenging activity, phenolics and volatile compounds of Arabica coffee beans (Coffea arabica L. cv. Catimor). *International Journal of Food Science & Technology* **2011**, *46*, 2287–2296.

14. Wang, H.Y.; Qian, H.; Yao, W.R. Melanoidins produced by the Maillard reaction: Structure and biological activity. *Food chemistry* **2011**, *128*, 573–584.

15. de Oliveira, E.M.; Leme, D.S.; Barbosa, B.H.G.; Rodarte, M.P.; Pereira, R.G.F.A. A computer vision system for coffee beans classification based on computational intelligence techniques. *Journal of Food Engineering* **2016**, *171*, 22–27.

16. Arboleda, E.R.; Fajardo, A.C.; Medina, R.P. 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). IEEE, 2018, pp. 1–5.

17. Arboleda, E.R. Comparing performances of data mining algorithms for classification of green coffee beans. *Int. J. Eng. Adv. Technol* **2019**, *8*, 1563–1567.

18. Pizzaia, J.P.L.; Salcides, I.R.; de Almeida, G.M.; Contarato, R.; de Almeida, R. Arabica coffee samples classification using a Multilayer Perceptron neural network. 2018 13th IEEE International Conference on Industry Applications (INDUSCON). IEEE, 2018, pp. 80–84.

19. Barbin, D.F.; Felicio, A.L.d.S.M.; Sun, D.W.; Nixdorf, S.L.; Hirooka, E.Y. Application of infrared spectral techniques on quality and compositional attributes of coffee: An overview. *Food Research International* **2014**, *61*, 23–32.

20. Calvini, R.; Ulrici, A.; Amigo, J.M. Practical comparison of sparse methods for classification of Arabica and Robusta coffee species using near infrared hyperspectral imaging. *Chemometrics and Intelligent Laboratory Systems* **2015**, *146*, 503–511.

21. Link, J.V.; Lemes, A.L.G.; Marquetti, I.; dos Santos Scholz, M.B.; Bona, E. Geographical and genotypic classification of arabica coffee using Fourier transform infrared spectroscopy and radial-basis function networks. *Chemometrics and Intelligent Laboratory Systems* **2014**, *135*, 150–156.

22. Mutz, Y.S.; do Rosario, D.; Galvan, D.; Schwan, R.F.; Bernardes, P.C.; Conte-Junior, C.A. Feasibility of NIR spectroscopy coupled with chemometrics for classification of Brazilian specialty coffee. *Food Control* **2023**, *149*, 109696.

23. Okubo, N.; Kurata, Y. Nondestructive classification analysis of green coffee beans by using near-infrared spectroscopy. *Foods* **2019**, *8*, 82.

24. Makimori, G.Y.F.; Bona, E. Commercial instant coffee classification using an electronic nose in tandem with the ComDim-LDA approach. *Food Analytical Methods* **2019**, *12*, 1067–1076.

25. Bona, E.; da Silva, R.S.d.S.F.; Borsato, D.; Bassoli, D.G. Optimized neural network for instant coffee classification through an electronic nose. *International Journal of Food Engineering* **2011**, *7*.

26. Tang, C.L.; Chou, T.I.; Yang, S.R.; Lin, Y.J.; Ye, Z.K.; Chiu, S.W.; Lee, S.W.; Tang, K.T. Development of a Nondestructive Moldy Coffee Beans Detection System Based on Electronic Nose. *IEEE Sensors Letters* **2023**, *7*, 1–4.

27. Luo, S.; Yan, C.; Chen, D. Preliminary study on coffee type identification and coffee mixture analysis by light emitting diode induced fluorescence spectroscopy. *Food Control* **2022**, *138*, 109044.

28. Hu, Q.; Sellers, C.; Kwon, J.S.I.; Wu, H.J. Integration of surface-enhanced Raman spectroscopy (SERS) and machine learning tools for coffee beverage classification. *Digital Chemical Engineering* **2022**, *3*, 100020.

29. Belchior, V.; Botelho, B.G.; Franca, A.S. Comparison of spectroscopy-based methods and chemometrics to confirm classification of specialty coffees. *Foods* **2022**, *11*, 1655.

30. Tamayo-Monsalve, M.A.; Mercado-Ruiz, E.; Villa-Pulgarin, J.P.; Bravo-Ortíz, M.A.; Arteaga-Arteaga, H.B.; Mora-Rubio, A.; Alzate-Grisales, J.A.; Arias-Garzon, D.; Romero-Cano, V.; Orozco-Arias, S.; others. Coffee maturity classification using convolutional neural networks and transfer learning. *IEEE Access* **2022**, *10*, 42971–42982.

31. Manuel, M.N.B.; da Silva, A.C.; Lopes, G.S.; Ribeiro, L.P.D. One-class classification of special agroforestry Brazilian coffee using NIR spectrometry and chemometric tools. *Food Chemistry* **2022**, *366*, 130480.

32. Gope, H.L.; Fukai, H. Peaberry and normal coffee bean classification using CNN, SVM, and KNN: Their implementation in and the limitations of Raspberry Pi 3. *AIMS Agriculture and Food* **2022**, *7*, 149–167.

33. Adiwijaya, N.O.; Romadhon, H.I.; Putra, J.A.; Kuswanto, D.P. The quality of coffee bean classification system based on color by using k-nearest neighbor method. Journal of Physics: Conference Series. IOP Publishing, 2022, Vol. 2157, p. 012034.

34. Pahlawan, M.F.R.; Masithoh, R.E. Vis-NIR Spectroscopy and PLS-Da model for classification of Arabica and robusta roasted coffee bean. Advances in Science and Technology. Trans Tech Publ, 2022, Vol. 115, pp. 45–52.

35. Figueroa, R.L.; Zeng-Treitler, Q.; Kandula, S.; Ngo, L.H. Predicting sample size required for classification performance. *BMC medical informatics and decision making* **2012**, *12*, 1–10.

36. Tarawneh, A.S.; Hassanat, A.B.; Chetverikov, D.; Lendak, I.; Verma, C. Invoice classification using deep features and machine learning techniques. 2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT). IEEE, 2019, pp. 855–859.

37. Hassanat, A.B.; Prasath, V.S.; Al-kasassbeh, M.; Tarawneh, A.S.; Al-shamailh, A.J. Magnetic energy-based feature extraction for low-quality fingerprint images. *Signal, Image and Video Processing* **2018**, *12*, 1471–1478.

38. Hassanat, A.B. On identifying terrorists using their victory signs. *Data Science Journal* **2018**, *17*.

39. Hassanat, A.B.; Btoush, E.; Abbadi, M.A.; Al-Mahadeen, B.M.; Al-Awadi, M.; Mseidein, K.I.; Almseden, A.M.; Tarawneh, A.S.; Alhasanat, M.B.; Prasath, V.S.; others. Victory sign biometrie for terrorists identification: Preliminary results. 2017 8th International Conference on Information and Communication Systems (ICICS). IEEE, 2017, pp. 182–187.

40. Bohn, B.; Garcke, J.; Iza-Teran, R.; Paprotny, A.; Peherstorfer, B.; Schepsmeier, U.; Thole, C.A. Analysis of car crash simulation data with nonlinear machine learning methods. *Procedia Computer Science* **2013**, *18*, 621–630.

41. Chen, H.; Kätelhön, E.; Compton, R.G. Machine Learning in Fundamental Electrochemistry: Recent Advances and Future Opportunities. *Current Opinion in Electrochemistry* **2023**, p. 101214.

42. Hassanat, A.B.; Mnasri, S.; Aseeri, M.A.; Alhazmi, K.; Cheikhrouhou, O.; Altarawneh, G.; Alrashidi, M.; Tarawneh, A.S.; Almohammadi, K.S.; Almoamari, H. A simulation model for forecasting covid-19 pandemic spread: Analytical results based on the current saudi covid-19 data. *Sustainability* **2021**, *13*, 4888.

43. Koutsoupakis, J.; Seventekidis, P.; Giagopoulos, D. Machine learning based condition monitoring for gear transmission systems using data generated by optimal multibody dynamics models. *Mechanical Systems and Signal Processing* **2023**, *190*, 110130.

44. Tarawneh, A.S.; Hassanat, A.B.; Altarawneh, G.A.; Almuhaimeed, A. Stop Oversampling for Class Imbalance Learning: A Review. *IEEE Access* **2022**.

45. Elreedy, D.; Atiya, A.F.; Kamalov, F. A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning. *Machine Learning* **2023**, pp. 1–21.

46. Hagtvedt, R.; Jones, G.T.; Jones, K. Pedagogical simulation of sampling distributions and the central limit theorem. *Teaching Statistics* **2007**, *29*, 94–97.

47. Witten, I.H.; Frank, E. *Data Mining: Practical Machine Learning Tools and Techniques*, 2 ed.; Morgan Kaufmann, 2005.

48. Gupta, S.C.; Goel, N. Predictive Modeling and Analytics for Diabetes using Hyperparameter tuned Machine Learning Techniques. *Procedia Computer Science* **2023**, *218*, 1257–1269.

49. Ahamad, G.N.; Shafiullah.; Fatima, H.; Imdadullah.; Zakariya, S.; Abbas, M.; Alqahtani, M.S.; Usman, M. Influence of Optimal Hyperparameters on the Performance of Machine Learning Algorithms for Predicting Heart Disease. *Processes* **2023**, *11*, 734.

50. Sikora, M.; Złotek, U.; Kordowska-Wiater, M.; Świeca, M. Spicy Herb Extracts as a Potential Improver of the Antioxidant Properties and Inhibitor of Enzymatic Browning and Endogenous Microbiota Growth in Stored Mung Bean Sprouts. *Antioxidants* **2021**, *10*, 425.

51. Hassanat, A.; Tarawneh, A.S. Fusion of Color and Statistic Features for Enhancing Content-Based Image Retrieval Systems. *Journal of Theoretical & Applied Information Technology* **2016**, *88*.

52. Hassanat, A.B.; Alkasassbeh, M.; Al-awadi, M.; Esra'a, A. Color-based object segmentation method using artificial neural network. *Simulation Modelling Practice and Theory* **2016**, *64*, 3–17.

53. Hassanat, A.B.; Jassim, S. Color-based lip localization method. Mobile Multimedia/Image Processing, Security, and Applications 2010. SPIE, 2010, Vol. 7708, pp. 313–324.

54. Hassanat, A.B.; Alkasassbeh, M.; Al-awadi, M.; Esra'a, A. Colour-based lips segmentation method using artificial neural networks. 2015 6th international conference on information and communication systems (ICICS). IEEE, 2015, pp. 188–193.

55. Kozubal, J.V.; Hassanat, A.; Tarawneh, A.S.; Wróblewski, R.J.; Anysz, H.; Valença, J.; Júlio, E. Automatic strength assessment of the virtually modelled concrete interfaces based on shadow-light images. *Construction and Building Materials* **2022**, *359*, 129296.

56. Hassanat, A.B.; Tarawneh, A.S.; Abed, S.S.; Altarawneh, G.A.; Alrashidi, M.; Alghamdi, M. Rdpvr: Random data partitioning with voting rule for machine learning from class-imbalanced datasets. *Electronics* **2022**, *11*, 228.

57. Hassanat, A.B.; Ali, H.N.; Tarawneh, A.S.; Alrashidi, M.; Alghamdi, M.; Altarawneh, G.A.; Abbadi, M.A. Magnetic Force Classifier: A Novel Method for Big Data Classification. *IEEE Access* **2022**, *10*, 12592–12606.

58. Hassanat, A.; Alkafaween, E.; Tarawneh, A.S.; Elmougy, S. Applications Review of Hassanat Distance Metric. 2022 International Conference on Emerging Trends in Computing and Engineering Applications (ETCEA). IEEE, 2022, pp. 1–6.