Hindawi
Scientific Programming
Volume 2023, Article ID 2556066, 21 pages
https://doi.org/10.1155/2023/2556066

Research Article

Identification of Dry Bean Varieties Based on Multiple Attributes Using CatBoost Machine Learning Algorithm

S. Krishnan,[1] **S. K. Aruna**,[2] **Karthick Kanagarathinam**,[3] **and Ellappan Venugopal**[4]

[1]*Department of EEE, Mahendra Engineering College (Autonomous), Namakkal, Tamil Nadu, India*
[2]*Department of Computer Science and Engineering, School of Engineering and Technology, CHRIST (Deemed to be University), Bangalore, Karnataka, India*
[3]*Department of Electrical and Electronics Engineering, GMR Institute of Technology, Rajam, Andhra Pradesh, India*
[4]*Department of Electronics and Communication Engineering, School of Electrical Engineering and Computing, Adama Science and Technology University, Adama, Ethiopia*

Correspondence should be addressed to Ellappan Venugopal; ellappan.venugopal@astu.edu.et

Received 12 December 2022; Revised 11 February 2023; Accepted 3 March 2023; Published 21 April 2023

Academic Editor: Sadiq Hussain

Dry beans are the most widely grown edible legume crop worldwide, with high genetic diversity. Crop production is strongly influenced by seed quality. So, seed classification is important for both marketing and production because it helps build sustainable farming systems. The major contribution of this research is to develop a multiclass classification model using machine learning (ML) algorithms to classify the seven varieties of dry beans. The balanced dataset was created using the random undersampling method to avoid classification bias of ML algorithms towards the majority group caused by the unbalanced multiclass dataset. The dataset from the UCI ML repository is utilised for developing the multiclass classification model, and the dataset includes the features of seven distinct varieties of dried beans. To address the skewness of the dataset, a Box-Cox transformation (BCT) was performed on the dataset's attributes. The 22 ML classification algorithms have been applied to the balanced and preprocessed dataset to identify the best ML algorithm. The ML algorithm results have been validated with a 10-fold cross-validation approach, and during validation, the CatBoost ML algorithm achieved the highest overall mean accuracy of 93.8 percent, with a range of 92.05 percent to 95.35 percent.

1. Introduction

People eat dry beans, which are a type of legume that is self-pollinated. Beans are a significant crop on a global scale and are popular with both farmers and consumers. Dry beans account for nearly 50 percent of the grain legumes consumed directly by humans in the majority of developing countries [1]. Beans are a staple food in Sub-Saharan Africa, where they are consumed by more than 200 million people [2]. A system of quality control makes sure that approved seed meets national and global quality benchmarks. For the majority of food products, visual characteristics are the primary criterion used by consumers when making purchasing decisions [3]. Like other legume species, common beans show the most variation in terms of growth patterns, physical features (size, shape, and shading), maturity, and ability to grow and adapt [4, 5]. Sorting and classifying bean seeds manually is a time-consuming process. Additionally, this method is inefficient and tedious, particularly when working with large production volumes. Human inspectors are usually in charge of checking raw materials, and it is difficult to streamline the inspectors' findings. These considerations reaffirm the importance of objective measurement systems. As a result, automatic grading and classification methods are required.

Recent technological changes have helped researchers in this field a lot. Computer vision systems (CVSs) are being used for quality control and have recently begun to be used as an objective measurement and evaluation system [6–9]. CVS technology, which is primarily camera cum computer based, has been considered for sensory characteristics of agricultural products. This system consists of a light source, an image acquisition device, and computer peripherals and software. The digital repository systems provide this information widely with various attributes.

Equal numbers of input samples represent each output class (or target class), which is known as a balanced dataset. Imbalanced training data has a major negative impact on real-time performance [10]. The majority of the reported studies used a target class with an uneven distribution of observations, i.e., an imbalanced dataset.

The main contribution of this research is to develop the unbiased ML based multiclass classification model to identify the dry bean variety with the best accuracy using the balanced dry bean dataset available at the UCI ML digital repository [11]. Using the preprocessed balanced dataset, the dry bean types such as "Dermason," "Sira," "Seker," "CAli," "Bombay," "Horoz," and "Barbunya" have been identified without losing any features available in the dataset. The 22 ML algorithms have been evaluated with 10-fold cross-validation to identify the best ML multiclass dry bean classification model. To make the model more accurate, the BCT was used to reduce the skewness of the dataset's attributes, making them almost identical to a normal distribution.

2. Related Work

Kilic et al. [12] used computer vision to develop the classification system for bean varieties. The system consisted of hardware and software. The hardware was developed to capture a standard image from the samples. The software part discusses segmentation, morphological operation, and colour quantification of the samples. The 69 samples have been used in their artificial neural network (ANN) model. The system's overall performance in classifying beans was 90.56 percent.

Using an infrared hyperspectral imagery method that works in the wavelength range of 390–1050 nm, Sun et al. [13] examined a quick and nondestructive method for categorising black bean variants. The primary component of the image was used to extract 16 textural and 6 morphological features by using ray level co-occurrence matrix analysis. Hasan et al. [14] examined various categories of dry beans and used a deep neural network-based method to categorise them. The outcomes indicate that their approach was 93.44 percent accurate and had an F-1 score of 94.57 percent when applied to a dataset of seven varieties of dry beans.

Giza3, Giza461, Misr1, Nobarya1, and Sakha1 are the five varieties of Egyptian faba-bean seeds studied by Abdulwahed et al. [15]. This method uses morphological features and an ANN to grade and classify the quality of Egyptian faba-bean seeds. Based on 15 physical traits of the seeds, artificial neural networks separated faba beans into different types.

It was presented by Araújo et al. [16] to develop a computer-based visual inspection system for beans that used correlation-based multishape granulometry in order to locate each grain in an image as well as its size and eccentricity. Using this method, their system correctly located 29,993 out of 30,000 grains, even when there were a lot of "glued" grains in the image.

De Oliveira et al. [17] used ANN as the transformation model and the Bayes as the classifier to identify the coffee beans types such as whitish, cane green, green, and bluish-green. The neural network models achieved a generalisation error of 1.15 percent, and the Bayesian classifier identified all samples.

Gope and Fukai [18] discussed the assessment of the Raspberry Pi 3 system's capacity in low-income countries for classifying peaberries and normal beans. They discovered that due to hardware constraints in the case of large-sized images, the Raspberry Pi 3 could not complete computation with linear support vector machines (SVMs) and k-nearest neighbors (kNNs).

Arboleda et al. [19] created the classification model for identifying coffee bean species. From 195 training images and 60 testing images, significant coffee bean morphology attributes such as bean area, perimeter, equivalent diameter, and percentage of roundness were extracted. The coffee beans were automatically classified using ANN and kNN. ANN obtained classification scores of 96.66 percent.

Koklu and Ozkan [11] used CVS to develop a multiclass classification of dry beans. The CVS-derived bean images were subjected to segmentation and feature extraction stages, yielding a total of 16 features, 12 dimensions, and 4 shape forms from the grains. With 10-fold cross validation, multilayer perceptron (MLP), SVM, kNN, and decision tree (DT) classification models were developed, achieving overall classification rates of 91.73 percent, 93.13 percent, 87.92 percent, and 92.52 percent for MLP, SVM, kNN, and DT, respectively. Table 1 shows the methodology and performance of various classification approaches for bean variety classification.

In this article, the proposed multiclass classification model uses the balanced dataset with 16 features and 7 varieties of dry beans. To avoid classification biassing of ML algorithms towards the majority group due to the unbalanced multiclass dataset, each dry bean type has 522 instances (522 $*$ 7) with 16 features in the processed dataset.

3. Exploratory Data Analysis and Methodology

The proposed multi-class classification model is depicted in Figure 1. The model's initial stage is data preprocessing. The second stage of the model is the Box-Cox transformation, and the final stage is ML model development.

3.1. Data. The data science process is a methodical way to address a data problem. In most scenarios, a data science project will have to go through five critical stages: problem definition, data processing, modelling, evaluation, and implementation. The dry bean dataset for this research was obtained from the UCI ML repository, which is accessible at [11]. It is also available as a supplementary file with this

Table 1: Recent research activity on bean variety classification.

Research works	Methodologies	Findings
Hasan et al. [14]	Deep neural network was implemented to identify the various categories of dry beans	93.44 percent accurate and had an F1 score of 94.57 percent when applied to a dataset of seven varieties of dry beans
Koklu and Ozkan [11]	Dry beans with 7 varieties were identified using ML algorithms	Achieved overall classification rates of 91.73 percent, 93.13 percent, 87.92 percent, and 92.52 percent for MLP, SVM, kNN, and DT, respectively
Arboleda et al. [19]	195 training images and 60 testing images coffee bean species were used with ANN	Obtained classification scores of 96.66 percent
De Oliveira et al. [17]	Employed ANN as the transformation model and the Bayes as classifier to identify the coffee beans types such as whitish, cane green, green, and bluish-green	Achieved a generalisation error of 1.15 percent
Kilic et al. [12]	69 samples of beans were used to develop the neural network-based classification system for beans	The system's overall performance in classifying beans was 90.56 percent

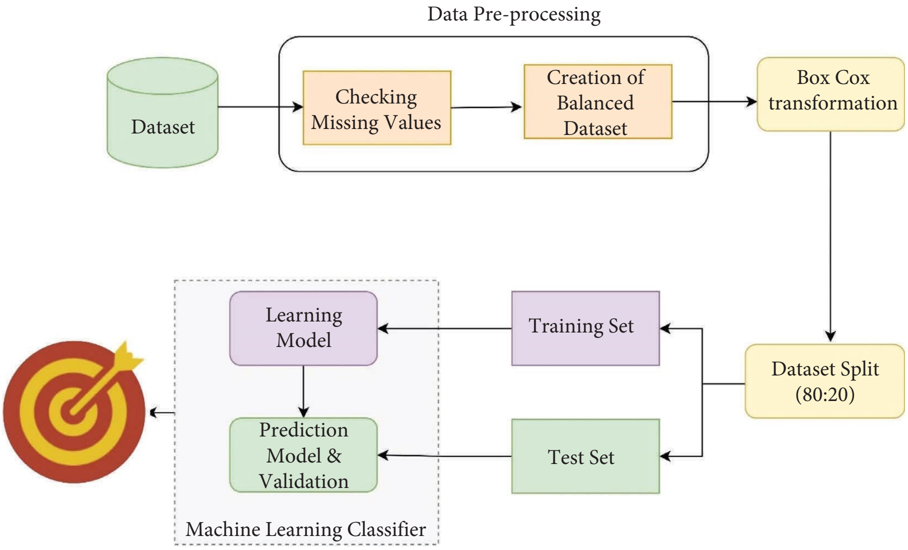


Figure 1: Proposed classification model.

article. The dataset contains information about the images taken with a high-resolution camera of 13,611 grains of seven different registered dry beans. From the grains, a total of 16 features were extracted. This study examined seven distinct varieties of dried beans, with market conditions dictating features such as aspect, shape, category, and structure. The dataset is available in.csv format for the dry bean varieties "Dermason," "Sira," "Seker," "CAli," "Bombay," "Horoz," and "Barbunya" with a total of 13611 instances. Table 2 shows quantile and descriptive statistics for 16 features of the dry bean dataset.

3.2. Data Preprocessing. Preprocessing strategies improve the performance of classifiers [20]. The information extraction (IE) method of extracting structured content such as entities, interactions, facts, and terms, as well as other kinds of information that aid the data analysis pipeline in prepping the data for the study [21]. The distribution in the dry bean variants of dry bean dataset is shown in Figure 2. Figure 2(a) shows the percentage of distribution of seven dry bean varieties, and Figure 2(b) shows the individual dry bean variety count in the raw dataset. It is observed that the dry bean type "DERMASON" has appeared at a maximum of 26.1 percent and the dry bean type "BOMBAY" at a minimum of 3.84 percent. The most frequently encountered problem in data quality is the absence of feature values in some entries. The missing values for each instance have been checked. The total data set instances become 13543 from 13611 instances after dropping the duplicate instances. Classification is a process that can be applied to structured or unstructured data. The class wise count of dry bean dataset is 3546, 2636, 2027, 1860, 1630, 1322 and 522 for DERMASON, SIRA, SEKER, HOROZ, CALI, BARBUNYA, and BOMBAY, respectively, after dropping the duplicate instances. Except for the target "Class," all feature data types have been converted to float.

3.2.1. Creation of a Balanced Dataset. Because of the unbalanced multiclass dataset, learning algorithms will be influenced towards the majority population. In contrast, the minority class is typically more significant from the perspective of data mining, as it may contain valuable information amidst its rarity. When encountered with such disparities, the researchers should design an effective model capable of handling the bias. This is referred to as learning from unbalanced data [22]. In terms of balancing distributions, there are methods for creating new objects for the minority group (over sampling) and methods that eliminate instances from the majority group (under sampling) [23]. Overfitting may result from the creation of new instances for the minority group. As a result, the random undersampling method used in this article will make the majority group of instances in the dry beans dataset matchable with the minority dry bean group. All of the dry bean types of instances were brought to 522 instances uniformly using the random undersampling method. This can be observed in Figure 3. To develop the model, a balanced dataset with 3654 instances has been considered. Each bean variety has 522 instances.

The steps followed in the creation of a balanced dataset are as follows:

(i) Step 1: The majority and minority classes in the dataset have been identified. The majority class index in the preprocessed dataset is "DERMASON," and the minority class index is "BOMBAY," with 3546 and 522 instances, respectively.

(ii) Step 2: The number of instances of "BOMBAY" is less by comparing all other classes. It is decided that the maximum number of instances is 522 for each variety of bean.

(iii) Step 3: The random samples of other bean varieties have been chosen.

(iv) Step 4: All the samples have been concatenated, and the balanced dataset has been created.

Table 2: Dataset statistics.

Features	Area	Perimeter	MajorAxisLength	MinorAxisLength	AspectRation	Eccentricity	ConvexArea	EquivDiameter	Extent	Solidity	Roundness	Compactness	ShapeFactor1	ShapeFactor2	ShapeFactor3	ShapeFactor4
Quantile statistics																
Minimum	20786.00	534.72	192.80	122.51	1.025	0.2190	21057.00	162.68	0.5724	0.9446	0.490	0.641	0.0028	0.0006	0.410	0.957
5-th percentile	29398.35	636.31	229.51	158.86	1.213	0.5656	29732.30	193.47	0.6554	0.9775	0.764	0.691	0.0034	0.0008	0.478	0.986
Q1	40329.00	743.56	268.51	184.64	1.443	0.7209	40761.25	226.60	0.7226	0.9848	0.823	0.758	0.0053	0.0010	0.574	0.993
Median	52981.50	911.06	350.62	203.69	1.575	0.7724	53858.50	259.73	0.7637	0.9880	0.866	0.795	0.0063	0.0014	0.632	0.996
Q3	75098.75	1070.86	405.75	242.33	1.727	0.8154	76264.75	309.22	0.7888	0.9900	0.908	0.831	0.0069	0.0020	0.690	0.998
95-th percentile	181172.55	1627.96	615.47	381.17	2.078	0.8766	183753.30	480.29	0.8150	0.9921	0.958	0.907	0.0080	0.0027	0.823	0.999
Maximum	254616	1985.37	738.86	460.20	2.430	0.9114	263261	569.37	0.8584	0.9947	0.988	0.987	0.0105	0.0037	0.975	1.000
Range	233830	1450.65	546.06	337.69	1.405	0.6925	242204	406.69	0.2860	0.0501	0.498	0.347	0.0077	0.0031	0.564	0.042
Interquartile range (IQR)	34769.75	327.30	137.24	57.70	0.284	0.0945	35503.50	82.62	0.0662	0.0052	0.085	0.073	0.0016	0.0010	0.116	0.005
Descriptive statistics																
Standard deviation	46096.53	296.60	115.00	66.84	0.246	0.0915	46750.50	85.16	0.0495	0.0049	0.0600	0.061	0.001	0.001	0.099	0.005
Coefficient of variation (CV)	0.66	0.31	0.32	0.29	0.154	0.1211	0.66	0.30	0.0658	0.0050	0.0695	0.077	0.229	0.401	0.155	0.005
Kurtosis	2.01	0.56	0.46	1.09	0.000	1.6840	2.02	0.89	0.7746	7.0853	0.4222	−0.223	−0.289	−0.587	−0.104	6.656
Mean	69718.99	969.54	362.43	227.47	1.597	0.7558	70681.22	285.52	0.7525	0.9868	0.8637	0.796	0.006	0.002	0.638	0.994
Median absolute deviation (MAD)	15923.50	164.53	70.33	27.03	0.141	0.0460	16298.00	39.65	0.0301	0.0024	0.0424	0.037	0.001	0	0.058	0.002
Skewness	1.73	1.14	1.02	1.43	0.465	−1.1675	1.73	1.32	−0.9632	−1.996	−0.442	0.156	−0.458	0.653	0.358	−2.035

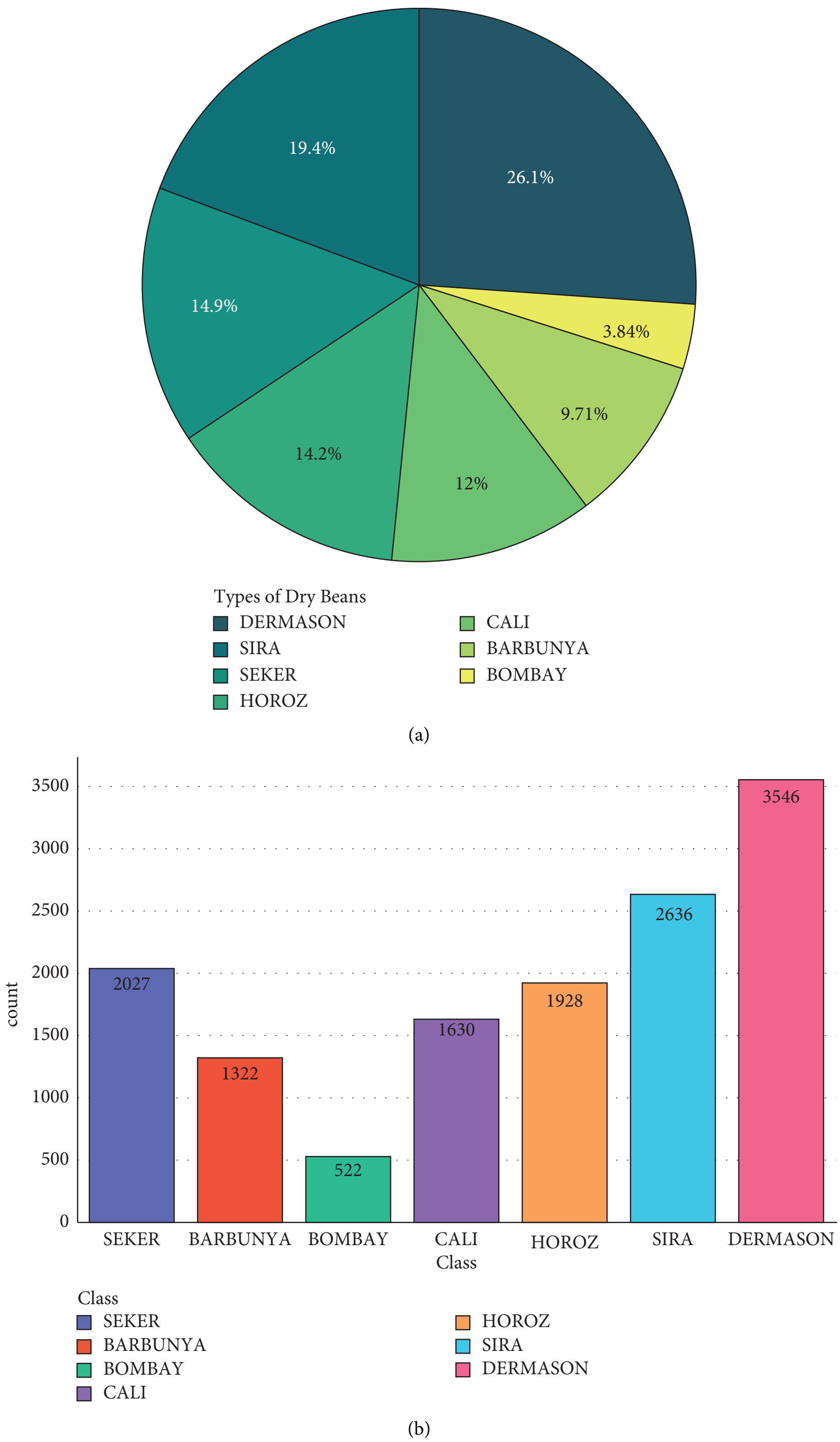


Figure 2: Dry bean raw dataset (a) percentage of distribution of dry bean varieties (b) dry bean varieties count of the raw dataset.

3.3. Box-Cox Transformation. When handling with a skewed outcome, investigators use log transformation to normalise the data before applying standard statistical tests, such as the *t*-test, linear regression, etc. Nevertheless, log-transformed data will not always be normal. In such instances, BCT can be implemented to normalise skewed data [24]. Initially, the dry bean dataset features were applied with log transformation. It fails with a reduction in negative skewness. As shown in Figures 4(a)–4(p),

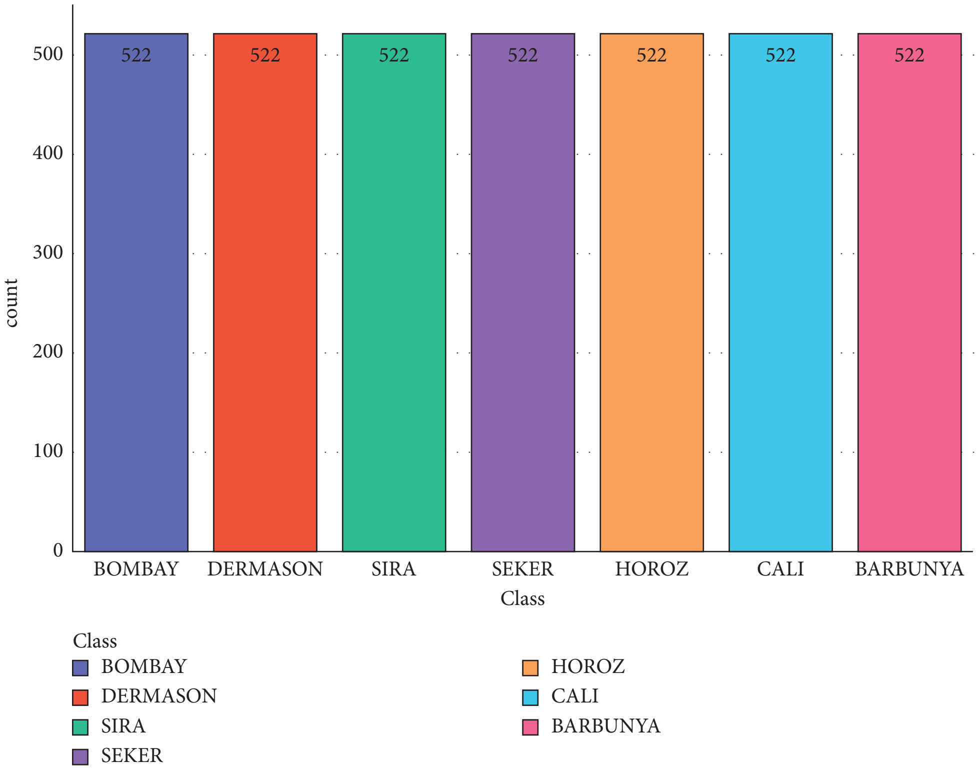


Figure 3: Balanced dry bean dataset.

the BCT was applied to all of the features of the dataset for transforming the skewed data into a normal distribution. For each attribute, the figure on the left shows the distribution before BCT, and the figure on the right shows the distribution after BCT. The skewness can be found at the top right corner of the figure. Y represents the dependent (continuous) variable, while X represents the independent variables ($1, x_1, x_2, \ldots, x_k$). In the equation, the BCT [24] used to transform the skewed distribution into a normal distribution without the original scale is given (1). The maximum likelihood technique is commonly used to determine the parameter lambda (λ).

$$Y_{\mathrm{BCT}}(Y,\lambda) = X\beta + \sigma\varepsilon, \tag{1}$$

where

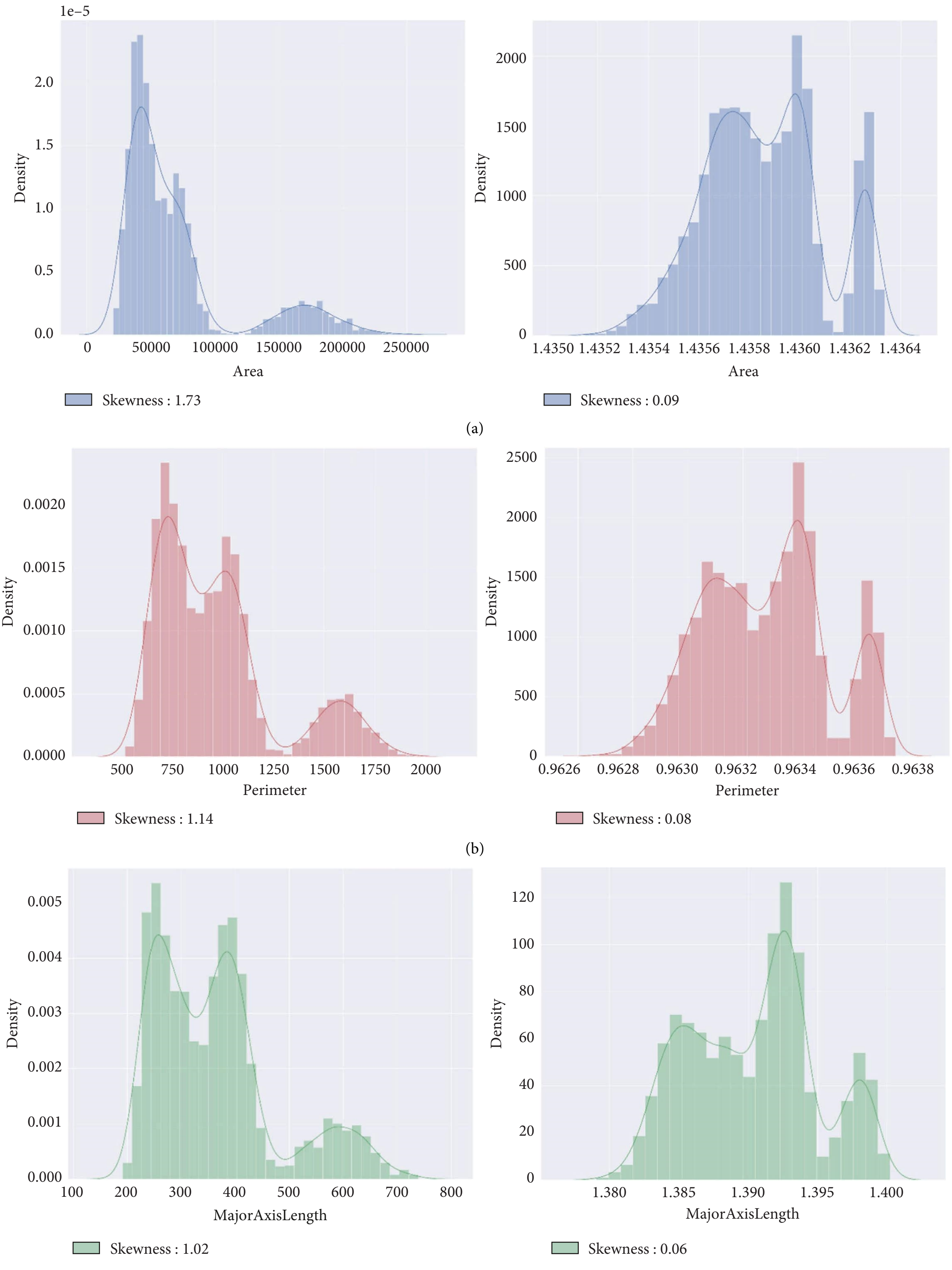
$$Y_{\mathrm{BCT}}(Y,\lambda) = \begin{cases} \dfrac{Y^{\lambda}-1}{\lambda}, & \text{if } \lambda \neq 0, \\ \log Y, & \text{if } \lambda = 0. \end{cases} \tag{2}$$

X is the covariate matrix, which includes the intercept. β is a regression coefficient vector. σ is the variance of random error. ε is a random error.

3.4. Machine Learning Model

3.4.1. Training and Test Dataset. The training dataset is the set of data used to construct the model, which contains known features and a target. The created model will also need to be validated against another well-known dataset known as the test dataset or validation dataset. To meet this challenge, the entire known dataset can be divided into training and a test set [25]. The dry bean categorical classes, namely "SIRA," "BOMBAY," "DERMASON," "BARBUNYA," "HOROZ," "CALI," and "SEKER" were converted into integer types as 1–7, respectively. The training and test sets have been split in an 80 : 20 ratio, with 2923 and 731 instances with 16 features, respectively.

3.4.2. Machine Learning Algorithm (MLA) Selection. A model built with a single method may not offer the best prediction for a specific dataset. Each machine learning technique has its own constraints and creating a model with significant accuracy is difficult. The 22 MLAs were used to determine the accuracy of various MLAs on a balanced dataset. It helps us to bring out a better predictive model. The 10-fold cross validation has been performed and the mean accuracy of 19 MLAs has been listed in Table 3. Ensemble methods [26] such as AdaBoost classifier, Bagging classifier, and extra tree classifier, generalised linear models [27] like logistic regression, passive aggressive classifier, Ridge classifier, stochastic gradient descent classifier, and perceptron, Navies Bayes models [28] like Bernoulli and Gaussian MLA, kNN, and SVM algorithms [29], tree-based methods [30] such as DT classifier and extra tree classifier, and discriminant analysis methods [31] such as linear and quadratic discriminant analysis. Gaussian process MLAs have been evaluated with 10-fold cross-validation. Figure 5 displays the mean accuracy of MLA performance with 10-fold cross validation in descending order. The logistic regression

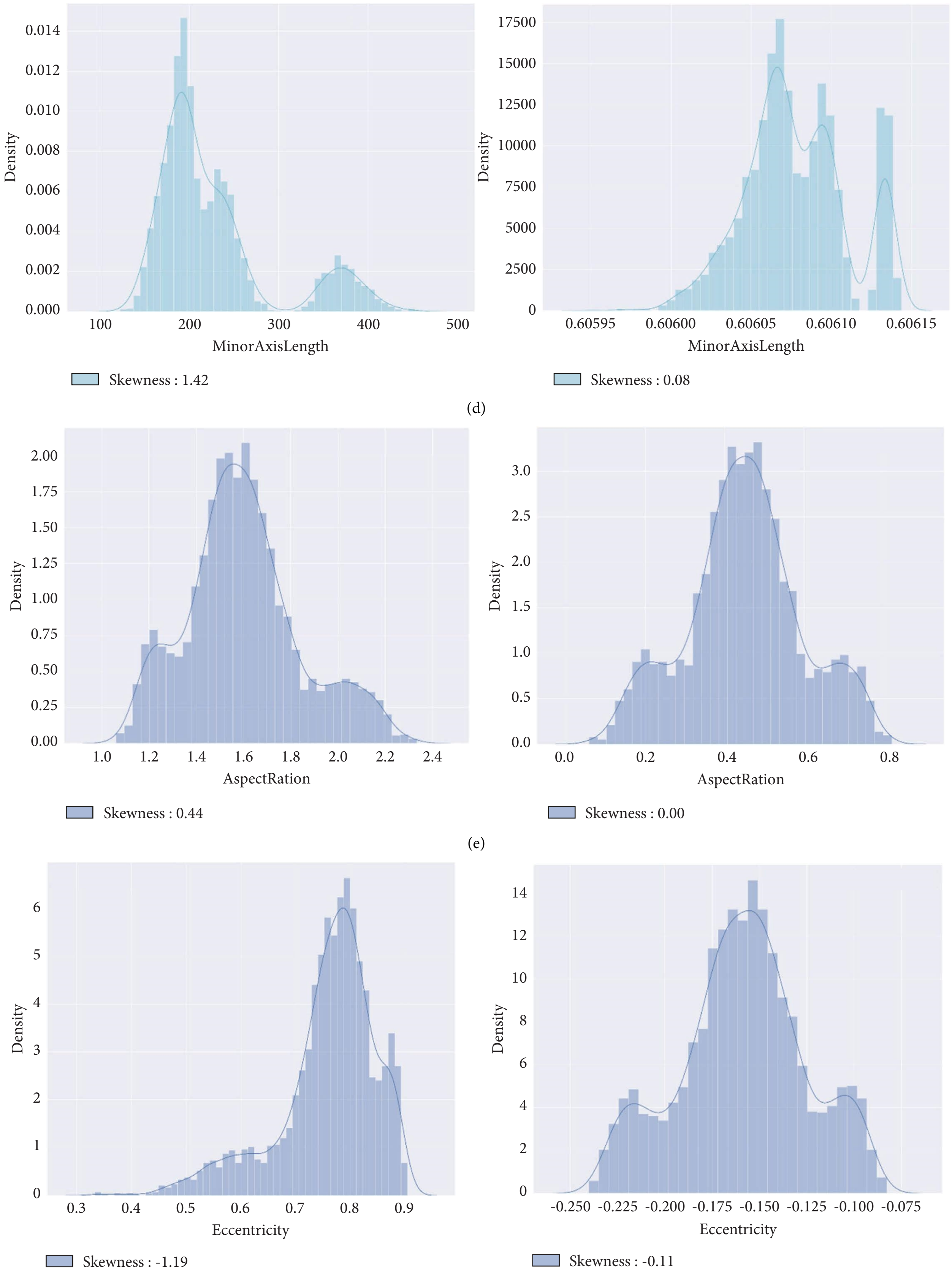


(a)

(b)

(c)

Figure 4: Continued.

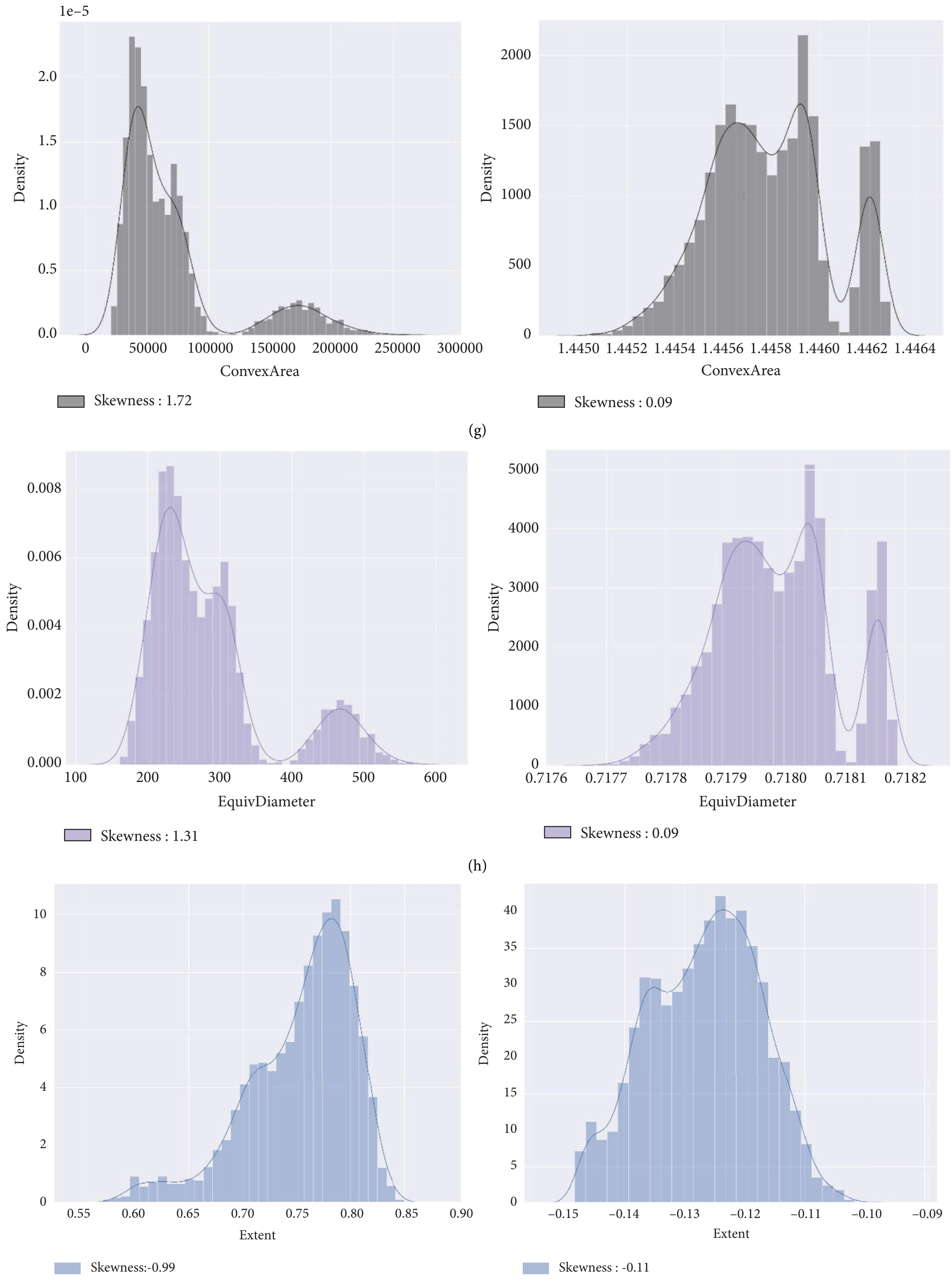


(d)

(e)

(f)

Figure 4: Continued.

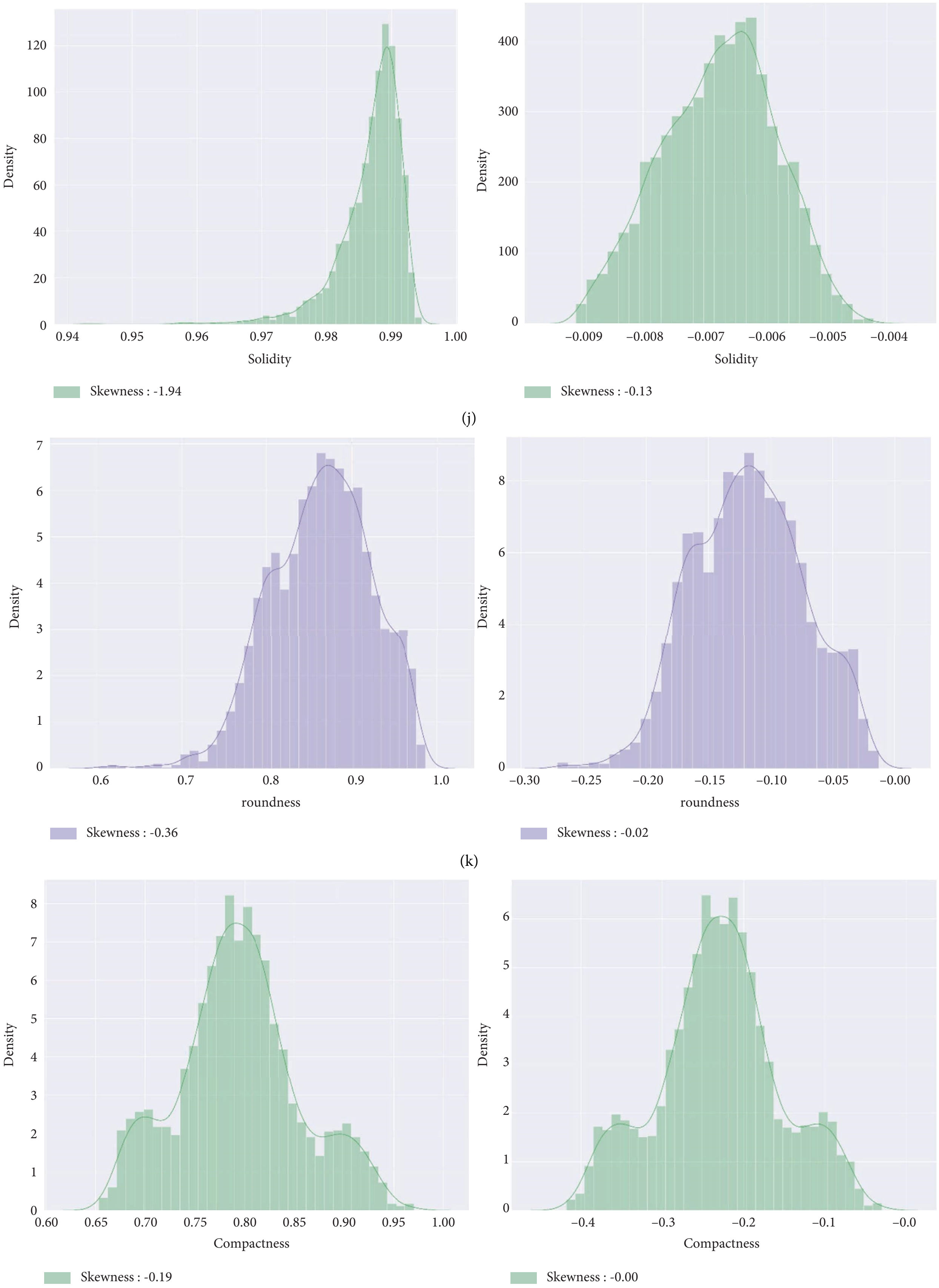


(g)

(h)

(i)

Figure 4: Continued.

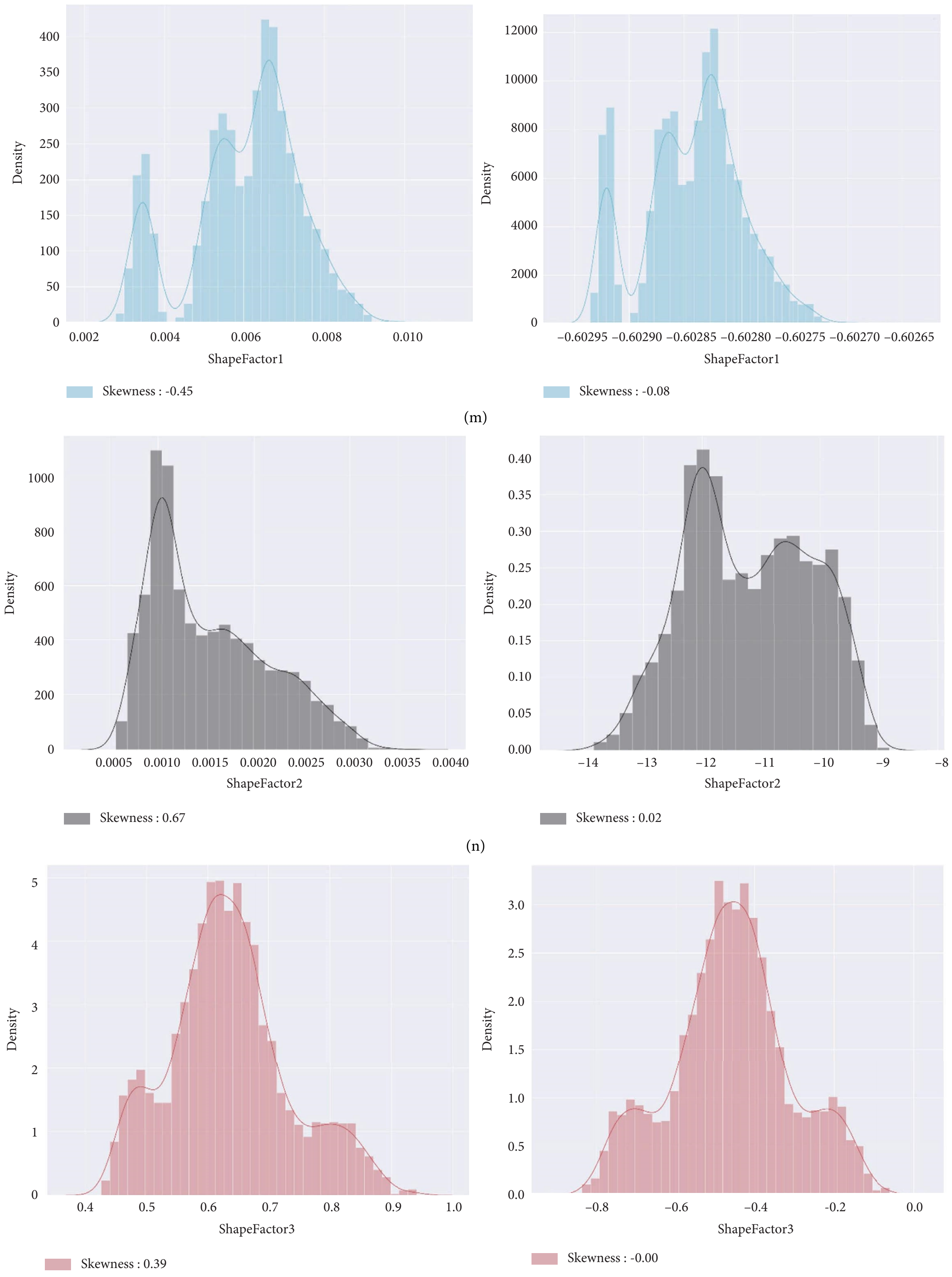


(j)

(k)

(l)

Figure 4: Continued.

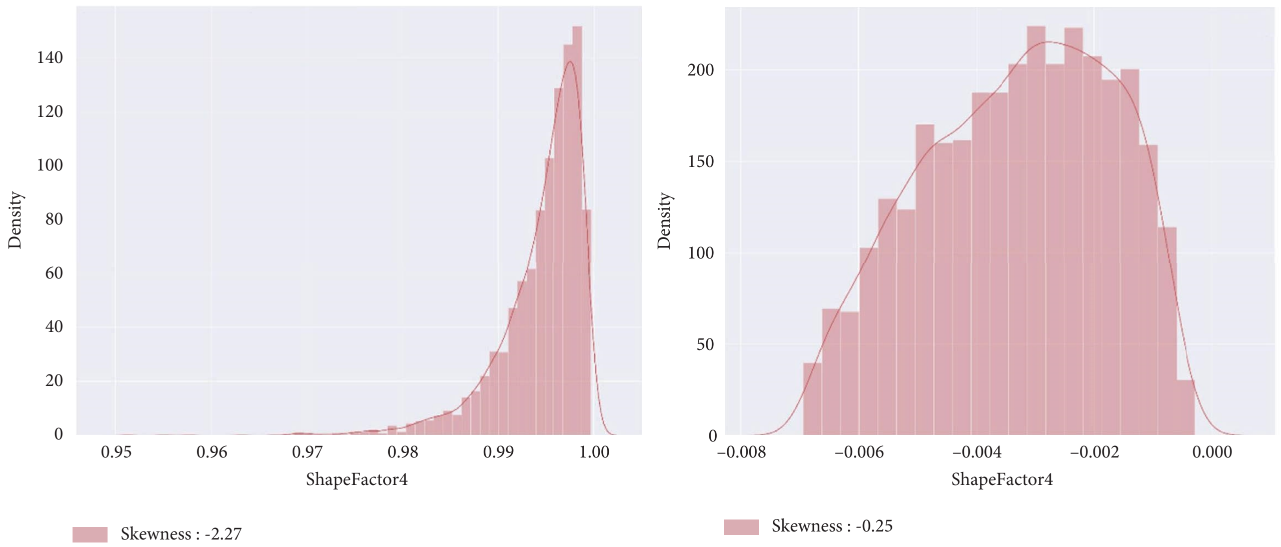


(m)

(n)

(o)

Figure 4: Continued.



(p)

Figure 4: Distribution plot of BCT for the balanced and preprocessed dry bean dataset. (a) Area. (b) Perimeter. (c) MajorAxisLength. (d) MinorAxisLength. (e) AspectRatio. (f) Eccentricity. (g) ConvexArea. (h) EquivDiameter. (i) Extent. (j) Solidity. (k) Roundness. (l) Compactness. (m) ShapeFactor1. (n) ShapeFactor2 (o) ShapeFactor3. (p) ShapeFactor4.

Table 3: MLA accuracy.

S.Nos	MLA names	MLA parameters	MLA test accuracy mean
1	LogisticRegressionCV	{“Cs”: 10, “class_weight”: None, “cv”: None, “. . .”	0.926949
2	ExtraTreesClassifier	{“bootstrap”: False, “ccp_alpha”: 0.0, “class_. . .”	0.921751
3	BaggingClassifier	{“base_estimator”: None, “bootstrap”: True, “b. . .”	0.921614
4	LinearDiscriminantAnalysis	{“covariance_estimator”: None, “n_components”:. . .	0.911628
5	KNeighborsClassifier	{“algorithm”: “auto”, “leaf_size”: 30, “metric. . .”	0.902326
6	DecisionTreeClassifier	{“ccp_alpha”: 0.0, “class_weight”: None, “crit. . .”	0.898222
7	GaussianNB	{“priors”: None, “var_smoothing”: 1e-09}	0.897127
8	ExtraTreeClassifier	{“ccp_alpha”: 0.0, “class_weight”: None, “crit. . .”	0.88632
9	LinearSVC	{“C”: 1.0, “class_weight”: None, “dual”: True,. . .	0.877839
10	GaussianProcessClassifier	{“copy_X_train”: True, “kernel”: None, “max_it. . .”	0.842681
11	NuSVC	{“break_ties”: False, “cache_size”: 200, “clas. . .”	0.835021
12	SGDClassifier	{“alpha”: 0.0001, “average”: False, “class_wei. . .”	0.677018
13	SVC	{“C”: 1.0, “break_ties”: False, “cache_size”:. . .	0.664432
14	PassiveAggressiveClassifier	{“C”: 1.0, “average”: False, “class_weight”: N. . .	0.647469
15	RidgeClassifierCV	{“alphas”: array([0.1, 1. , 10.]), “class_w. . .”	0.647196
16	AdaBoostClassifier	{“algorithm”: “SAMME.R′, “base_estimator”: Non. . .	0.597127
17	Perceptron	{“alpha”: 0.0001, “class_weight”: None, “early. . .”	0.518878
18	QuadraticDiscriminantAnalysis	{“priors”: None, “reg_param”: 0.0, “store_cova. . .”	0.412038
19	BernoulliNB	{“alpha”: 1.0, “binarize”: 0.0, “class_prior”:. . .	0.12777

provides the highest test accuracy of 92.69 percent in the 19 MLAs and the lowest accuracy of 12.77 percent found with the Bernoulli Naive Bayes ML classifier.

From the initial screening during validation, it is observed that the XGBoost, RF, and CatBoost algorithms offer greater precision. Therefore, in the following sections, the performance of these three algorithms with an 80 : 20 balanced dry bean dataset and with 10-fold cross validation is described.

3.4.3. Random Forest Algorithm. The DT modelling is an important part of RF. It is used on several samples of the original data obtained by the bootstrap method. Samples of the original data are used to make the bootstrap samples, and each sample has the same number of data points as the original data. The RF [32] constructs multiple DTs as well as merges them to produce more precise and stable predictions. The node's importance is calculated as follows:

$$ni_j = w_j C_j - w_{\mathrm{left}(j)} C_{\mathrm{left}(j)} - w_{\mathrm{right}(j)} C_{\mathrm{right}(j)}, \qquad (3)$$

where C_j = node j's impurity value, w_j = the weighted sample size arriving at the node j, and right(j) and left(j) are the child node from right and left split on node j, respectively.

An individual attribute's feature importance is

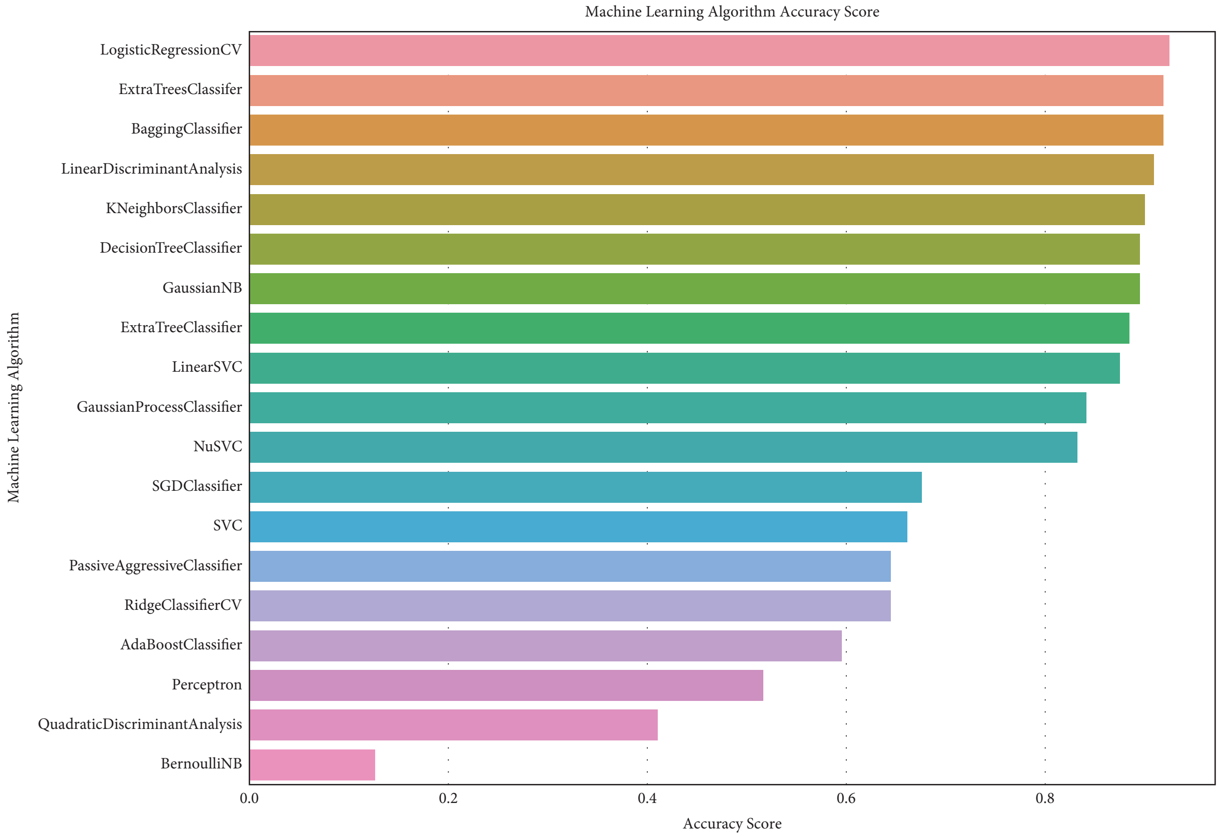


Figure 5: MLA mean accuracy with a balanced dry bean dataset.

$$fi_i = \frac{\sum j\text{: node } j \text{ splits on featurei } \mathrm{ni}_j}{\sum k \in \text{all nodes } \mathrm{ni}_k}. \tag{4}$$

3.4.4. Extreme Gradient Boost. XGBoost [33] is a framework of the gradient boosting machine (GBM), a well-known algorithm for supervised learning. It is appropriate to both classification and regression tasks.

If DS is the set of data containing "m" attributes, then for "n" occurrences

$$\mathrm{DS} = \left\{ (x_i, y_i) \colon i = 1 \ldots, n, x_i \in \mathbb{R}^m, y_i \in \mathbb{R} \right\}. \tag{5}$$

Let $\hat{y}_i$ be the ensemble tree model's target value constructed using the equation.

$$\hat{\mathrm{y}}_{\mathrm{i}=\phi}\left(x_i\right)_= \sum_{k=1}^{K} f_k\left(x_i\right), f_k \in \mathcal{F}. \tag{6}$$

Here K denotes the model's total number of trees and f_k denotes the model's k^{th} tree. Classification and Regression Trees (CART) serve as the base learner for Gradient Boosted Trees, which is a popular machine learning algorithm for both classification and regression problems. F's functional space is f, and the set of feasible CARTs is F.

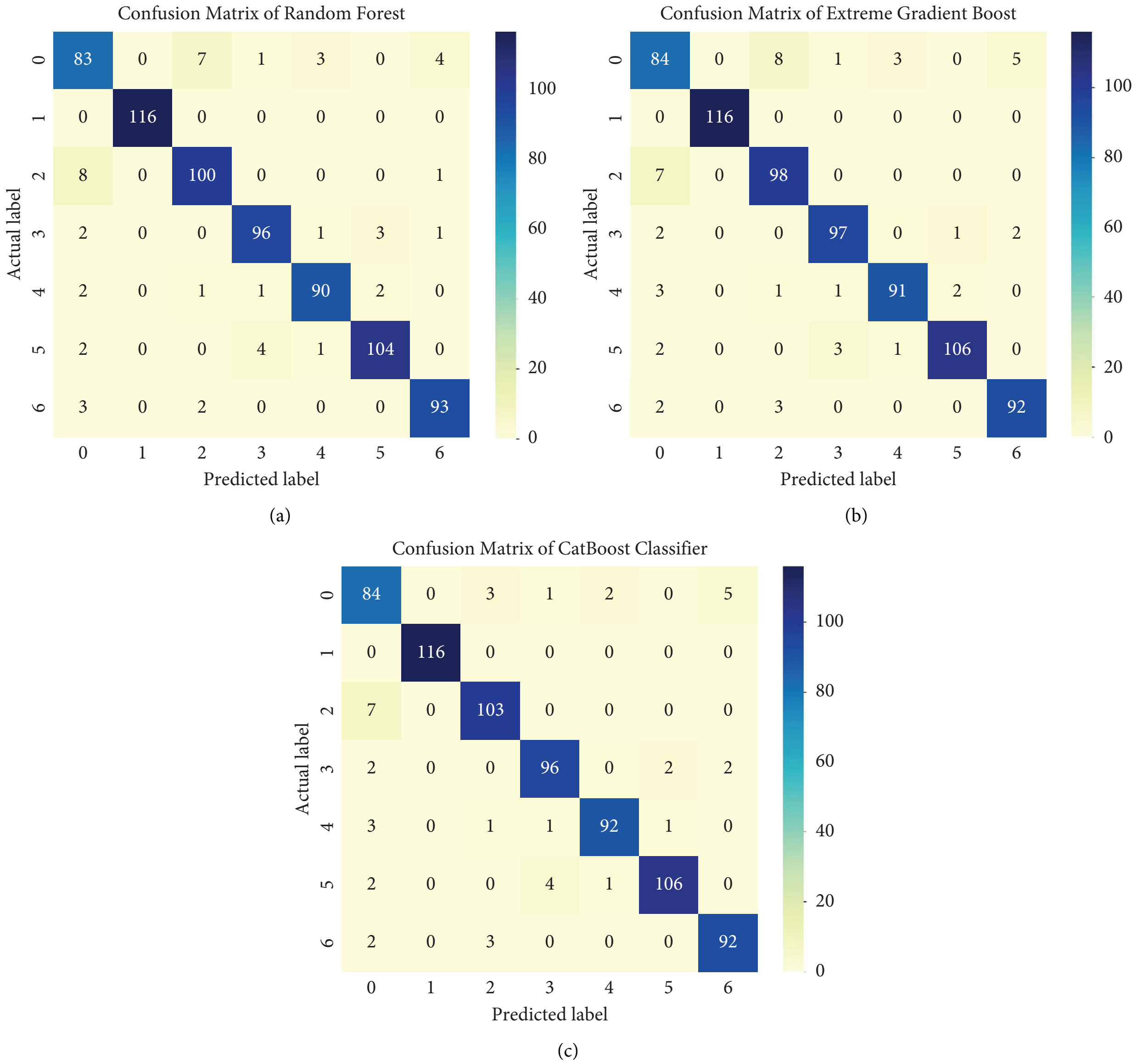
3.4.5. Cat Boost Classifier. Categorical boosting (CatBoost) is a Yandex-developed open-source boosting library [34]. CatBoost implements oblivious DTs (binary trees in which the same features have been used to create left and right splits for every level of the tree), thereby limiting the number of features split per level to a single instance, which aids in reducing prediction time. In the dataset "D" of dry beans, for every instance has "m" features in a vector "x" and the target dry bean class type, y.

Mathematically, the target assessment of the i^{th} categorical data of the k^{th} element of dry bean dataset D for dry beans can be expressed as follows:

$$\hat{x}_k^i = \frac{\sum x_j \in D_k \, 1_{x_k^i = x_k^j} \cdot y_j + ap}{\sum x_j \in D_k \, 1_{x_k^i = x_k^j} + a}; \text{if } D_k = \left\{ x_j \colon \sigma(j) < \sigma(i) \right\}, \tag{7}$$

when $a > 0$. When the i^{th} component of CatBoost's input vector x_j is equal to the i^{th} component of input vector x_k, the indicator function $1_{x_k^i = x_k^j}$ returns the value 1. The parameters "a" and "p" (prior) prevent underflowing in the equation. σ is a permutation at random.

3.5. Results and Discussion. The use of diverse bean varieties in dry bean cultivation actually inhibits the production of uniform crops. As a result, the resulting product, which includes a set of dried bean species, incurs economic losses. To address this issue, the purpose of this study is to distinguish the seven classes of dry beans cultivated in Turkey, as determined by the Turkish Standards Institute (TSE). The dry beans dataset has been processed through the developed model. The confusion matrix



Confusion Matrix of Random Forest

	0	1	2	3	4	5	6
0	83	0	7	1	3	0	4
1	0	116	0	0	0	0	0
2	8	0	100	0	0	0	1
3	2	0	0	96	1	3	1
4	2	0	1	1	90	2	0
5	2	0	0	4	1	104	0
6	3	0	2	0	0	0	93

(a)

Confusion Matrix of Extreme Gradient Boost

	0	1	2	3	4	5	6
0	84	0	8	1	3	0	5
1	0	116	0	0	0	0	0
2	7	0	98	0	0	0	0
3	2	0	0	97	0	1	2
4	3	0	1	1	91	2	0
5	2	0	0	3	1	106	0
6	2	0	3	0	0	0	92

(b)

Confusion Matrix of CatBoost Classifier

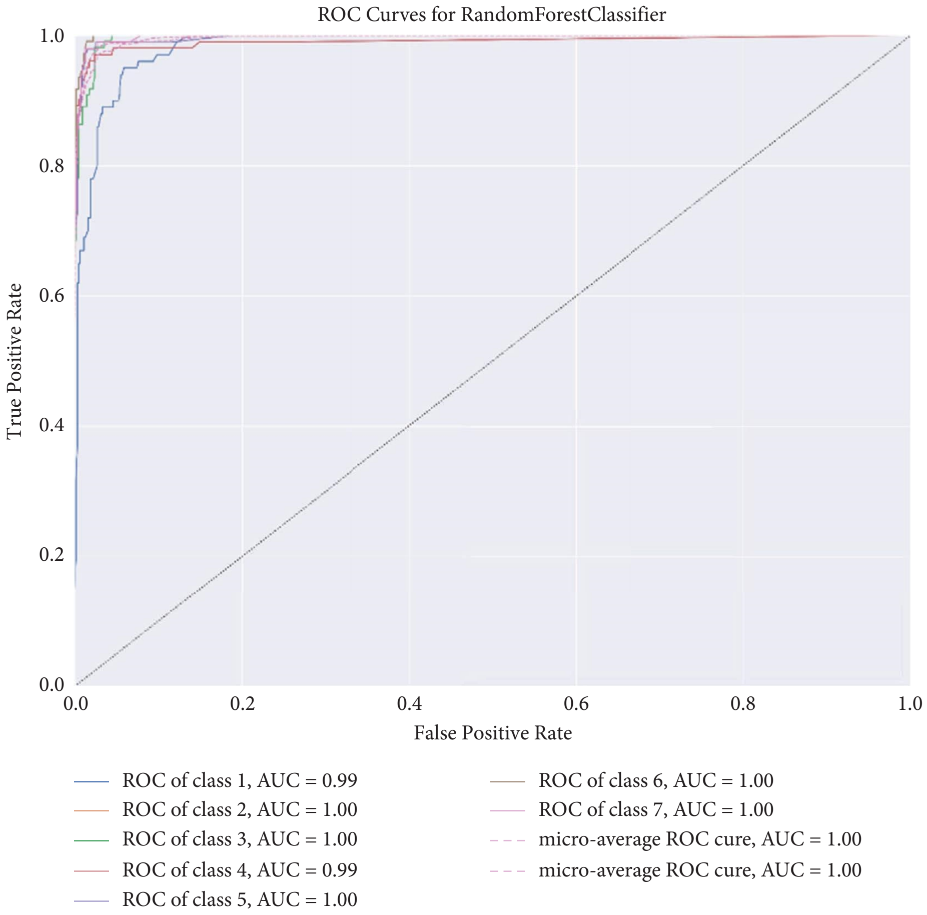
	0	1	2	3	4	5	6
0	84	0	3	1	2	0	5
1	0	116	0	0	0	0	0
2	7	0	103	0	0	0	0
3	2	0	0	96	0	2	2
4	3	0	1	1	92	1	0
5	2	0	0	4	1	106	0
6	2	0	3	0	0	0	92

(c)

Figure 6: Confusion matrix of (a) random forest, (b) XGBoost, and (c) CatBoost ML classification algorithms.

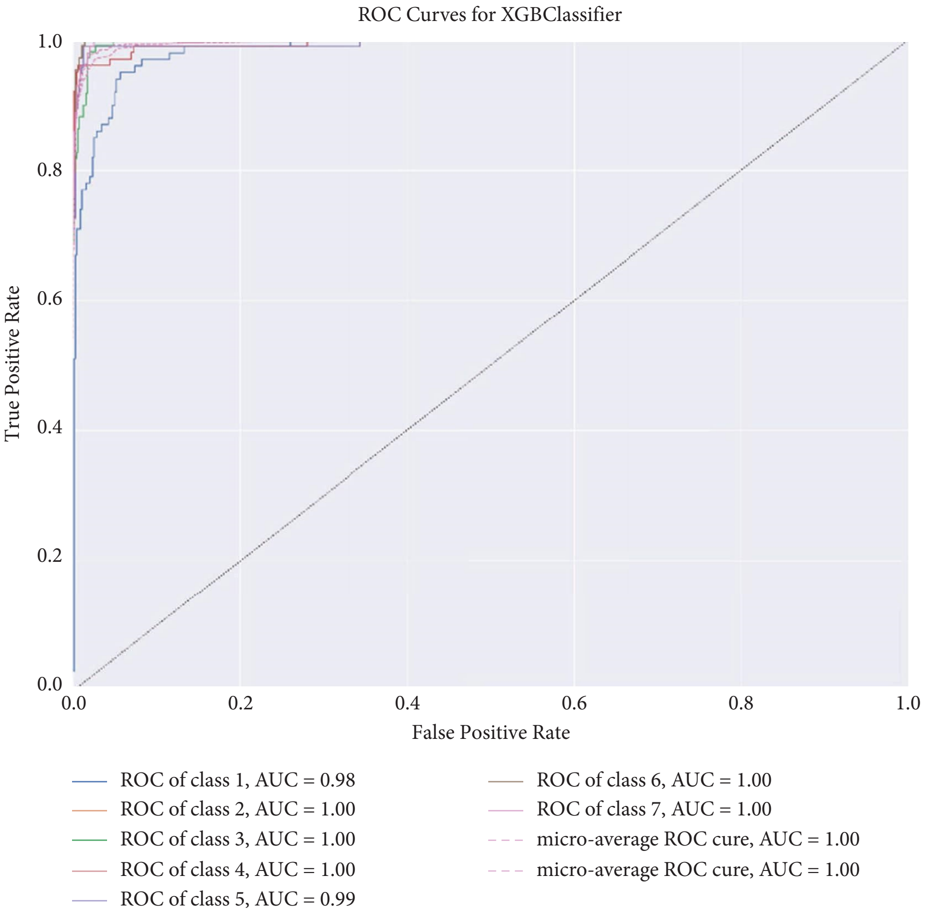
of three MLAs, namely RF, XGBoost, and CatBoost, is shown in Figure 6. Confusion matrices enable a more detailed visualisation of results and a comparison of actual and predicted values. In Figure 6, “SIRA,” “BOMBAY,” “DERMASON,” “BARBUNYA,” “HOROZ,” “CALI,” and “SEKER” are denoted as 0, 1, 2, 3, 4, 5, and 6. The correctly predicted sample numbers can be found in the diagonal part of the confusion matrix. The misclassified instances are available in other parts of the confusion matrix. For example, in Figure 6(c), for the dry bean variety “SIRA,” the correctly identified test set instances were 84. Seven test instances were identified as “DERMASON,” two instances were identified as “BARBUNYA,” three instances were identified as “HOROZ,” two instances were identified as “CALI,” and two instances were identified as “SEKER.” Figure 7 shows the receiver operating characteristic (ROC) curve that shows the performance of the RF, XGBoost, and CatBoost ML classification algorithms. ROC is the plot between true positive and false positive. In ROC, the area under the curve (AUC) represents the degree or measure of separability. It shows the model’s capability of distinguishing between dry bean classes. It is observed that the CatBoost algorithm provides the AUC value for the “SIRA” dry bean type as 0.99, and for other dry bean types such as “BOMBAY”, “DERMASON”, “BARBUNYA”, “HOROZ”, “CALI”, and “SEKER” has an AUC value of 1. Table 4 provides the performance metrics like precision, recall, and f1-score of the three ML algorithms, and Table 5 provides the ML model accuracy with an 80 : 20 dataset. The accuracy of the model has been improved by about 1.49 percent using the balance dataset and the CatBoost ML algorithm.

Among the 22 MLAs tested, it is observed that the CatBoost ML classifier provides the best performance. Table 6 shows the performance comparison with the existing method. The CatBoost ML classifier performs well as compared to the existing method under balanced instances for seven dry bean types.



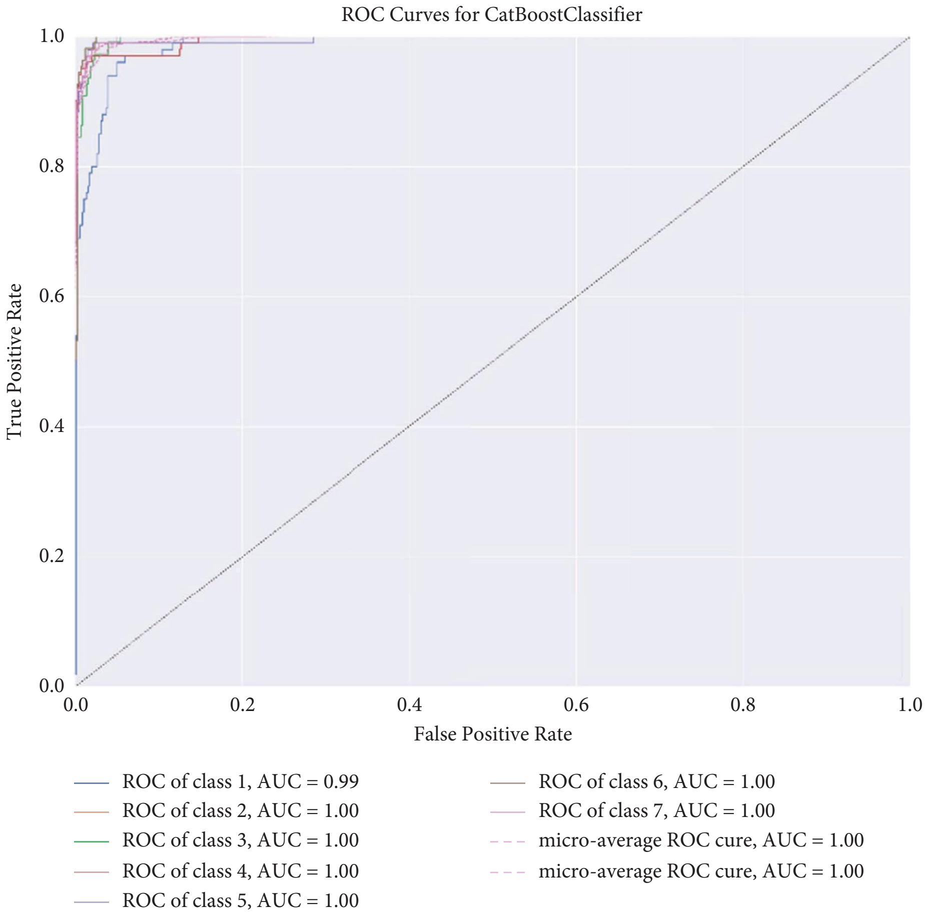
(a)

Figure 7: Continued.



(b)

Figure 7: Continued.



(c)

Figure 7: ROC curve of (a) random forest (b) XGBoost, and (c) CatBoost ML classification algorithms.

Table 4: Performance metrics of proposed MLAs for the seven varieties of dry beans.

ML classifiers	Classes	Precision	Recall	f1-score
Random forest	Sira	0.85	0.83	0.84
	Bombay	1	1	1
	Dermason	0.92	0.91	0.91
	Barbunya	0.93	0.94	0.94
	Horoz	0.94	0.95	0.94
	CAli	0.94	0.95	0.95
	Seker	0.95	0.94	0.94
XGBoost	Sira	0.83	0.84	0.84
	Bombay	1	1	1
	Dermason	0.93	0.89	0.91
	Barbunya	0.95	0.95	0.95
	Horoz	0.93	0.96	0.94
	Cali	0.95	0.97	0.96
	Seker	0.95	0.93	0.94
CatBoost	Sira	0.88	0.84	0.86
	Bombay	1	1	1
	Dermason	0.94	0.94	0.94
	Barbunya	0.94	0.94	0.94
	Horoz	0.94	0.97	0.95
	Cali	0.94	0.97	0.95
	Seker	0.95	0.93	0.94

Table 5: ML model accuracy for 80 : 20 data split.

ML models	Accuracy in percentage	
	Unbalanced dataset	Balanced dataset
Random forest	92.06	93.29
XGBoost	92.10	93.57
CatBoost	92.76	94.25

Table 6: Performance comparison with the existing method with 80 : 20 split.

Models	ML algorithms	Precision	Recall	f1-score	Accuracy in percentage
Koklu and Ozkan [11]	MLP	0.93	0.93	0.93	91.73
	SVM	0.94	0.94	0.94	93.13
	DT	0.89	0.88	0.88	87.92
	kNN	0.93	0.93	0.93	92.52
Proposed method	Random forest	0.93	0.93	0.93	93.29
	XGBoost	0.93	0.93	0.93	93.57
	CatBoost	0.94	0.94	0.94	94.25

Table 7: 10-Fold cross validation accuracy.

Fold nos	No. of instances (90 : 10)		Random forest	XGBoost	CatBoost
	Training sets	Test sets			
1	3288	366	92.89	94.53	94.53
2	3288	366	91.25	93.16	95.35
3	3288	366	92.34	92.07	94.81
4	3288	366	93.44	91.25	92.89
5	3289	365	92.05	92.87	93.97
6	3289	365	93.42	93.15	93.42
7	3289	365	95.06	93.15	93.69
8	3289	365	92.32	90.95	92.87
9	3289	365	93.97	95.06	94.52
10	3289	365	91.78	92.05	92.05
Mean accuracy in percentage			92.9	92.8	93.8

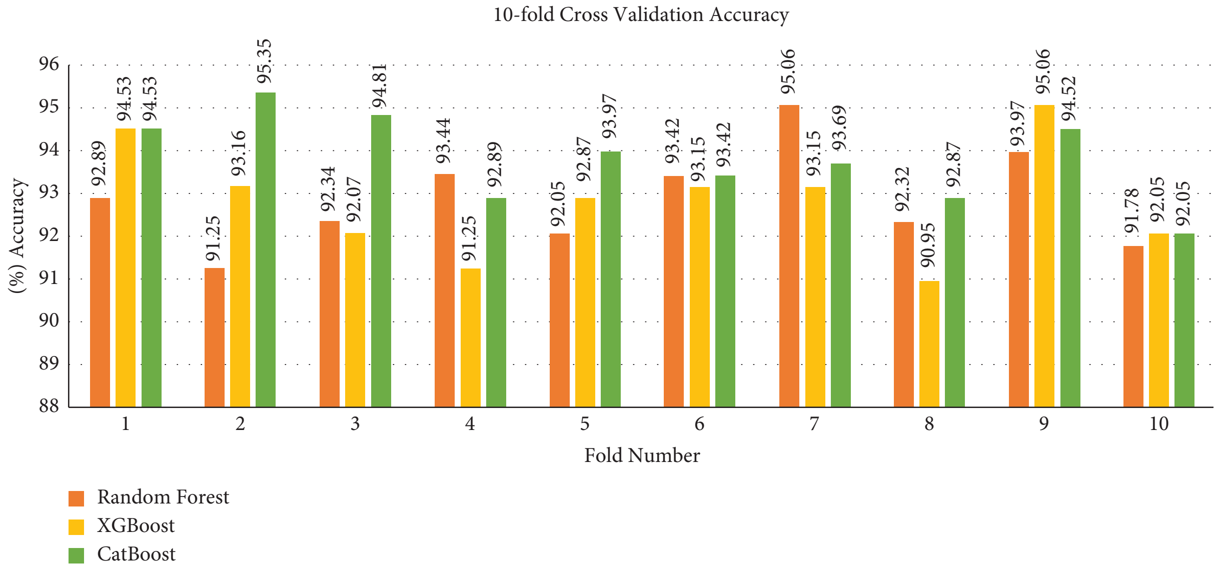


Figure 8: k-Fold cross validation accuracy of ML classification algorithms.

CatBoost ML excels at solving classification problems with heterogeneous data.

3.5.1. Model Performance with Cross-Validation (CV). The three algorithms RF, XGBoost, and CatBoost have been validated with 10-fold cross validation with 90 : 10 data split. In cross-validation with k folds, the original dataset is randomly subdivided into "k" mutually exclusive subgroups or "folds" (F_1, F_2, $\ldots F_k$) of roughly equal size. There are k training and testing iterations. In iteration "i" the test set is partition F_i, while the remaining segments, subgroups, or folds are used to train the model collectively [29]. Table 7 and Figure 8 show the 10-fold cross validation accuracy of the three MLAs. In 10-fold cross validation, the CatBoost ML algorithm achieves the highest overall mean accuracy of 93.8 percent, with a range of 92.05 percent to 95.35 percent.

4. Conclusion

Classification of dry bean seed varieties is critical for seed uniformity and quality assurance. Compared to human inspectors, the system possessed two significant advantages. It produces higher, reproducible, and objective sample classification, and also excludes the possibility of human inspectors misclassifying specimens. Initially, the dry bean dataset features has been applied with log transformation. It fails with a reduction in negative skewness. The BCT was applied to all of the features of the dataset for transforming the skewed data into a normal distribution. A model constructed using a single method may not provide the best forecast for a given data set. Each machine learning technique has its own set of restrictions, making it challenging to create a model with substantial accuracy. The accuracy of various MLAs on a balanced dataset was determined using the 22 MLAs. It supports us in developing a more accurate predictive model. The accuracy of the model has been improved by about 1.49 percent using the balance dataset and the CatBoost ML algorithm. The developed models' high success rates across all metrics indicate that they are effective at classification. The overall system mean accuracy of a balanced dataset is obtained as 93.8 percent for the CatBoost ML model. The results indicate that the proposed CatBoost ML classifier can be used effectively to classify a variety of dry bean variants. Additionally, this developed framework can be applied to various kinds of dry beans from various regions. The model is developed without losing any features from the dataset. The ML model can be upgraded further by combining ML, deep learning, and novel algorithms.

Data Availability

The dataset are available in a publicly accessible database.

Conflicts of Interest

The authors declare that there are no conflicts of interest.

Supplementary Materials

Dry Bean Data Set. (*Supplementary Materials*)

References

[1] S. Mamidi, M. Rossi, D. Annam et al., "Investigation of the domestication of common bean (Phaseolus vulgaris) using multilocus sequence data," *Functional Plant Biology*, vol. 38, no. 12, pp. 953–967, 2011.

[2] C. Larochelle, E. Katungi, and Z. Cheng, "Household consumption and demand for bean in Uganda," *Determinants and implications for nutrition security*, vol. 23, 2016.

[3] F. Maalouf, S. Ahmed, and Z. Bishaw, "Chapter 6 - faba bean," in *The Beans and the Peas*, A. Pratap and S. Gupta, Eds., Wood head Publishing, Sawston, UK, 2021.

[4] A. A. S. Palilo, B. A. Majaja, and B. Kichonge, "Physical and mechanical properties of selected common beans (Phaseolus vulgaris L.) cultivated in Tanzania," *Journal of Engineering*, vol. 2018, Article ID 8134975, 9 pages, 2018.

[5] P. Gepts and D. Debouck, "Origin, domestication, and evolution of the common bean (Phaseolus vulgaris L.)," *Common beans: Research for Crop Improvement*, vol. 7, p. 53, 1991.

[6] J. Lukinac, K. Mastanjević, K. Mastanjević, G. Nakov, and M. Jukić, "Computer vision method in beer quality evaluation—a review," *Beverages*, vol. 5, no. 2, p. 38, 2019.

[7] F. J. Rodríguez-Pulido, A. B. Mora-Garrido, M. L. González-Miret, and F. J. Heredia, "Research progress in imaging technology for assessing quality in wine grapes and seeds," *Foods*, vol. 11, no. 3, p. 254, 2022.

[8] M. J. Cejudo-Bastante, F. J. Rodríguez-Pulido, F. J. Heredia, and M. L. González-Miret, "Assessment of sensory and texture profiles of grape seeds at real maturity stages using image analysis," *Foods*, vol. 10, no. 5, p. 1098, 2021.

[9] W. H. Su and H. Xue, "Imaging spectroscopy and machine learning for intelligent determination of potato and sweet potato quality," *Foods*, vol. 10, no. 9, p. 2146, 2021.

[10] B. Krawczyk, "Learning from imbalanced data: open challenges and future directions," *Prog Artif Intell*, vol. 5, no. 4, pp. 221–232, 2016.

[11] M. Koklu and I. A. Ozkan, "Multiclass classification of dry beans using computer vision and machine learning techniques," *Computers and Electronics in Agriculture*, vol. 174, Article ID 105507, 2020.

[12] K. Kilic, I. H. Boyaci, H. Köksel, and I. Küsmenoĝlu, "A classification system for beans using computer vision system and artificial neural networks," *Journal of Food Engineering*, vol. 78, no. 3, pp. 897–904, 2007.

[13] J. Sun, S. Jiang, H. Mao, X. Wu, and Q. Li, "Classification of black beans using visible and near infrared hyperspectral imaging," *International Journal of Food Properties*, vol. 19, no. 8, pp. 1687–1695, 2016.

[14] M. M. Hasan, M. U. Islam, and M. J. Sadeq, "A deep neural network for multi-class dry beans classification," in *Proceedings of the 2021 24th International Conference on Computer and Information Technology (ICCIT)*, pp. 1–5, Tabuk City, Saudi Arabia, December 2021.

[15] A. Aboukarima, M. El-Marazky, H. Elsoury, M. Zayed, and M. Minyawi, "Artificial neural network-based method to identify five varieties of Egyptian faba bean according to seed morphological features," *Engenharia Agrícola*, vol. 40, no. 6, pp. 791–799, 2020.

[16] S. A. D. Araújo, J. H. Pessota, and H. Y. Kim, "Beans quality inspection using correlation-based granulometry," *Engineering Applications of Artificial Intelligence*, vol. 40, pp. 84–94, 2015.

[17] E. M. De Oliveira, D. S. Leme, B. H. G. Barbosa, M. P. Rodarte, and R. G. F. A. Pereira, "A computer vision system for coffee

beans classification based on computational intelligence techniques," *Journal of Food Engineering*, vol. 171, pp. 22–27, 2016.

[18] H. L. Gope and H. Fukai, "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*, vol. 7, no. 1, pp. 149–167, 2022.

[19] E. R. Arboleda, A. C. Fajardo, and R. P. Medina, "Classification of coffee bean species using image processing, artificial neural network and K nearest neighbors," in *Proceedings of the 2018 IEEE International Conference on Innovative Research and Development (ICIRD)*, pp. 1–5, Bangkok, Thailand, May 2018.

[20] R. Duwairi and M. El-Orfali, "A study of the effects of preprocessing strategies on sentiment analysis for Arabic text," *Journal of Information Science*, vol. 40, no. 4, pp. 501–513, 2014.

[21] K. Adnan and R. Akbar, "Limitations of information extraction methods and techniques for heterogeneous unstructured big data," *International Journal of Engineering Business Management*, vol. 11, Article ID 184797901989077, 2019.

[22] G. Wu and E. Y. Chang, "KBA: kernel boundary alignment considering imbalanced data distribution," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 6, pp. 786–795, 2005.

[23] X.-Y. Liu, J. Wu, and Z.-H. Zhou, "Exploratory undersampling for class-imbalance learning," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 39, no. 2, pp. 539–550, 2009.

[24] S. Marimuthu, T. Mani, T. D. Sudarsanam, S. George, and L. Jeyaseelan, "Preferring Box-Cox transformation, instead of log transformation to convert skewed distribution of outcomes to normal in medical research," *Clinical Epidemiology and Global Health*, vol. 15, Article ID 101043, 2022.

[25] V. R. Joseph and A. Vakayil, "SPlit: an optimal method for data splitting," *Technometrics*, vol. 64, no. 2, pp. 166–176, 2021.

[26] H. Jafarzadeh, M. Mahdianpari, E. Gill, F. Mohammadimanesh, and S. Homayouni, "Bagging and boosting ensemble classifiers for classification of multispectral, hyperspectral and PolSAR data: a comparative evaluation," *Remote Sensing*, vol. 13, no. 21, p. 4405, 2021.

[27] S. Holm, "Generalized linear models for ordered categorical data," *Communications in Statistics - Theory and Methods*, vol. 52, no. 3, pp. 670–683, 2021.

[28] S. K. Singh, R. W. Taylor, B. Pradhan, A. Shirzadi, and B. T. Pham, "Predicting sustainable arsenic mitigation using machine learning techniques," *Ecotoxicology and Environmental Safety*, vol. 232, Article ID 113271, 2022.

[29] K. Karthick, S. K. Aruna, R. Samikannu, R. Kuppusamy, Y. Teekaraman, and A. R. Thelkar, "Implementation of a heart disease risk prediction model using machine learning," *Computational and Mathematical Methods in Medicine*, vol. 2022, Article ID 6517716, 14 pages, 2022.

[30] I. Jenhani, N. B. Amor, and Z. Elouedi, "Decision trees as possibilistic classifiers," *International Journal of Approximate Reasoning*, vol. 48, no. 3, pp. 784–807, 2008.

[31] P. Boedeker and N. T. Kearns, "Linear discriminant analysis for prediction of group membership: a user-friendly primer," *Advances in Methods and Practices in Psychological Science*, vol. 12, pp. 250–263, 2019.

[32] K. Kanagarathinam, D. Sankaran, and R. Manikandan, "Machine learning-based risk prediction model for cardiovascular disease using a hybrid dataset," *Data and Knowledge Engineering*, vol. 140, Article ID 102042, 2022.

[33] A. Ibrahem Ahmed Osman, A. Najah Ahmed, M. F. Chow, Y. Feng Huang, and A. El-Shafie, "Extreme gradient boosting (Xgboost) model to predict the groundwater levels in Selangor Malaysia," *Ain Shams Engineering Journal*, vol. 12, no. 2, pp. 1545–1556, 2021.

[34] J. T. Hancock and T. M. Khoshgoftaar, "CatBoost for big data: an interdisciplinary review," *J Big Data*, vol. 7, no. 1, p. 94, 2020.