Journal of Engineering Science and Technology Review 16 (2) (2023) 107 - 115

Research Article

JOURNAL OF
Engineering Science and Technology Review

www.jestr.org

An Advance Boosting Approach for Multiclass Dry Bean Classification

Janmenjoy Nayak[1,*], Pandit Byomokesha Dash[2] and Bighnaraj Naik[3]

[1]*Dept. of Computer Science, Maharaja Sriram Chandra Bhanja Deo University, Baripada-757003, Odisha, India*
[2]*Dept. of Information Technology, Aditya Institute of Technology and Management, Tekkali, AP-532201*
[3]*Dept. of Computer Application, Veer Surendra Sai University of Technology, Burla, India-768018.*

Received 17 November 2022; Accepted 10 April 2023

Abstract

Dry beans, which are produced in large quantities, have the highest level of genetic diversity. The quality of seeds has a significant impact on crop yield. The importance of seed classification to both marketing and production can be shown by realizing that sustainable agricultural systems depend on these principles. This research is primarily aimed at providing a means to generate uniform seed varieties, as seed is not certified as a single variety. To achieve consistent seed classification, we have proposed Extreme Gradient Boosting ensembles using the Synthetic Minority Over-Sampling Methodology (SMOTE) to differentiate seven distinct registered types of dry beans with similar characteristics. There were a total of 13,611 grains from seven different varieties of dry beans sampled for the classification model. Classification algorithms based on machine learning like Decision Tree (DT), Support Vector Machine (SVM), Multilayer Perceptron (MLP), Adaptive Boosting classifier, Bagging Classifier, and Extreme Gradient Boosting ensembles using the Synthetic Minority Over-Sampling Methodology (SMOTE) were developed and compared. Overall correct classification rates for SVM, MLP, DT, Adaboost, Bagging, and Extreme Gradient Boost classifiers were 94.44%, 94.48%, 96.53%, 96.35%, 96.89%, and 97.32%, respectively. Extreme Gradient Boosting ensembles using the SMOTE classification model have the best accuracy. The results of this study satisfy the producers' and customers' demand for uniform bean varieties.

Keywords: Machine Learning (ML), SMOTE, Dry-beans, Extreme Gradient Boosting, Uniform seed classification.

1. Introduction

As humans travelled around the world throughout history, they brought seeds of food crops with them to support them and grow in new locations. Humans rely on grains for both dietary and economic reasons; they provide a wide range of benefits including protein and energy [1]. In many developing countries, beans are a crucial food staple because they provide a high concentration of plant-based proteins, including 20–25 percent protein and 56–56 percent carbs [2]. Due to the plant's sensitivity to climatic changes, these fluctuations will have a large impact on the plant. The utilization of novel seed cultivars and their seed properties is one of the most essential aspects in the plant's success [3].

Seed verification is a major issue for dry bean farmers and markets because the quality of the seed is crucial. Producing crops with lesser-quality seeds will always result in lower yields, no matter what else is done in the field to improve things. According to farmers, seed is the most critical input when it comes to the total cost of agriculture. Beans are among the primary contributors to the overall cost of food production. While worldwide, there is substantial genetic diversity, the most common origin of dry beans is from the United States [4]. Some definitions, according to the Turkish Standards Institution, include terms such as " Dermason, Bombay, Cali, Horoz, Tombul, Selanik, Barbunya, Battal, and Seker", which are referred to as "barley, lentils, mung beans, peas, rice, lentils, black eyed peas, pea pods, adzuki beans, soybeans, and green beans" (Turkish Standards Institution, 2009) [5].

Crop production relies on one of the most basic processes of genotyping and analysis: the genotyping and analysis of dry bean varieties, which are common all over the world [6]. When information on the appearance, size, colour, health, and variety of products made available to consumers, this leads to an increase in market value. Additionally, it aids farmers in identifying bean varieties and standard seed use. In terms of yield and disease resistance, maintaining high seed quality is critical for successful bean cultivation. It is necessary to go through a lengthy process that involves manually sorting and classifying bean seeds. This method wastes a lot of time and resources. When producing in larger quantities, the amount of time and resources wasted is greatly increased. Automatic methods are required to adequately grade and classify.

In the agricultural sector, the characteristics of seeds have a significant impact on crop productivity. A range of computational equipment is available for monitoring the quality of agricultural and food products. Yet, the majority of them are accomplished using conventional methods. For instance, seed categorization is typically relied on human comprehension, and defining the type of dry beans manually needs a skilled individual and a substantial amount of time. Since the variety of seeds appears so identical, classifying them manually becomes difficult. It is nearly impossible for a human operator to interpret or manage such seeds without the use of specialized equipment or automated software techniques.

The goal of seed classification is to provide increased yields of high-quality food. The dry bean is the most nutritionally and frequently produced vegetable worldwide.

*E-mail address: jnayak@ieee.org
ISSN: 1791-2377
doi:10.25103/jestr.162.14

The purification of dry beans is vital to the economies of agriculture-based nations such as Bangladesh, India, and Pakistan, among others. Due to the effects of a changing climate and other environmental conditions, seed quality might deteriorate at any time during the plant's development, beginning with fertilization. Breeding new seed cultivars and identifying their characteristics, which are the significant variables for good plant growth, can enhance the reactivity and/or tolerance of plants to environmental stimuli. The seed identification process is time-consuming and open to multiple interpretations. In terms of business and technical issues, the situation gets more difficult from a practical standpoint. Specifically, different species of dry beans tend to vary in colour, and the geometric data contain no information regarding bean colour. Therefore, it is not only economically but also technically essential to develop an automated method for detecting and classifying seed characteristics fast and repeatedly. With the aid of machine learning, inspections of the quality of seeds, fruits, and vegetables, as well as examination and categorization of seeds and grains, are undertaken globally to satisfy these demands. The goal of seed classification is to provide increased yields of high-quality food.

To conduct this research, we relied on seven common dry bean kinds gathered by the Turkish Standards Institute. Some samples of seven species are, however, much lower in count, resulting in a natural imbalance between the seven species. This imbalance is challenging for machine learning models, as the majority of the learning models assume equal amounts of samples or of class. Because of this, the ability to identify a single variety of bean is hampered by poor predictive performance. This research is designed to develop a process for the acquisition of uniform seed varieties from crops. It will find new products which are available on the market and will help you identify products that maintain the market at the desired price. The main goal of this study is to design a powerful learning engine that takes advantage of Extreme Gradient Boosting ensembles using the SMOTE for classifying simple dry-bean types described by the Institution of Turkish Standards. The remaining sections are broken down as follows: Section 2 describe the literature study of various machine learning and deep learning-based uniform seed classification methods, as well as their drawbacks. The proposed methodology is described in Section 3. The details of the experiment and results analysis are discussed in Section 4. Section 5 wraps up the proposed work with some recommendations for the future.

2. Literature Study

To improve yield, beans grown in Turkey are divided into varieties according to the characteristics of their form, shape, type, and structure, as well as market demand. It's these seven bean varieties that are most well-known among customers: Dermason, Horoz, Seker, Barbunya, Bombay, Cali, and Sira [7]. The technology for classifying bean seed species was first developed some years ago, but ML and artificial intelligence are widely used in research investigation to identify dry bean seed species. For the quality management of the beans, Kılıc et al. [8] have created a computer vision system (CVS) which takes account of the samples' dimensions and colour quantities. Artificial neural network (ANN) utilised to determine the colour of beans. According to the criteria established by the system and the experts, the samples were classified into five groups. 371 samples were tested for ANN. There is a total 90.6 % accuracy in the system categorization. To classify wheat grains, Sabanci et al. [9] utilized an ANN classifier. After examining 21 bread variants and 100 durum wheat varieties, they found 21 distinct characteristics in a total of 100 unique products. The Holdout approach has been used to distribute 90% of training data and 10% of test data for the purpose of distinguishing the two varieties of wheat. Their class accuracy rate found 92.92% after using an ANN.

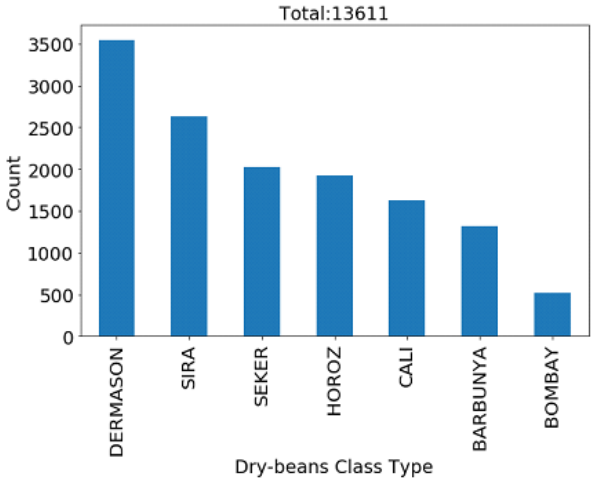
The same bean variety is used in the research by Araujo et al. [10], when using the multivariate granulometry approach based on correlation. The computer system that was created in this study is aimed at aiding the visual inspection of beans, primarily for the purpose of differentiating various shades of the bean colour. The three modules comprising this system were grain sorting, pixel colour mapping and grain partitioning. Employing the k-Nearest Neighbor classification methodology, they were able to get a 99.88% accuracy. For the classification of six Italian landraces of beans, Venora et. al. [11] proposed a linear discriminant analysis (LDA) method utilizing KS-400, a commercial image analysis library. In the experiments, features like grain size, shape, colour, and texture were all used and the results were impressive, with a success rate of 99.56%. In their follow-up study Venora et al. [12] conducted further trials on fifteen Italian traditional landraces of beans, with a 98.49% success rate.

A CVS based on artificial intelligence have been proposed by Koklu et al. [13] for the Turkish Standards Institutes to classify basic dry bean varieties with anatomically identical features but no distinguishing colour. The model classification has been compared to 10-fold cross validation of several machine learning algorithms like kNN, SVM, MLP and DT. Overall, the correct classification rates for DT, SVM, kNN, and MLP were 92.52%, 93.13%, 87.92%, and 91.73% respectively. Various species of seeds will be contained in the final products as different populations with various genotypes are cultivated. Dry bean seeds derived from large population agriculture cannot be segregated on a species basis since they are not isolated from each other, their market value is drastically decreasing [14]. The goal is to eliminate the disadvantages of population cultivation and obtain a uniform type of bean for producers. A study of retail goods on the market reveals that the evaluation and pricing of those items is the work of experts. The process has an inherent human element, and so it is error-prone. In the last decade, there has been an uptick in the use of machine learning and artificial intelligence (AI) techniques to assist in solving problems related to forecasting and classification.

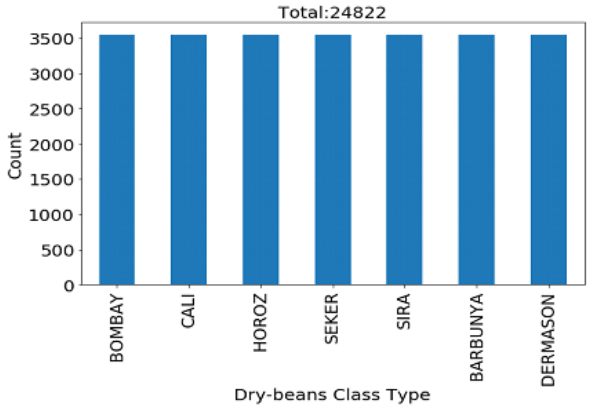
3. Proposed Methodology

Proposals for this research work are broken down into two parts: (a) using SMOTE to obtain the balanced structure of dry bean varieties from original imbalanced dry bean samples, and (b) designing Extreme Gradient Boosting ensembles for the classification of basic dry bean types. In this work, we have used dry beans dataset from Turkish Standards Institute for the model evaluation. There are 13611 samples in all, each with a different set of 16 characteristics. This dataset covers seven types of dry beans those are “Cali, Dermosan, Horoz, Seker, Barbunya, Bombay and Sira”. The details percentage of distribution of dry beans type “Cali, Dermosan, Horoz, Seker, Barbunya, Bombay and Sira” having the distribution 11.97%, 26.05%,16.16%,14.89%,5.71%,3.83% and 16.89% respectively. When using machine learning algorithms, Predictive accuracy of a model is significantly impacted by

how the data is organized. As a result of data imbalance, we used SMOTE-oversampling to compensate for the fact that the underrepresented classes had fewer votes than the majority class. As seen in Figure 1 A, B, the original data sample distribution is shown as distribution (A) while the sample profile of transformed dry beans is shown as (B).



(A)



(B)

Fig. 1. (A) original dataset_ Class distribution (B) SMOTE_Oversampled dataset_ Class distribution.

There is a total of seven recognized dry bean varieties, and the dataset used to create the model has 13,611 grains representing all of them and their unique characteristics. As this an imbalance dataset, we have used SMOTE method to oversampling the dataset in order to have a properly balanced dataset. The employed SMOTE method has following major steps: i) Initializing the amount of oversampling in percentage, ii) Iteratively selection of i^{th}minority sample and selections of its 'k' no. of nearest neighbors, iii) Saving the indices of the k no. of nearest neighbors and generations of required amount samples (based on amount of oversampling in percentage) by using nearest neighbors.

Algorithm for XGBoost classification model

Initialize $f_0(x)$;
For K = 1,2,3…….,M do
Compute $g_k = \frac{\partial L(y,f)}{\partial f}$;
Compute $h_k = \frac{\partial^2 L(y,f)}{\partial f^2}$;
Establish the structure by choosing splits with maximize gain
$A = \frac{1}{2}[\frac{G_L^2}{H_L} + \frac{G_R^2}{H_R} - \frac{G^2}{H}]$;
Decide the leaf weights w* = $-\frac{G}{H}$;
Decide base learner $\hat{b}(x) = \sum_{i=1}^{T} wI$;
Add trees = $f_k(x) = f_{k-1}(x) + \hat{b}(x)$;
End

Output: $f(x) = \sum_{k=0}^{M} f_k(x)$

Where,
M = number of base learners
W* = optimized leaf weight
G and H are 1st and 2nd Gradient respectively
For the left and right branches of a tree, the subscripts L and R have been used.

Assume that the obtained balanced dataset after applying SMOTE is $D = \{(X_i, y_i)\}_{i=1}^{n}$, which is consisting of with 'n' number of instances with 'm' number of features. Initially, a tree ensemble model is created in additive manner that uses 'K' number of functions to predict output as shown in Eq.1.

$$\hat{y}_i = P(X_i) = \sum_{k=1}^{K} f_k(X_i) \quad (1)$$

In Eq. (1), $\hat{y}_i$ is the output prediction ofX_i , $P(X_i)$is the tree ensemble model's prediction onX_i, $f_k$is one of the tree with structure $q$and weight q, $F$is the tree space (Figure 2). The collection of functions f_k used in the model is learnt by decreasing the regularized objective in Eq. (2).

$$L(P) = \sum_i l\left(\hat{y}_i, y_i\right) + \sum_k \Omega(f_k) \quad (2)$$

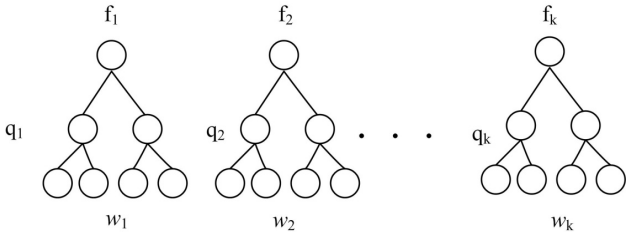


Fig. 2. Tree space.

In Eq. (2), $l(\ )$ is the differentiable loss function and $\Omega(f) = \gamma T + \frac{1}{2}\lambda\|w\|^2$ is the term that penalizes for complexity of model. The tree ensemble model has been trained in an additive manner as the objective function in Eq. (2) consists of functions as parameters and cannot be optimized using conventional approaches. This is achieved by adding f_t greedily Eq. (3) that certainly improve the model. This can be quickly optimized by having second order approximation as presented in Eq. (4).

$$L^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)}\right) + f_t(X_i) + \Omega(f_k) \quad (3)$$

$$L^{(t)} = \sum_{i=1}^{n} [g_i \times f_t(X_i) + 0.5 \times h_i \times {f_t}^2(X_i)] + \Omega(f_k) \quad (4)$$

In Eq. (4),

$$g_i = \partial_{\hat{y}^{(t-1)}} l\left(y_i, \hat{y}_i^{(t-1)}\right)$$

and

$$h_i = \partial^2_{\hat{y}^{(t-1)}} l\left(y_i, \hat{y}_i^{(t-1)}\right)$$

are 1st and second gradient respectively.

Letting$I_j = \{i|q(X_i) = j\}$as the sample set of leaf j, the Eq. (5) can be rewritten as Eq. (5) by expanding it.

$$L^{(t)} = \sum_{i=1}^{n}\left[g_i \times f_t(X_i) + 0.5 \times h_i \times {f_t}^2(X_i)\right] + \gamma T + 0.5 \times \lambda \sum_{j=1}^{T} w_j^2 = \sum_{j=1}^{T}\left[\left(\sum_{i\in I_j} g_i\right) \times w_j + 0.5 \times \left(\sum_{i\in I_j} h_i + \lambda\right) \times w_j^2\right] + \gamma T \quad (5)$$

The optimal weight 'w_j' of the 'j'leaf for a constant tree structure is computed as in Eq. (6) and the corresponding optimal value has been obtained Eq. (7). The scoring function (impurity score) has been used 't' evaluate the quality of the structure 'q'.

$$w_j^* = -\frac{\sum_{i\in I_j} g_i}{\sum_{i\in I_j} h_i + \lambda}, \quad (6)$$

$$L^t(q) = 0.5 \times \sum_{j=1}^{T} \frac{\left(\sum_{i\in I_j} {g_i}^2\right)}{\left(\sum_{i\in I_j} h_i + \lambda\right)} + \gamma T \quad (7)$$

As it is not possible to evaluate all the possible tress structure 'q', a greedy procedure is employed to derive optimal tree structure in additive and iterative manner. Assuming$I = IL \cup IR$, where $IL$and IR are the set of instances of left and right nodes, the reduction of loss after the split is computed as in Eq. (8).

$$L_{split} = 0.5 \times \left(\frac{\left(\sum_{i\in I_L} g_i\right)^2}{\left(\sum_{i\in I_L} h_i + \lambda\right)} + \frac{\left(\sum_{i\in I_R} g_i\right)^2}{\left(\sum_{i\in I_R} h_i + \lambda\right)} - \frac{(\sum_{i\in I} g_i)^2}{(\sum_{i\in I} h_i + \lambda)}\right) - \gamma \quad (8)$$

The final tree obtained through this iterative and additive procedure is used to differentiate between seven distinct registered types of dry beans. The Dry Beans classification employed in this study is depicted in Figure 3.

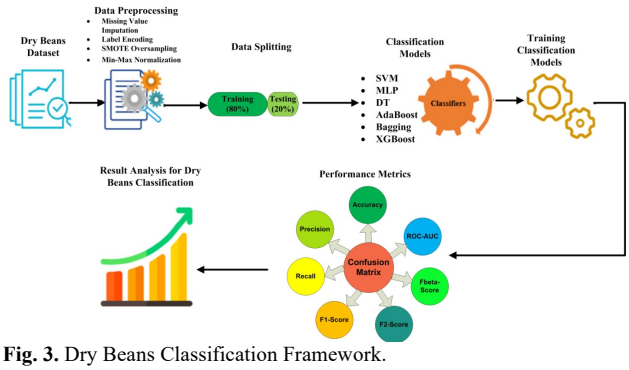


Fig. 3. Dry Beans Classification Framework.

The Dry Beans dataset is loaded at the initial phase. In the second phase, multiple data preprocessing techniques are employed to process the dataset. Using the relevant pandas library framework function, missing values, string values, and NaN values are eliminated. To convert categorical text data into model comprehensible numerical data, we employ the Label Encoder approach. Several columns in the dataset contain a large number of values that are extremely dissimilar to one another, resulting in biased model performance. We confined all features to the interval 0 to 1. Due to the low occurrence of most types of Dry Beans, the distribution of Dry Beans datasets is often unbalanced, with large variances in the number of samples from different categories. When the model is trained with unbalanced datasets, its performance and reliability are degraded. To address this imbalance, we have employed SMOTE, a technique of oversampling in which synthetic samples are generated for the minority group. This approach aids in overcoming the problem of overfitting caused by random oversampling. In the third stage, the Dry Beans dataset is split into a training set of 80% and a test set of 20%. In the fourth stage, the models are instantiated (SVM, MLP, DT, AdaBoost, Bagging and Proposed XGBoost Classifier). Once the models have been initialized, they are trained (using the training set) and evaluated (using the testing set). After the completion of the proposed XGBoost Classifier in conjunction with the other ML models that were compared, the models are trained and tested. The performance of classifying the various kinds of beans is evaluated by the module for Dry Beans classification. The proposed model performed extremely well in comparison to other approaches.

3. Proposed System

This section presents an overview of the experimental setup and the results of the experiment, along with explanations of the results.

3.1. Simulation Setup

Spyder, the Scientific Python Development Environment, which is based on Python 3.9.6 and includes a wide variety of Python packages like the Pandas framework, the Imblearn framework, and the Numpy framework, is used to create and test all of the models that are analysed and proposed. The Matplotlib and Mlxtend frameworks have been used for data visualization, while the sklearn framework and classification-metrics frameworks have been utilized for data analysis, using a machine having an Intel Core i5-6700 processor at 3.40 GHz (8 CPUs), 4 GB of RAM, and Windows 8.1 Pro on a 64-bit architecture. All parameters of the classifiers are specified by picking appropriate test and error values shown in Table 1.

Table 1. Parameters of all considered models.

Models	Parameters
SVM	degree=3, gamma='scale' ,shrinking=True, tol=0.001, decision_function_shape='ovr', C=1.0, kernel='rbf'
MLP	activation='relu', hidden_layer_sizes=100, solver='adam', alpha=0.0001, learning_rate_init=0.001
DT	max,criterion='gini', splitter='best', min_samples_split=2, min_samples_leaf=1
Adaboost	Base_estimator=DecisionTreeClassifier(),n_ estimators=100,learning_rate=1.0, algorithm='SAMME.R'
Bagging	n_estimators=100, base_estimator=DecisionTreeClassifier()
XGBoost	n_estimators=100

3.2. Description of Dataset

The Turkish Standards Institute determined the shape, form, kind, and layout of the beans, as well as the present market circumstances, in this research, which employed seven distinct varieties of dry beans. Cali, Dermosan, Horoz, Seker, Barbunya, Bombay and Sira are some of the names that have been given to them. In 2009, the Turkish Standards Institution established the following general qualities for dry beans.

3.3. Findings and Discussion

Several performance metrics, including Accuracy Eq. (9), Precision Eq. (10), Recall Eq. (11), F1-score Eq. (12), F2-score Eq. (13), F-beta score Eq. (14) and AUC-ROC, have been used to evaluate the performance of the suggested model. The mathematical formulas used to express the aforementioned performance metrics are provided below.

1. Accuracy: Accuracy refers to the proportion of all sample points for which predictions are accurate.

Accuracy = (TP + TN) / (TP + FP + TN + FN) * 100 (9)

2. Precision: It quantifies the proportion of true positive cases to all Dry Bean types.

Precision=TP / (TP + FN) *100 (10)

3. Recall: It is the ratio of true negative cases to all Dry Beans types.

Recall=TN / (TN + FP) *100 (11)

4. F1-Score: F1-score is the harmonic mean of the model's precision and recall.

F1-Score=2 * (Precision * Recall) / (Precision + Recall) (12)

5. F2-Score: The F2 score represents the weighted harmonic mean of precision and recall (given a threshold value). Contrary to the F1 score, which assigns equal weight to precision and recall, the F2 score assigns greater weight to recall than to precision.

F2-Score= 5 * (Precision * Recall) / (4 * Precision + Recall) (13)

6. F-beta Score: The F-beta score is the weighted harmonic mean of precision and recall, with its best value being 1 and its worst value being 0.

F-beta Score= (1 + beta**2) * (Precision * Recall) / (beta**2 * Precision + Recall) (14)

Where, TP and TN are the true positive and true negative predictions of the Dry Beans Classification model. FP and FN represent the model's false positive and false negative predictions.

Various ML and Ensemble-based approaches are investigated for performance evaluation and comparison in this study. This section provides extensive information regarding the analysis of the results obtained. This study explored a variety of machine learning techniques (SVM, MLP, and DT) as well as ensemble methods (AdaBoost, Bagging and XGBoost). The main challenge of experimenting is applying various techniques to an imbalanced data sample. When applied to the same imbalance dataset, the various techniques give different performance outcomes. Due to the imbalanced composition of the Dry Beans dataset, the design of machine learning models is a challenge, as the majority of machine learning models assume an equal number of samples for each class, resulting inadequate predictive accuracy for identifying Dry Bean type. To deal with this issue, this study provides all investigated ML approaches with SMOTE for predicting Dry Bean type. All SMOTE-enabled models have a higher overall correct classification rate and F1 score than non-SMOTE models. The XGBoost ensemble classification model with SMOTE has the highest accuracy (97.32%), F1-score (0.9732), F2-score (0.9731), F-beta score (0.9731), Precision (0.9732), recall (0.9732), and ROC-AUC (0.9841) among all models. The results demonstrate that the proposed method outperforms conventional ML and ensemble techniques in terms of precision, recall, F1 score, F2 score, Fbeta score, and AUC-ROC. The results emphasize the significance of the suggested model in addressing the Dry Beans Classification problem. To assess the effectiveness of the proposed method, various performance metrics such as accuracy, precision, recall, F1-score, F2-score, F-beta-score and receiver operating characteristic (ROC)-area under the ROC curve (AUC) curve have been shown in Table 2.

Table 2. Performance Comparison among all the models.

Prediction Models	Performance Metrics						
	Precision	Recall	F1 Score	F2 Score	Fbeta Score	ROC-AUC	Accuracy
SVM	0.9312	0.9312	0.9312	0.9309	0.9320	0.9646	93.20
MLP	0.9191	0.9191	0.9191	0.9166	0.9210	0.9550	92.02
DT	0.8964	0.8964	0.8964	0.8962	0.8963	0.9464	89.64
Adaboost	0.9004	0.9004	0.9004	0.9003	0.9005	0.9482	90.04
Bagging	0.9269	0.9269	0.9269	0.9266	0.9270	0.9620	92.69
XGBoost	0.9309	0.9309	0.9309	0.9309	0.9313	0.9636	93.09
SVM_SMOTE	0.9444	0.9444	0.9444	0.9443	0.9442	0.9672	94.44
MLP_SMOTE	0.9448	0.9448	0.9448	0.9443	0.9451	0.9675	94.48
DT_SMOTE	0.9653	0.9653	0.9653	0.9653	0.9652	0.9795	96.53
Adaboost_SMOTE	0.9635	0.9635	0.9635	0.9635	0.9635	0.9784	96.35
Bagging_SMOTE	0.9689	0.9689	0.9689	0.9689	0.9689	0.9817	96.89
Proposed XGBoost_SMOTE	0.9732	0.9732	0.9732	0.9731	0.9731	0.9841	97.32

Figure 4 (A) to (F) depicts the ROC-AUC curves of SVM, MLP, DT, Adaboost, Bagging, and XGBoost without SMOTE. The ROC AUC curves of SVM, MLP, DT, Adaboost, Bagging, and XGBoost with SMOTE are shown in Figure 5 (A) to (F). Table 2 shows the comparisons of all of the models considered based on various performance metrics. When performance measures are compared, it is observed that the suggested strategy outperforms previous models in terms of F1-score, sensitivity and ROC-AUC, indicating productivity and effectiveness in recognizing uniform seed types as well as the capacity to handle unbalanced dry bean datasets.

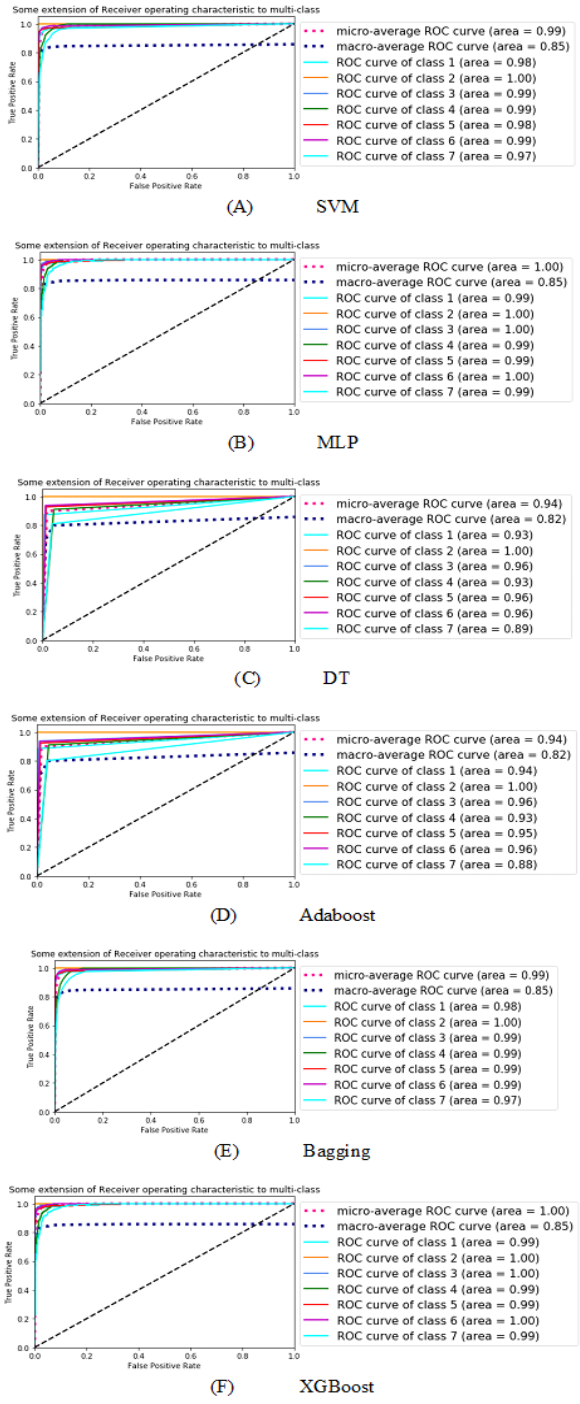


Fig. 4. ROC-AUC of (A) SVM (B) MLP (C) DT (D) Adaboost (E) Bagging (F) XGBoost.

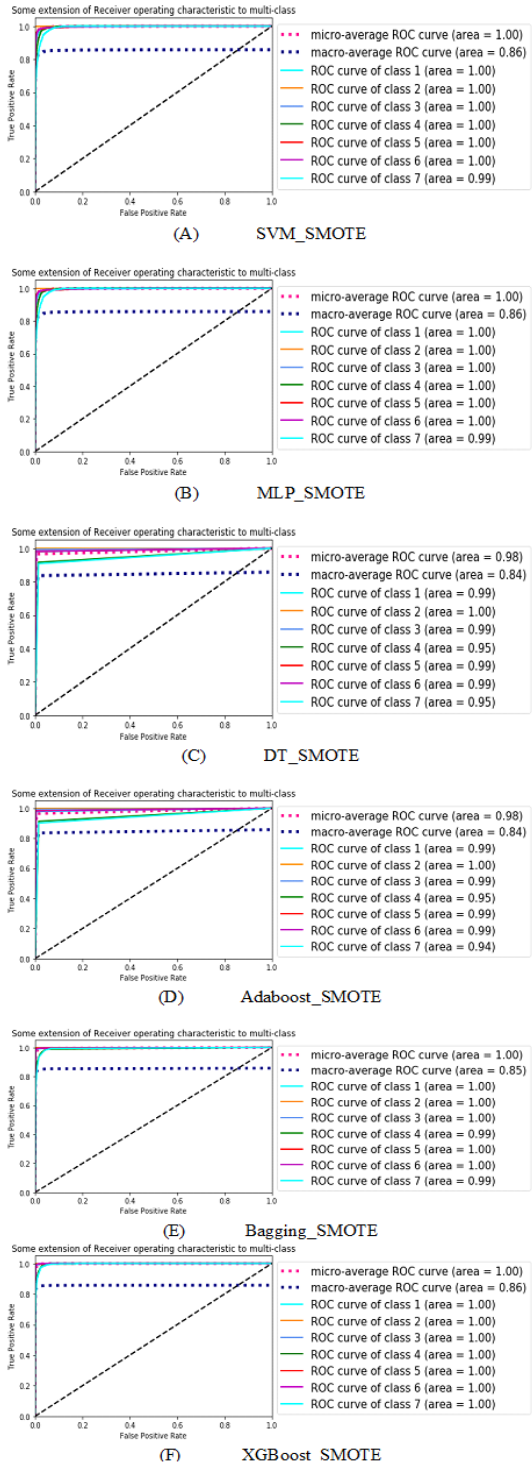


Fig. 5. ROC-AUC of (A) SVM_SMOTE (B) MLP_SMOTE (C) DT_SMOTE (D) Adaboost_SMOTE (E) Bagging_SMOTE (F) XGBoost_SMOTE.

Tables 3–8 show the confusion matrices for the SVM_SMOTE, MLP_SMOTE, DT_SMOTE, Adaboost_SMOTE, Bagging_SMOTE, and XGBoost_SMOTE models developed. Figure 6 depicts the classification accuracy performance comparison for dry bean varieties. The accuracy of the classifiers' classifications can be seen from the chart below, which clearly shows that the proposed classifiers' performance is superior to the other standard classifiers. The analysis of precision, recall and accuracy of different classification models have been depicted

in Figure 7. Table 9 shows the comparison of current research work with previous study.

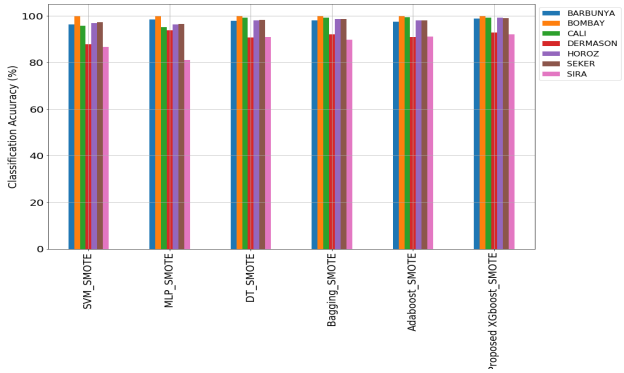


Fig. 6. For all bean varieties, the accuracy of classification models.

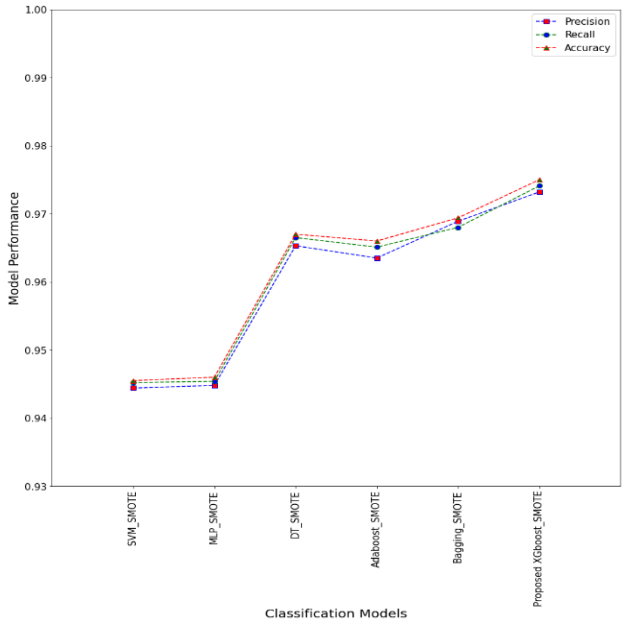


Fig. 7. Analysis of Precision, Recall and Accuracy of classification models.

Table 3. Confusion Matrix of SVM_SMOTE.

	BARBUNYA	BOMBAY	CALI	DERMASON	HOROZ	SEKER	SIRA
BARBUNYA	687	0	18	0	1	4	3
BOMBAY	1	758	0	0	0	0	0
CALI	21	0	678	0	5	1	3
DERMASON	0	0	0	603	2	20	62
HOROZ	1	0	5	3	674	0	12
SEKER	1	0	2	6	0	677	10
SIRA	5	0	1	60	15	14	612

Table 4. Confusion Matrix of MLP_SMOTE.

	BARBUNYA	BOMBAY	CALI	DERMASON	HOROZ	SEKER	SIRA
BARBUNYA	702	0	7	0	0	4	0
BOMBAY	1	758	0	0	0	0	0
CALI	30	0	673	0	4	1	0
DERMASON	0	0	0	645	0	12	30
HOROZ	9	0	14	2	669	0	1
SEKER	8	0	2	14	0	671	1
SIRA	7	0	2	88	22	15	573

Table 5. Confusion Matrix of DT_SMOTE.

	BARBUNYA	BOMBAY	CALI	DERMASON	HOROZ	SEKER	SIRA
BARBUNYA	697	0	11	0	0	4	1
BOMBAY	0	759	0	0	0	0	0
CALI	2	0	702	0	2	2	0
DERMASON	1	0	0	623	3	9	51
HOROZ	2	0	6	0	682	0	5
SEKER	4	0	0	2	0	684	6
SIRA	2	0	0	53	5	4	643

Table 6. Confusion Matrix of Adaboost_SMOTE.

	BARBUNYA	BOMBAY	CALI	DERMASON	HOROZ	SEKER	SIRA
BARBUNYA	695	0	11	0	0	3	4
BOMBAY	0	759	0	0	0	0	0
CALI	1	0	704	0	1	2	0
DERMASON	1	0	0	624	1	8	53
HOROZ	2	0	7	0	681	0	5
SEKER	3	0	2	2	0	683	6
SIRA	2	0	0	54	3	4	644

Table 7. Confusion Matrix of Bagging_SMOTE.

	BARBUNYA	BOMBAY	CALI	DERMASON	HOROZ	SEKER	SIRA
BARBUNYA	699	0	9	0	0	4	1
BOMBAY	0	759	0	0	0	0	0
CALI	4	0	702	0	1	1	0
DERMASON	0	0	0	632	2	11	42
HOROZ	0	0	2	1	686	0	6
SEKER	4	0	2	1	0	687	2
SIRA	0	0	1	55	8	8	635

Table 8. Confusion Matrix of XGBoost_SMOTE.

	BARBUNYA	BOMBAY	CALI	DERMASON	HOROZ	SEKER	SIRA
BARBUNYA	705	0	4	0	0	4	0
BOMBAY	0	759	0	0	0	0	0
CALI	3	0	702	0	1	2	0
DERMASON	0	0	0	638	2	8	39
HOROZ	1	0	0	1	689	0	4
SEKER	1	0	0	1	0	689	5
SIRA	1	0	1	42	6	6	651

Table 9. Comparison study of the proposed work with other research.

Method Name	Dataset used	No.of Class	Accuracy %	Ref.
ANN	Synthetic	5	90.06	[8]
ANN	Synthetic	2	92.92	[9]
SVM	Cocoa(Synthetic)	2	95.8	[15]
coarse tree algorithm	Cavite Coffee Beans	3	94.1	[16]
MLP,KNN,DT and SVM	Dry beans Dataset (Turkish Standards Institution)	7	91.73, 87.92,92.52, 93.13	[13]
XGBoost_SMOTE	Dry beans Dataset (Turkish Standards Institution	7	97.32	**Proposed method**

5. Conclusion

Dry bean seeds are impossible to identify based on their size or shape, making the process of classifying them complex. Improving the uniformity and quality of bean seeds is of prime importance when it comes to classifying and assigning varieties. It has always been a research challenge to create an intelligent system that has zero tolerance for misclassification. for researchers at all levels. Machine learning is one of the most effective methods for classifying different types of dry beans. Using SMOTE, this study provides an ensemble learning-based approach. Simulation findings show that the suggested strategy has been effective in dealing with the unbalanced data in the dry bean dataset. This proposed method has superior results when compared to other methods by various evaluation metrics. In other words, the results are more accurate, more precise, and the F1 score goes up. There were various evaluation metrics where we discovered that the suggested method having improved performance indexes. This new development structure is capable of being applied to different bean varieties from different regions. Further improvement of the model can be made through the incorporation of hybrid ML methods, deep learning, and advanced algorithms.

5.1 Limitations and Future Scope

Products from the agriculture sector are extremely reliant on the purity of seeds and the fertility of soil. In this investigation, a genetically heterogeneous dataset of dry beans is utilized and Many machine learning approaches are applied in this framework to standardize the classification of dry beans from crop output at a reasonable computing cost while also addressing the challenges posed by intra-class differences in beans. Many researchers have developed new algorithms recently, most of which are best suited to uniform distribution data sets. Multiclass unbalanced data, which includes the skewed data points, is particularly challenging to classify. Using the SMOTE algorithm, these existing disadvantages of classification can be addressed by distributing characteristics evenly across all classes.

The major goals of future work would be (1) expand the size of the dataset and (2) boost performance to achieve near-100% accuracy. In 1st point, the existing dataset is well-suited for machine learning approaches, but to do good deep learning-based categorization, more data of various types should be required. In addition to expanding the dataset, also in the 2nd point, the suggested model's hyper-parameters can be tuned to improve performance. Moreover, models of deep learning can be employed for improved classification of Dry Beans. When a multi classification model has been introduced based on deep learning. Adjusting the hyper-parameters of a deep learning model can significantly reduce the execution time to get successful outcomes.

References

1. Henry, C.J., "Pasting characteristics of starches in flours of chickpea (Cicer Arietinum L.) and faba bean (Vicia Faba L.) as affected by sorting and dehulling practices". *African Journal of Food Science*, 9 (12), 2015, pp.555–559.
2. Zhang, M., "Effects of ultrasound on the structure and physical properties of black bean protein isolates". *Food Research International*, 62(4), 2014, pp.595–601.
3. Onder, M., "The Impacts of Environment on Plant Products". International *Journal of Bioscience, Biochemistry and Bioinformatics*, 2(1), 2012, pp.48-51.
4. Gentry, H. S., "Origin of the Common Bean, Phaseolus vulgaris". *Economic Botany*, 23(1), 1969, pp.55–69.
5. Turkish Standards Institution, 2009. Dry Beans (Kuru Fasulye). *Turkish Stand. Inst. Off. Gaz. Repub*, Turkey.
6. Ceccatto, H.A., "Weed seeds identification by machine vision". *Computers and Electronics in Agriculture*, 33(2), 2002, pp.91-103.
7. Ceyhan, E., "Correlation and path analysis for yield and yield components in common bean genotypes (Phaseolus vulgaris L.)". *Ratarstvo i Povrtarstvo*, 50(2), 2013, pp.14-19.
8. Kusmenoglu, I., "A classification system for beans using computer vision system and artificial neural networks". *Journal of Food Engineering*, 78(3), 2007, pp.897-904.
9. Toktas, A., "Computer vision-based method for classification of wheat grains using artificial neural network". *Journal of the Science of Food and Agriculture*, 97(8), 2017, pp.2588-2593.
10. Kim, H.Y., "Beans quality inspection using correlation-based granulometry". *Engineering Applications of Artificial Intelligence,* 40(C), 2015, pp.84-94.
11. Cremonini, R., "Tuscany beans landraces, on-line identification from seeds inspection by image analysis and Linear Discriminant Analysis". *Agrochimica*, 51(4-5), 2007, pp.254-268.
12. Cremonini, R., "Identification of Italian landraces of bean (Phaseolus vulgaris L.) using an image analysis system". *Scientia horticulturae,* 121(4), 2009, pp.410-418.
13. Ozkan, I.A., "Multiclass classification of dry beans using computer vision and machine learning techniques". *Computers and Electronics in Agriculture*, 174, 2020, pp.105507.
14. Çelik, Ş., "Modelling and estimation of chickpea production in Turkey using Artificial Neural Networks and Time Series Analysis". International *Journal of Engineering and Science*, 10(11), 2020, pp.01-07.
15. Arenga, D.Z.H. and Cruz, J.C.D., "Ripeness classification of cocoa through acoustic sensing and machine learning". *In: Proceedings of the 9th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment and Management*, Manila, Philippines: IEEE, 2017, pp.1-6.
16. Arboleda, E.R., "Comparing performances of data mining algorithms for classification of green coffee beans". *International Journal of Engineering and Advanced Technology*, 8(5), 2019, pp.1563-1567.