Sigma Journal of Engineering and Natural Sciences
Web page info: https://sigma.yildiz.edu.tr
DOI: 10.14744/sigma.2026.2091

Research Article

Advancing agricultural quality control: An AI-based web application for dry bean classification using machine learning models

Aniket SUDKE[1,*], Harshada JADHAV[1,*], Aishwarya BANGAR[1], Yashraj HOLKAR[1], Smita S. KULKARNI[1,*]

[1]*Department of E&TC Engineering, MIT Academy of Engineering, Alandi, Pune, 412105 India*

ARTICLE INFO

Article history
Received: 25 August 2024
Revised: 17 October 2024
Accepted: 13 January 2025

Keywords:
Cat Boost; Dry Beans, Feature Engineering; Machine Learning; SVM

ABSTRACT

This research technique focuses on dry bean classification using a machine-learning (ML) algorithm to enhance the quality of agricultural products, improve market value, and automate processes, thereby overcoming the limitations of conventional techniques. This technique supports the food industry business by improving the quality of agricultural products. In this work, seven types of dry beans are classified using a machine-learning algorithm via a web-based AI model. The machine learning algorithms were trained on 13,611 samples of the dataset. The dataset is processed through feature normalisation and outlier removal to improve ML model performance. The performance of the ML model was checked through the performance matrices, F1 score, accuracy and area under the curve (AUC). Through this performance metric, it is observed that CatBoost performs best to classify dry beans accurately compared to other ML models. In the experiment, different feature engineering and hyperparameter tuning refer to improving the classification score for dry bean types. The results showed that this AI web-based application accuratly classify dey beans, which help automization and quality check in the agricultural sector.

Cite this article as: Sudke A, Jadhav H, Bangar A, Holkar Y, Kulkarni SS. Advancing agricultural quality control: An AI-based web application for dry bean classification using machine learning models. Sigma J Eng Nat Sci 2026;44(3):2219–2232.

INTRODUCTION

In agriculture, due to drastic changes in weather conditions and soil nature, it is very necessary to check the quality of dry beans to maintain the quality. Beans are essential to the global food chain, but their growth is highly susceptible to the effects of the changing climate. It is imperative to select premium seeds based on a range of factors, including genetic, physiological, hygienic, and physical characteristics, to increase bean yield and tolerance to unfavorable environments. Research projects therefore focus on providing farmers with state-of-the-art technologies to increase output in a range of environmental conditions [1]. This includes state-of-the-art methods for handling and selecting seeds, which surpass traditional practices that primarily depend on visual inspections through germination and

***Corresponding author.**
*E-mail address: sskulkarni@mitaoe.ac.in

This paper was recommended for publication in revised form by Editor-in-Chief Ahmet Selim Dalkilic

vigor tests to determine each set of seeds' physiological potential [2].

Assessing the vitality of the seeds is a crucial task in seed production programs. Producers examine dry seeds for impurities and foreign materials in order to select the best seeds for reproduction, either by population line selection or hybridization. The procedures involved in this classification are essential to importation, exportation, and manufacture. They are similar to those in the citrus fruit market. One of the biggest problems that rural producers and markets are currently facing is choosing the right seed kinds. In the past, seeds were manually sorted based on how they appeared on the outside. This method is labour-intensive, sluggish, and inefficient—especially when making a lot of seeds. Research is therefore desperately needed to develop new techniques for grading and categorising seeds [3].

The ML is a subset of artificial intelligence that trains on historical data to predictthe values for the new input data sets. The state-of-the-art work developed a machine learning algorithm to predict the quality of seeds in the agriculture sector. These ML models accurately produce results based on mathematical models and historical data, compared to traditional methods. In the paper [4], ML models classify seeds based on features such as size, shape and colour. Artificial intelligence techniques are promising for agriculture, particularly for high-quality production [5].

Machine learning techniques solve agricultural problems effectively by analysing seed features that enhance crop yields and quality of seed selection. In this research, AI enable web application development for dry beans classification to improve scalability. This technique helps farmers and other agricultural stakeholders to support advanced farming technology for import export business.

In this paper, Section 1 presents an introduction about dry beans classification using machine learning, follwoed to that Section 2 Liturature survey. In Section 3, the dataset of dry bean classsification expained. The methodology is explained in detail in Section 4, and Section 5 presents the results and discussion. At last, Section 6 describpe in concludes the findings of the research work.

Literature Survey

The quality of the crop is important in agriculture; for the same reason, there should be proper machinery to check the quality of seeds. In recent research, machine learning plays a vital role in accurately classifying seed types. In the research [6], sustainable agriculture was proposed through an ML algorithm, Support Vector Machine (SVM), to classify dry bean varieties. These research results are remarkable, with an accuracy of 93.13% for SVM. Additionally, another study applied eight different ML techniques to classify seed varieties, with the Extreme Gradient Boosting (XGB) classifier achieving the highest accuracy, significantly improved by the Adaptive Synthetic (ADASYN) algorithm for balancing the dataset [7]. These developments in ML technology have high potential in classifying different types of beans. The ML model relies on the agriculture dataset, which resolves classification through mathematical analysis. The Synthetic Minority Over-sampling Technique (SMOTE) technique is applied on an imbalanced dataset in [5] to improve ML classifier accuracy. A support vector machine model is used on an image dataset based on colour changes and shape to strengthen the feature set to get good results from the ML model. This study highlighted the effectiveness of Bayesian optimization in improving the classification accuracy of ML models [8].

One research project examined the physical and geometric characteristics of 21 dry beans and soybean seeds, emphasizing parameters such as water uptake, oil- and water-holding capacities, foam volume, and gel formation. In the state of the art, it is observed that to maintain food quality [9], it is necessary to check beans in terms of carbohydrates, proteins, and various minerals, to reduce the effect of diabetes, obesity and cancer. The review underscored the importance of these beans in human nutrition due to their antioxidant capacity and other health benefits [10]. In the existing research, work is either on nutritional or catarigarizatiion of beans. It is important to have an integrated method that incorporates nutritional or catarigarizatiion of beans. In [11], ML models were used for data analysis and classification for such usecases.

The efficiency of phosphorus (P) usage in genotypes of dry beans grown on Oxisols in South America has also been a topic of interest, given that P is among the elements that most restrict dry bean production in this area in terms of yield. A greenhouse experiment evaluated the P utilization effectiveness of 20 superior dry bean genotypes at low and sufficient soil P levels. The study revealed significant genotype variations in yield and yield components under different P levels, highlighting the importance of effective soil management practices to enhance bean production efficiency and soil health [9]. Moreover, another study utilized both machine learning (ML) and deep learning (DL) techniques to classify dry bean species, achieving accuracies ranging from 88.35% to 93.61%. This research demonstrated the potential of these technologies in improving species classification and contributing to agricultural sustainability [1].

The antinutritional effects of tannins in dry beans, which adversely affect protein digestibility and growth in animals, have also been examined. Tannins, primarily located in the seed coat, interact with proteins to form tannin-protein complexes, reducing protein solubility and digestibility. These effects can be mitigated through various processing methods such as dehulling, soaking, cooking, and germination, as well as through genetic selection to breed low-tannin varieties. This research highlights the importance of reducing tannin content to improve the nutritional quality of beans [12].

In addition to addressing nutritional and classification aspects, the sustainability of legume production is a critical issue. A study investigating the challenges faced by Konya

City farmers in bean farming revealed deficiencies in fertilization techniques, irrigation practices, disease and pest control, and seed quality. Addressing these issues is essential for enhancing bean yield and quality, promoting agricultural sustainability, and meeting the nutritional demands of the population. The study emphasized the need for establishing policies that support sustainable legume production and consumption [13].

Furthermore, optimising the drying process of Robusta coffee beans using hybrid solar systems and Wi-Fi-enabled moisture level indicators has shown significant improvements in drying times and maintained bean texture. This method, meeting Indonesian national standards, highlights the potential for technological advancements in improving agricultural practices and product quality [14]. Similarly, a study evaluating the effects of Ascophyllum nodosum extract on *'Alvorada'* bean seeds' germination and early development demonstrated increased vigor and potential for field establishment. This finding underscores the benefits of natural extracts in enhancing seed quality and crop yield [2].

This research aims to uncover the genetic factors influencing these nutritional traits, which are critical for human health, particularly in regions reliant on plant-based diets [10,15]. Additionally, the application of image processing techniques for classifying and measuring the geometric parameters of various beans has demonstrated a linear relationship between axial dimensions and pixel values, providing a reliable method for quality assessment [4,16].

Reliable preprocessing techniques and user-friendly interfaces are essential for improving the application of ML models in real-world agricultural situations. This research proposed an online web application along with data processing and ML classification to automate the user interface. As an agricultural dataset is irregular, to achieve accurate performance of the ML model, feature engineering and data normalisation are necessary for accutate predication. It is necessary to make sure that models can accurately categorise bean varieties on a real-time feature set in production. Then, high-quality data acquisition, which includes creating and using datasets that capture the many physical characteristics of dry beans with high resolution and well-annotated features. This enhancement would help models detect subtle quality attributes. And last is a user-friendly design. Combining ML models with a web-based platform to improve usability and accessibility, enabling people with little technical expertise to access advanced ML approaches.

In summary, the agricultural product quality check process can be completely transformed with the help of an automated web application integrated with the help of machine learning and dataset preprocessing. This will enhance agricultural production with quality by predicating quality and type of crop. Due to this, the agriculur filed, which is most sensitive to weather conditions possible to maintain the quality of crops to support global food security.

DATASET

The scientifically collected data set carried out by Koklu and Ozkan [6] was the source of the dataset used in the study, which correlates responses to rows of 13,611 dry beans from seven different types that the Turkish Standards Institute determined. The seeds were acquired from accredited seed suppliers, comprising Sira, Barbunya, Bombay, Cali, Dermason, Horoz, and Seker. [Dataset Link]

i. Cali: White in color, slightly plump, kidney-shaped seeds that are bigger than dry beans.
ii. Horoz: Medium-sized dry beans that are long and cylindrical in shape.
iii. Dermason: White dry beans that are fuller and flat, with one end round and the other end slightly round.
iv. Seker: Big, round, white seeds with an irregular shape.
v. Bombay: White seeds that are exceptionally huge, round, and bulging in structure.
vi. Barbunya: Beige-colored background with red stripes or variegation; seeds are big and oval, almost spherical in shape.
vii. Sira: Small, white seeds that are flat, with one end spherical and the other flat.
viii. Bombay: Bombay has a white tint, large seeds, and a bulging, oval morphological form.

Figure 1 describes the total number of pieces in each class as well as the frequency distribution of the dry bean classes. The class Bombay contains the fewest seed samples, with 522. Dermason has 3546 observations, which is the most. The analysis turned out a few useful dimension properties for seed separation. The seven class samples were therefore unbalanced. Various machine learning techniques are applied to assess classification performance in order to maintain parity between the dry bean classes.

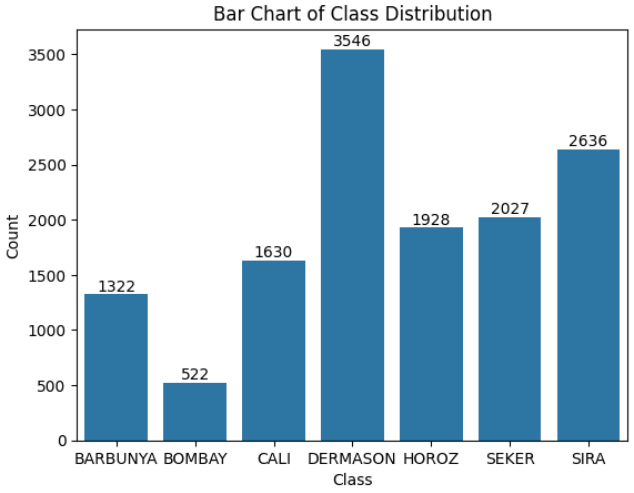

Figure 1. Distribution of frequencies for several types of dry beans.

Methodology for Dry Bean Classification

In Figure 2, the workflow till model deployment is shown. The collection of the dry bean dataset is the first stage in the approach, which is then followed by preparatory actions including normalization and data cleaning. Next, the dataset is divided into two parts: one is testing, and the second is training sets. The most important characteristics of the beans are determined by calculating feature importance. Afterwards, different machine learning models, such as KNN, SVM, Decision Tree, Random Forest, Logistic Regression, XGBoost, Linear SVC, BernoulliNB, Cat Boost Classifier, AdaBoostClassifier, Bagging Classifier, and Extra Tree Classifier, are chosen. Classification is carried out once the model has been trained with the chosen features, and performance measures are used to assess the outcomes. When the model produces findings that are satisfactory, it is finally deployed.

Feature Engineering

In dry bean classification, data preprocessing and feature engineering are very important to improve the performance and reliability of machine learning models. Data

Figure 2. Workflow of model deployment.

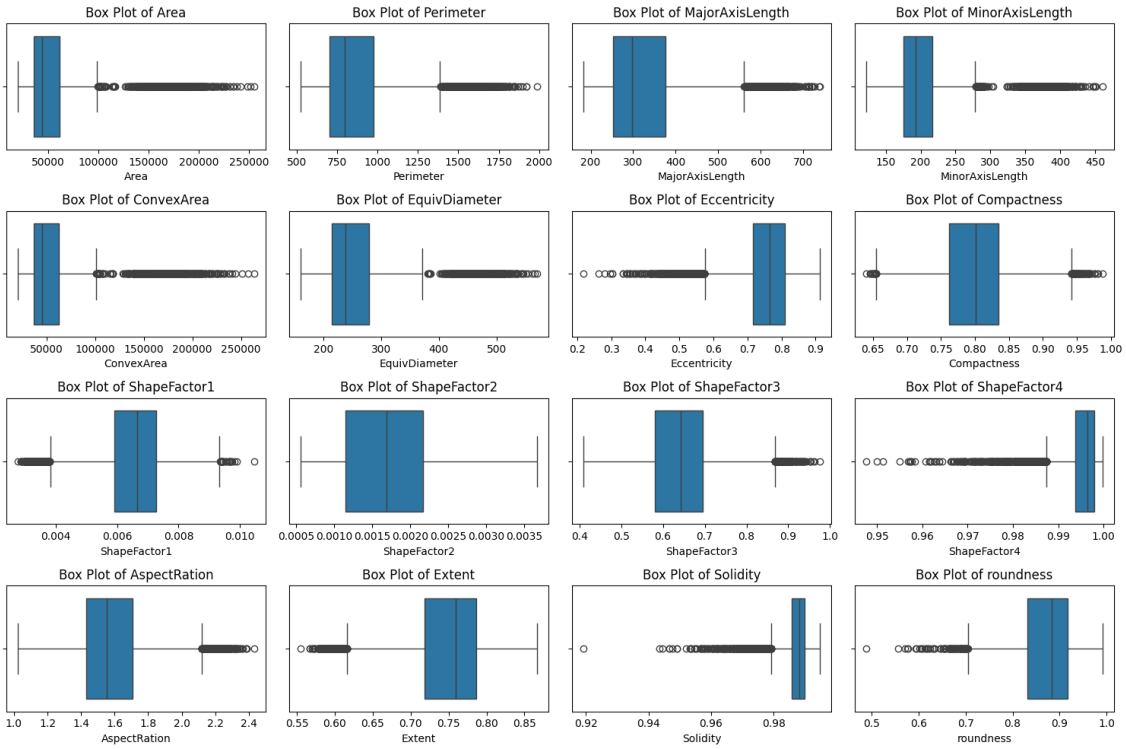

Figure 3. Box-plot for different features of dry beans.

preprocessing includes converting categorical data into numerical values, removing duplicates, handling missing dataset values, and normalising the data. By removing duplication in the dataset, bias is reduced, and the convergence speed of machine learning models is improved by applying normalisation techniques on the dataset, also handling missing values to complete the dataset to reduce model training error. In feature engineering, transformed data into important features to learn the dataset pattern in developing an accurate, robust machine learning model for dry beans classification.

In this subsection, we have used the statistical methods of interquartile range (IQR) and boxplot to identify outliers and missing values from the dataset. In exploratory data analysis, boxplots are a statistical technique that are usually used to identify outlier presence in the dataset, which impacts ML model performance, such as KNN, SVM. A box plot can find outliers with the help of a distribution of data on the Interquartile Range (IQR). The dataset point is below the "Lower Limit" = $Q1$ - 1.5(IQR) and above the ,"Upper Limit" = $Q3$ + 1.5(IQR).

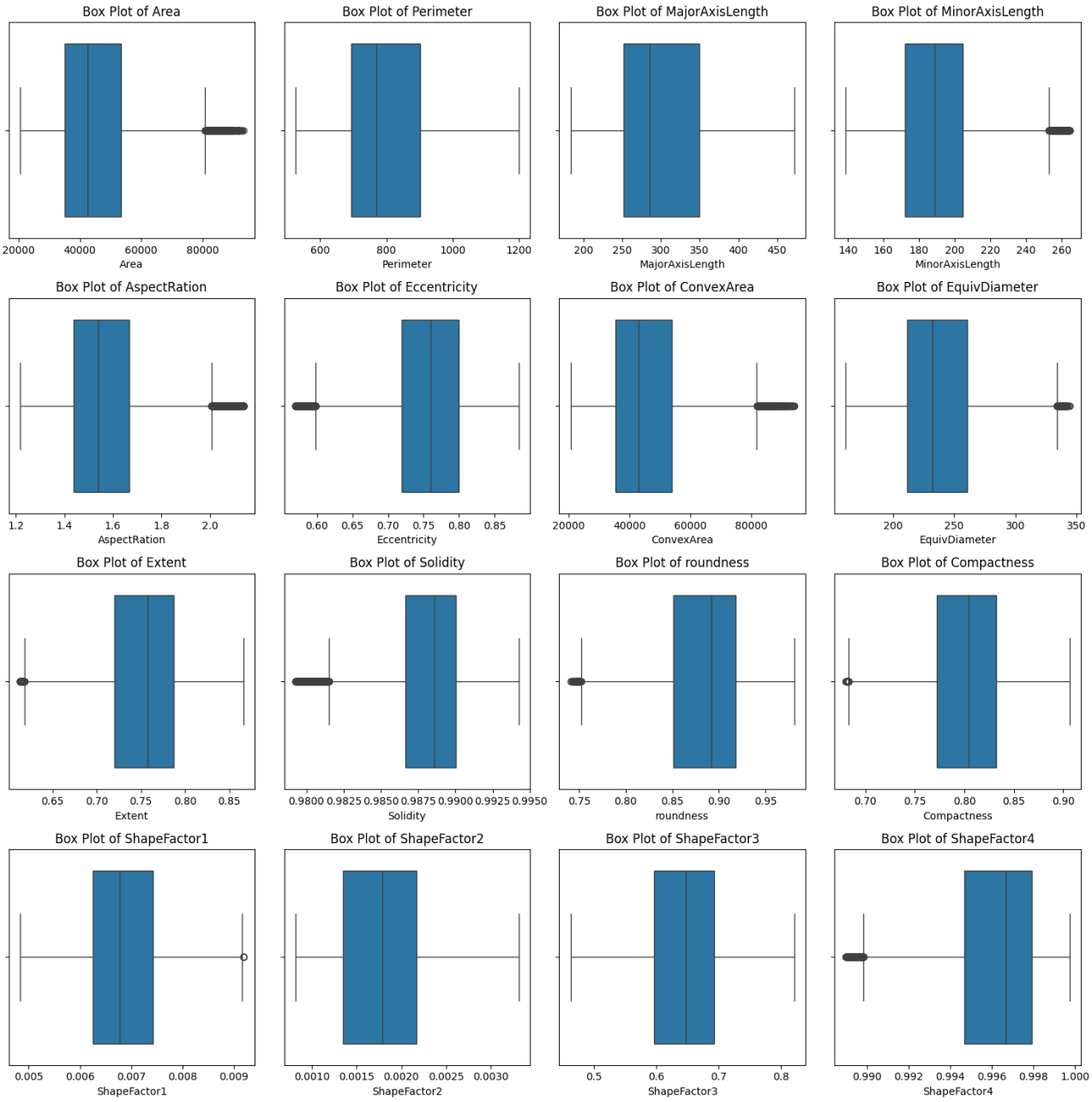

Figure 4. Outliers removal.

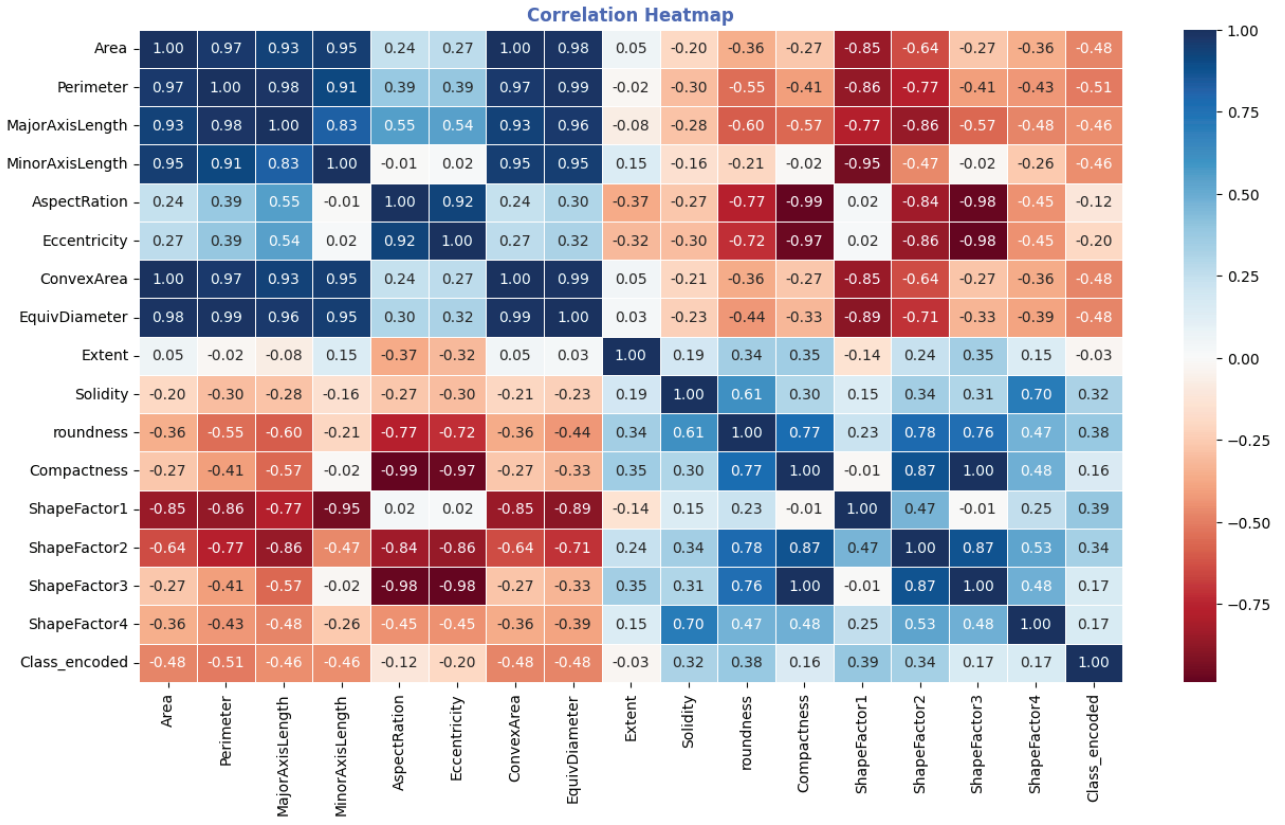

Figure 5. Correlation heatmap.

These outliers are shown in fig 3 for the feature EquivDiameter, Area, Perimeter, Minor Axis Length, Eccentricity, Convex Area, and ShapeFactor 4. As shown in Figure 4, outliers are removed using the interquartile method; as a result, the machine learning model performance is enhanced by reducing noisy outlier dataset values. Additionally, it helps to maintain data consistency to enhance generalisation, earlier convergence in training, and the process for more reliable der bean classification.

As shown in Figure 5, the correlation matrix and heatmap are presented for the dry bean dataset. It is an important tool for understanding relationships among dataset features. In the dataset, Area, Perimeter, MajorAxisLength, ConvexArea, and EquicDiameter exhibit multicollinearity, with correlations greater than 0.93. AspectRatio, Eccentricity & ShapeFactor represent identical details. ShapeFactor1 has an inverse relationship with Area, AspectRatio and Compatness. The lowest row of the heatmap in Figure 5 shows the correlation between all features and the target class, recommending that a multidimensional model is essential. From this analysis, we can infer that Compactness, ShapeFactor1, ShapeFactor2, ShapeFactor3, and ShapeFactor4 are the important features.

As shown in Figure 6, the histogram distribution of features Area, Perimeter, ConvexArea, and EquivDiameter. This shows it has a large variance with multiple peaks. Shape ratios are unimodal with the same distribution shape. Solidity, ShapeFactor4 shows a left-skewed distribution. These features, hence, required normalisation before applying to machine learning models such as KNN, SVM, etc. In this research, standardisation is used to normalise the feature, which improves the performance of the ML model.

The following is the formula to standardise a feature (x):

$$z = \frac{x - \mu}{\sigma} where\ the\ standard\ deviation\ is\ represented\ by\ \sigma$$

DRY BEANS CLASSIFICATION BY MACHINE LEARNING MODEL

K- Nearest Neighbors (KNN)

The K-Nearest Neighbours (KNN) algorithm works on distance based on similarity between shape-related features for classification. This algorithm required dataset preprocessing and feature engineering, such as standardisation and normalisation, to improve the classification accuracy. The category of new bean is classified based on the majority distance metric value among the KNN based on distinct feature patterns.

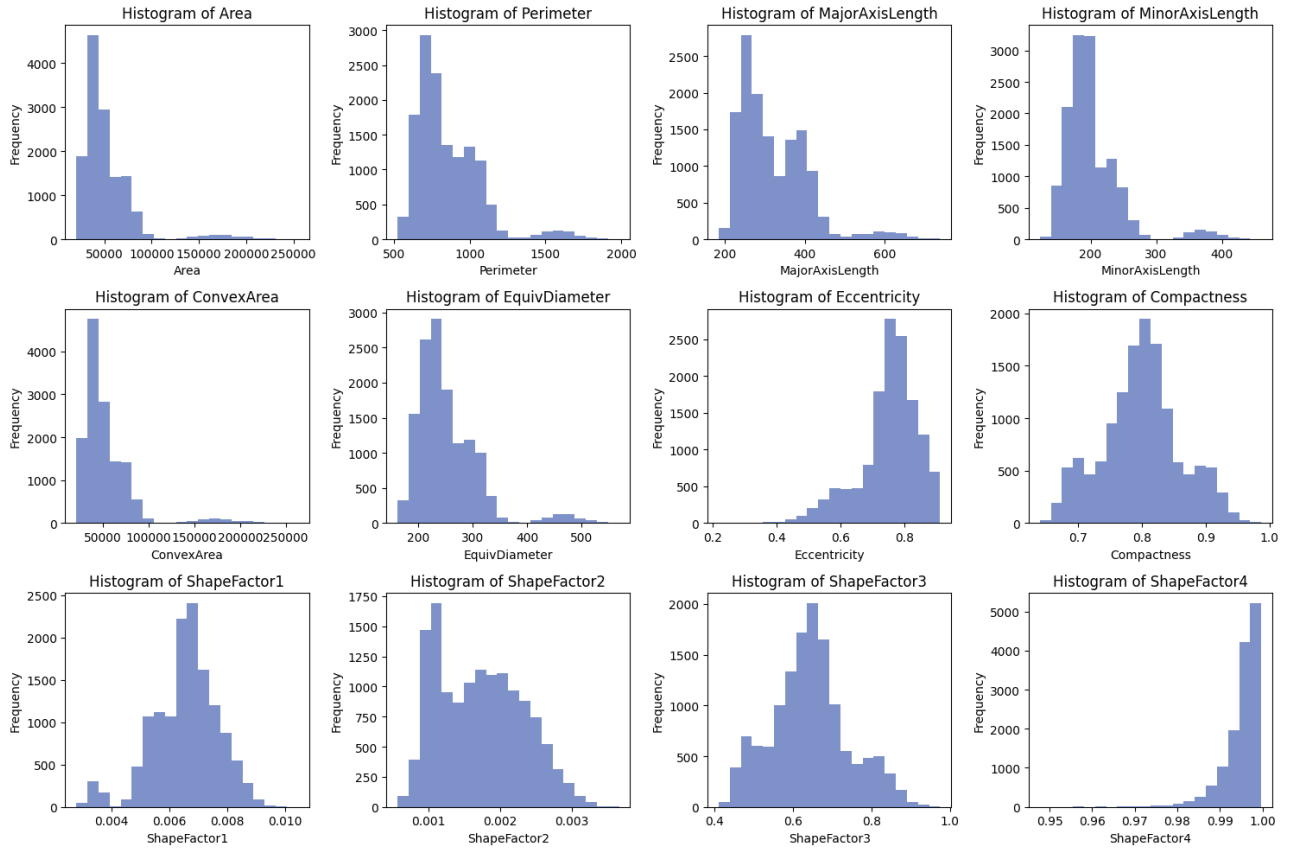

Figure 6. Multiple histograms plot.

$$d(p,q) = \sqrt{\sum_{i=1}^{k} (p_i - q_i)^2} \quad (1)$$

In Equation 1, p and q represent two points. k training set data points that are closest to the new point are then found. Assigning the most common class among these k neighbors is the algorithm's method of classification. The average of these neighbors values is calculated for regression.

Support Vector Machine (SVM)

Support Vector Machine (SVM) is an approach to supervised learning that is used for tasks related to regression and classification. The hyperplane that most effectively separates the data into different categories is identified. Optimizing the margin—a measure of the separation between the hyperplane and the nearest data point from each class—is the primary objective. For a linear SVM, the decision boundary is provided by:

$$w \cdot x + b = 0 \quad (2)$$

In Equation 2 the bias is represented by b and the weight vector by w. The sign of the function determines the classification rule for a new data point x, which is

$$class = sign(w.x + b) \quad (3)$$

Converting non-linear data is made possible by SVM's ability to find a linear separator in a higher-dimensional space through the application of kernel functions.

Decision Tree

The decision tree is a supervised learning technique that can be used for regression and classification issues. By splitting the data into subsets according to the supplied feature values, it builds a tree structure. An internal node in the tree represents a feature, a branch a decision rule, and a leaf node an outcome. When splitting data, the algorithm chooses the optimal feature and threshold to reduce impurities. Gini impurity is typically employed in classification.

$$Gini(j) = 1 - \sum_{\{i=1\}}^{n} p_i^2 \quad (4)$$

In Equation 4 p_i represents the percentage of class i samples in node j. The tree is constructed recursively until a stopping condition, such as the satisfaction of a minimum number of samples per leaf or a maximum depth, is met.

Random Forest

In order to increase accuracy and decrease overfitting, the Random Forest ensemble learning method builds several decision trees and then aggregates their predictions. The tree uses bootstrap sampling to generate subsets of the

original data and introduces unpredictability by including a random selection of features for each split. Usually, either a majority vote for classification or the average of the outputs from every single tree for regression is used to get the final forecast. The reduction in impurity can be used to determine the node's significance in a random forest. The significance of the node is determined by the following method:

$$si_j = w_j I_j - w_{left(j)} I_{left(j)} - w_{right(j)} I_{right(j)} \quad (5)$$

In Equation 5 the j^{th} node, the impurity value of the j^{th} child node is I_j node j's impurity value, w_j is the weighted sample size that reaches the j^{th} node, and right(j) and left(j) represent the child nodes that result from the right and left splits on the j^{th} node, respectively.

Logistic Regression

A parametric statistical model known as logistic regression is used for classification. To determine the likelihood that an input belongs to a specific class, the model makes use of the logistic function. The definition of the logistic function, sometimes referred to as the sigmoid function, is:

$$p(x) = \frac{1}{1+e^{-(w.x+b)}} \quad (6)$$

Here in Eq.6, w is the weight vector, and x stands for the input vector and b is the bias. Maximum Likelihood Estimation is used in the training of Logistic Regression to identify the values of w and b that maximises the likelihood of the observed data.

Bagging Classifier

In dry bean classification, bagging is an ensemble learning method implemented to reduce overfitting and enhance classification accuracy. This model was trained on multiple decision trees using a different subset of the dataset, termed as bootstrap sampling. In this, every model is trained independently and classifies the bean category; afterwards, final categorisation is based on majority voting. Bagging enhances classification stability by improving overall classification performance.

The combined prediction's formula is:

$$\hat{y} = \frac{1}{B}\sum_{b=1}^{B} \quad \hat{y}^b \quad (7)$$

In Equation 7 B is the number of base models and $\hat{y}^b$ This is the prediction from the base model.

AdaBoost

Adaboost is an ensemble algorithm that trains base learners sequentially. Base learner is a one-depth decision tree considered as a stump, treated as a weak learner. In this, multiple weak learners combine to develop a robust classifier. In sequential learning, AdaBoost assigned a higher weight to incorrectly classified data samples. This improves model performance by classifying all dataset samples. This model is beneficial mainly for morphological features with maximum accuracy.

$$\hat{y} = sign\left(\sum_{m=1}^{M} \quad \alpha_m \, h_m(x)\right) \quad (8)$$

In Equation 8 α_m is the weight assigned to the m-th weak learner h_m.

Linear SVC

A Support Vector Machine (SVM) optimises a hyperplane that identifies support vectors nearest to the data points from each class of the dry bean dataset. This is a supervised machine learning algorithm maximizes the margin based on the shape and size features. In the case of SVM, the role of preprocessing and feature engineering on the dataset is very important, else maximize margine will dominate due to large-scale features. As dry bean classification is multiclass, experimentation is performed through a multiclass SVM. One-vs-Rest strategies work well compared to One-vs-One in terms of computational efficiency. The regularisation parameter C controls missclassification between confused classes. Thus, SVM is an efficient technique in agricultural quality control and categorisation applications. In case of higher accuracy, ensemble methods are proposed to overcome class overlap.

The Linear SVM decision function is:

$$f(x) = w.x + b \quad (9)$$

Here id Equation 9, b represents the bias constant and w is the weight array. Regarding an entirely new sample x, the predicted outcome is:

$$class = sign(w.x + b) \quad (10)$$

CatBoost Classifier

The CatBoost algorithm was developed for categorical data competently based on the gradient boosting technique. This algorithm does not require wide-ranging preprocessing techniques on features. This algorithm prevents overfitting via the boosting technique through minimum parameter tuning and performs with high accuracy. Owing to this, CatBoost efficiatly clasify dry been structured datasets in minimum execution time.

$$\hat{y} = \sum_{t=1}^{T} \quad \gamma_t h_t(x) \quad (11)$$

In Equation 11, T is the number of trees, γ_t is the learning rate, and $h_t(x)$is the t-th tree. Gradient descent techniques are used to optimize the objective function.

XGBoost Classifier

XGBoost is an authoritative and competent machine learning algorithm built on gradient boosting decision trees. XGBoost is a robust, predictive model that integrates several weak decision trees to enhance model performance. This model includes provisions to handle missing values automatically with a regularisation technique & provisions for parallel processing to reduce overfitting and speed up

the training process. A regularisation term is included in the objective function:

$$Objective = \sum_{i=1}^{n} \; l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \; \Omega(f_k) \qquad (12)$$

In Equation 12, the true value is represented by y_i the predicted value by $\hat{y}_i$ the loss function by l and the regularisation term for the complexity of the model by $\Omega(f_k)$.

BernoulliNB

For binary characteristics, Bernoulli Naive Bayes (BernoulliNB) is a variation of the Naive Bayes algorithm. It uses the Bernoulli distribution and assumes that the features are binary-valued. Given a feature vector x, the likelihood that it belongs to class c is given by:

$$P(x)\alpha \; P(c) \prod_{i=1}^{n} \quad P(c) \qquad (13)$$

The Equation.13 represent probability of feature *i* being present in class *c* is denoted as *P*(*c*).

MODEL DEPLOYMENT

Based on predetermined parameters, the SVM classifier model for dry bean classification is intended to help farmers precisely identify various bean varieties. The SVM, as the best-performing model, is selected for deployment. This model's performance is evaluated by accuracy and F1 score. In the agriculture domain to acceblible for the farmer the Flask framework is implemented to deploy ML model into a web application.

The SVM modle is tested on all features along with prpeprocessing and feature engineering piplene. The modle is rbouste agasite real time feature veactor due to hyperparametr tumning. Once the model is trained pickel file is devloped for deployment on web application.

The Flask server is configured with the app.py script file for the deployment process. The frontend design is done with JavaScript, for arranging the physical style appearance with CSS, and the structure is developed with HTML. This helps farmers to access the web application in real-time easily.

This webpage is user-friendly for farmers to enter details about their beans, such as area, diameter, solidity, and shape, to categorise beans. Farmers can accurately categorise their beans using this web application in real-time, which enhances the quality and production of the agricultural sector. This increases yield production and crop management with a machine learning automated web application.

RESULTS AND DISCUSSION

Tables 1 and 2 shown exprimental results on different machine learning models on a preprocessed dataset for dry bean classification. Evaluation metrics such as F1 score, mean square error (MSE), accuracy (ACC), and area under the curve (AUC) were used to evaluate each model's performance.

Result Comparison

The results demonstrate that the Support Vector Machine (SVM) and CatBoost Classifier performed exceptionally well compared to the other models. As shown in Table 2, the SVM model categorises dry bean classes efficiently with the performance matrices precision, recall and accuracy as compared to other ML models. The CatBoost

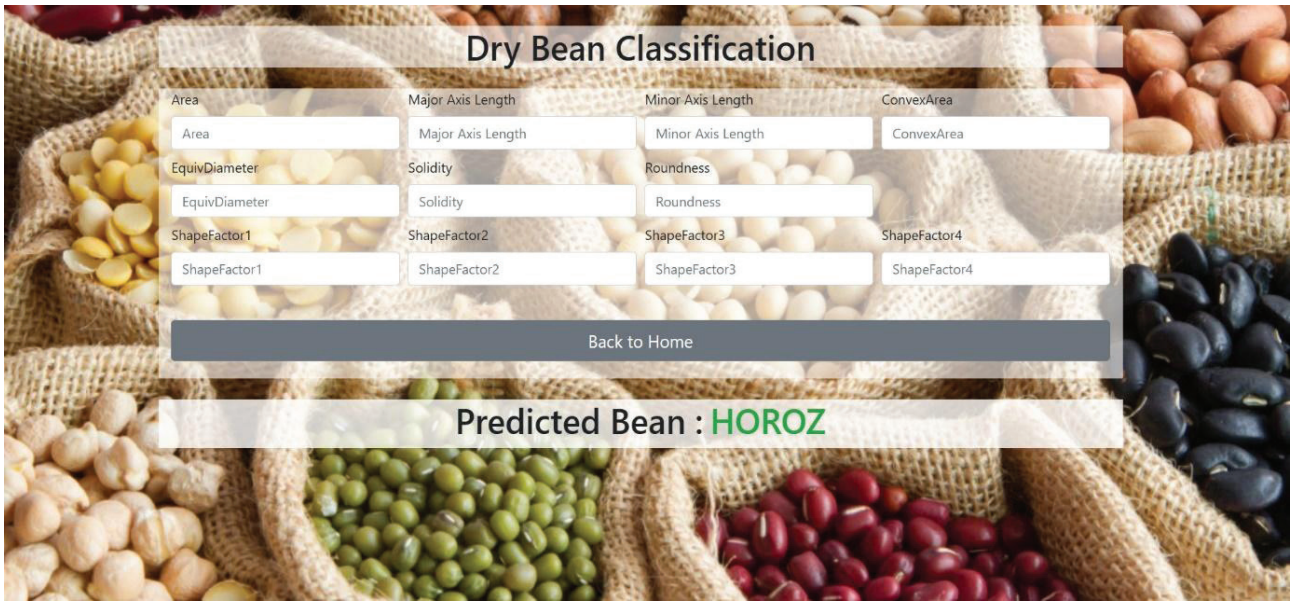

Figure 7. Web application to find the bean.

Table 1. Evaluation metrics for classification algorithms (percentage) on the dry bean dataset Without Preprocessing.

Without preprocessing				
Classifiers	**F1 Score**	**MSE**	**ACC**	**AUC**
KNN	0.7854	1.1873	0.8297	0.8649
SVM	0.8134	1.0821	0.8536	0.9041
Decision Tree	0.8561	0.9164	0.8834	0.9026
Random Forest	0.8712	0.8001	0.9058	0.9236
Logistic Regression	0.6013	1.453	0.7103	0.7563
Extra Trees Classifier	0.8645	0.8536	0.8934	0.9125
Bagging Classifier	0.8812	0.8236	0.9125	0.9364
AdaBoost Classifier	0.5015	2.0120	0.6070	0.7123
Linear SVC	0.6538	1.3672	0.7183	0.7501
Cat Boost Classifier	0.8964	0.7523	0.9289	0.9369
xgboost	0.8823	0.7796	0.9189	0.9241
BernoulliNB	0.5016	2.1456	0.6093	0.7200

Table 2. Evaluation metrics for classification algorithms (percentage) on the dry bean dataset With Preprocessing

With Preprocessing				
Classifiers	**F1 Score**	**MSE**	**ACC**	**AUC**
KNN	0.923236	0.616797	0.923115	0.992761
SVM	0.933080	0.530362	0.932909	1
Decision Tree	0.892811	0.822478	0.892752	0.944750
Random Forest	0.917538	0.615818	0.917728	0.991200
Logistic Regression	0.923255	0.631489	0.922870	0.995136
Extra Trees Classifier	0.918049	0.624633	0.917973	0.994469
Bagging Classifier	0.923629	0.580558	0.923604	0.993322
AdaBoost Classifier	0.590513	1.911851	0.646670	0.887356
Linear SVC	0.918121	0.700049	0.917483	1
Cat Boost Classifier	0.925432	0.592801	0.925318	0.995525
XGboost	0.927403	0.603330	0.927277	0.995416
BernoulliNB	0.702912	1.734819	0.709354	0.943200

algorithm improves the results by presenting robustness through controlling overfitting by handling categorical features efficiently. Overall, SVM results as the best classifier with the highest value of F1 0.933, accuracy 0.933 and AUC of 1.0. However, XGBoost and CatBoost are just lower than SVM and the Bagging classifier. Though pre-processing and feature engineering techniques apply to the dataset, KNN, Logistic Regression, and Random Forest models' performance did not improve. AdaBoost, BernoulliNB and Decision Tree underperformed due to high variance. The hyperparameter and diverse preprocessing benefit the SVM model as the top ranking here for dry bean classification.

Table 3 shows the average test accuracy of several machine learning algorithms (MLAs) about how well they performed in a different study. In the table, mainly KNN, SVM, Decision Tree and Random Forest ML algorithms performed well for accuracy. As compared to all ML models, SVM performance is best with an accuracy of 0.933. CatBoost, which is based on gradient boosting decision trees, also achieved good accuracy. In view of automated dry beans classification, SVM, CatBoost and XGBoost provide robustness as per the experimental result.

Confusion Matrix

In this study, confusion matrices were used to illustrate and compile the results of different classifiers for the classification of dry beans. The actual class is represented by every row in the confusion matrix, while each column symbolizes

Table 3. MLA Accuracy comparison with other paper

MLA Names	MLA test accuracy mean	MLA test accuracy mean of other paper
KNN	0.923	0.902
SVM	0.933	0.931
Decision Tree	0.893	0.898
Random Forest	0.918	0.932
Logistic Regression	0.923	0.926
Extra Trees Classifier	0.918	0.886
Bagging Classifier	0.922	0.921
Adaboost Classifier	0.647	0.597
Linear SVC	0.917	0.877
Cat Boost Classifier	0.925	0.902
XGboost	0.927	0.935
BernoulliNB	0.709	0.127

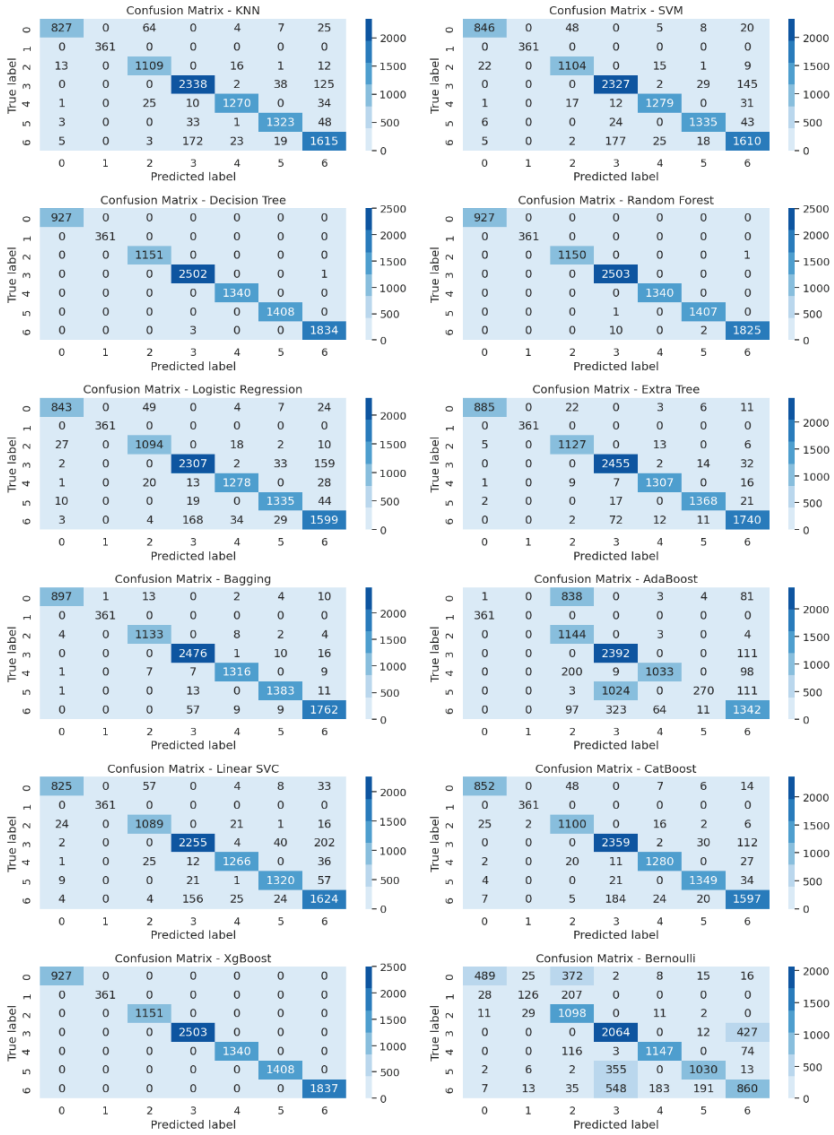

Figure 8. Confusion matrix of selected models.

the anticipated class. The diagonal elements show instances that have been successfully classified, and the off-diagonal elements are examples of incorrectly classified cases.

Overall, both SVM and CatBoost Classifier proved to be highly effective classifiers for this specific dataset, achieving perfect classification performance and significantly outperforming other models.

ROC Curve

As this research problem is multi-classification, the receiver operating characteristic (ROC) curve is presented to evaluate the performance of the ML model for different classification threshold values. In view of the methodology, the ROC-AUC investigation compares accuracy, recall, precision, and F1-score, owing to these evaluation metrics being robust. The ROC analysis's area under the curve, or AUC, indicates the degree or amount of separability. It illustrates the model's ability to distinguish between different dry bean classes. Figure. 9 shows the ROC curve that shows the performance of the K-Nearest Neighbors (KNN), Extra Tree Classifier, Random Forests, Decision Trees, Support Vector Machines (SVM), and Logistic Regression, Bagging Classifier, AdaBoost Classifier, XGBoost, CatBoost, and BernoulliNB ML classification algorithms.

The models' relative performance amongst the selected classifiers is shown in the 4b figure. The AUC value in this figure is one for the SVM, Logistic Regression, Extra Tree, CatBoost, and XgBoost models [3]. While the KNN, Random Forest, and Bagging Classifier are relatively better with AUC values of 0.99.

Discriminative ability is sometimes determined by computing the AUC, or area under the curve. AUC gives an aggregated metric and is equal to the probability in the ROC curve. The model is excellent and possesses a high degree of separation when the AUC values are close to 1. On the other hand, models with AUC values close to 0 suggest poorer performance. The test will have a more pleasing appearance if the ROC curve is closer to the upper left corner.

Model Deployment on AWS

A user-friendly machine learning model is developed and trained for the categorization of dry beans, as covered in the literature review. AWS was chosen because of its reliable, scalable infrastructure, which makes it possible for the application to effectively manage several users at once. The web application for dry bean classification is an ML deployed using Amazon Web Services (AWS) for real-world agricultural product quality control. The Flask and lightweight Python web framework are proposed for deploying an ML model that processes web requests submitt by users. An Amazon Elastic Compute Cloud (EC2) service hosts the application to maintain computational resources as per user request. This solution provides cost-effectiveness and reduces computational complexity. The deployment pipeline is presented as follows: the ML model was transferred along with supporting libraries through an EC2 instance. The HTTP access is used to make a web application publicly available. API layer constructed through Flask, a lightweight Python framework for prediction-ready models. This pipeline allows the user to enter dry bean features through a web request to classify beans in real-time. This web application provides many benefits to agricultural stakeholders, which enhance the quality and production. Additionally, the AWS set-up offers essential fault tolerance

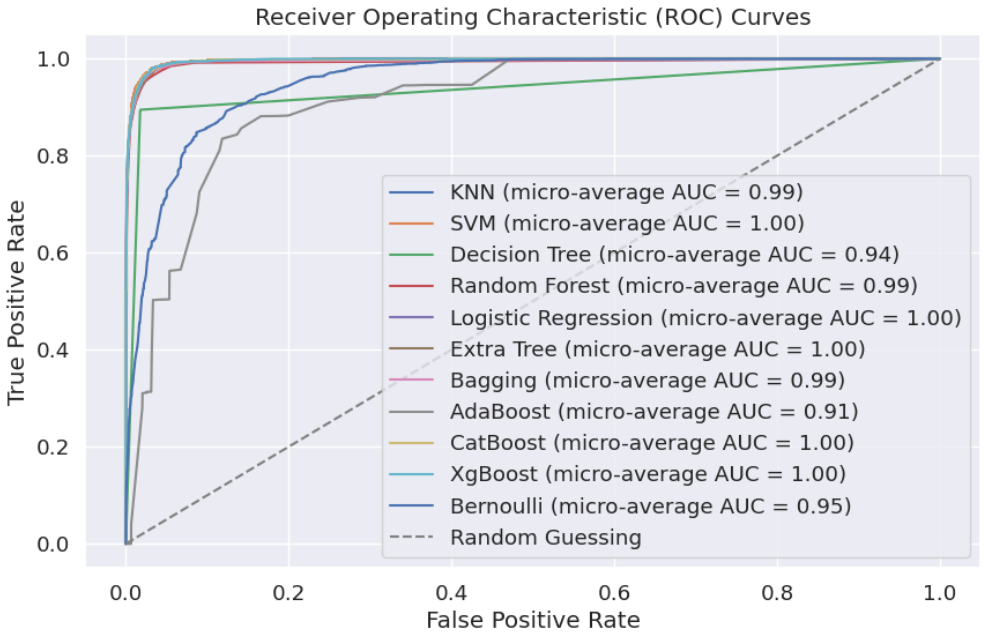

Figure. 9 ROC curve of all models.

and handles fluctuating user loads, confirming a consistent and approachable user experience. In this work, mainly this research work is not limited to laboratory purpose trained modle but deploy model at production-ready for real-time applications.

The categorization tool is available to farmers and other users at any time and from any device thanks to the application's dependable cloud-based server hosting. This configuration offers a safe and effectively manageable platform while maintaining a smooth user experience even as user demand increases.

CONCLUSION

In the agriculture domain, it is necessary to test the quality of crop products to maint sustanability of food. In earlier days, its rely on mainly on manual inspection, which is insufficient due to human error. This is challenging for dry beans classification to enhance the agriculture business. In current trades to improve market cap, it is necessary to automate dry bean sorting to maintain quality. In this research work AI- based web application was implemented to classify dry beans through machine learning algorithms. In the expriemnetaion different machine learning modles used to classify the beans by applying feature engineering techniques on the dataset. The preprocessing techniques, such as feature normalization, outlier removal, and importance of the feature enhace performance of ML models. In the expremantaion SVM and CatBoost performance results were mainly compared with AUC and accuracy score. The CatBoost performance is outstanding and selected for deployment in an AI web application. The proposed system overcomes convenation techniquies problem of conventional techniques to classify different types of dry beans by developing automated ML tecnquies. Due to this acrgiculture filed get supports to improve qaulity of food, enhance productivity and support sustainable farming. The feature scope involved integration of IoT to get real-time seed analysis. As well as propose deep learning models for accurate classification to improve maerket cap of the agriculture sector.

AUTHORSHIP CONTRIBUTIONS

Authors equally contributed to this work.

DATA AVAILABILITY STATEMENT

The authors confirm that the data that supports the findings of this study are available within the article. Raw data that support the finding of this study are available from the corresponding author, upon reasonable request.

CONFLICT OF INTEREST

The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

ETHICS

There are no ethical issues with the publication of this manuscript.

STATEMENT ON THE USE OF ARTIFICIAL INTELLIGENCE

Artificial intelligence was not used in the preparation of the article.

REFERENCES

[1] Słowiński G. Dry beans classification using machine learning. In: Proceedings of the International Workshop on Concurrency, Specification and Programming (CS&P 2021). Berlin: Humboldt University of Berlin; 2021. p. 192–203.

[2] Carvalho M, Castro P, Novembre AD, Chamma H. Seaweed extract improves the vigor and provides the rapid emergence of dry bean seeds. Am–Eurasian J Agric Environ Sci 2013;13:1104–1107.

[3] Krishnan S, Aruna SK, Kanagarathinam K, Venugopal E. Identification of dry bean varieties based on multiple attributes using CatBoost machine learning algorithm. Sci Program 2023;2023:2556066. [CrossRef]

[4] Kumar M, Bora G, Lin D. Image processing technique to estimate geometric parameters and volume of selected dry beans. J Food Meas Charact 2013;7:81–89. [CrossRef]

[5] Macuácua JC, Centeno JAS, Amisse C. Data mining approach for dry bean seeds classification. Smart Agric Technol 2023;5:100240. [CrossRef]

[6] Koklu M, Ozkan IA. Multiclass classification of dry beans using computer vision and machine learning techniques. Comput Electron Agric 2020;174:105507. [CrossRef]

[7] Khan M, Deb Nath T, Hossain M, Mukherjee A, Hasnath H, Meem TM, et al. Comparison of multiclass classification techniques using dry bean dataset. Int J Cogn Comput Eng 2023;4:92–100. [CrossRef]

[8] Naik N, Sethy P, Amat R, Behera S, Biswas P. Evaluation of optimization techniques with support vector machine for identification of dry beans. Indonesian J Electr Eng Comput Sci 2023;32:704–714. [CrossRef]

[9] Gupta S, Chhabra GS, Liu C, Bakshi JS, Sathe SK. Functional properties of select dry bean seeds and flours. J Food Sci 2018;83:2052–2061. [CrossRef]

[10] Suárez–Martínez SE, Ferriz–Martínez RA, Campos–Vega R, Elton–Puente JE, de la Torre Carbot K, García–Gasca T. Bean seeds: Leading nutraceutical source for human health. CyTA – J Food 2016;14:131–137. [CrossRef]

[11] Kumararaja K, Sivaraman B, Saravanan S. Performance evaluation of hybrid nanofluid–filled cylindrical heat pipe by machine learning algorithms. J Therm Eng 2024;10:286–298. [CrossRef]

[12] Reddy NR, Pierson MD, Sathe SK, Salunkhe DK. Dry bean tannins: A review of nutritional implications. J Am Oil Chem Soc 1985;62:541–549. [CrossRef]

[13] Fageria NK, Baligar VC, Moreira A, Portes TA. Dry bean genotypes evaluation for growth, yield components and phosphorus use efficiency. J Plant Nutr 2010;33:2167–2181. [CrossRef]

[14] Larasati DA, Kalandro GD, Fibriani I, Hadi W, Herdiyanto DW, Sarwono CS. Optimization of coffee bean drying using hybrid solar systems and Wi-Fi data communication. In: 2018 International Conference on Electrical Engineering and Computer Science (ICECOS). Piscataway: IEEE; 2018. p. 29–32. [CrossRef]

[15] Santur Y. Candlestick chart based trading system using ensemble learning for financial assets. Sigma J Eng Nat Sci 2022;40:370–379. [CrossRef]

[16] Das O. Prediction of the natural frequencies of various beams using regression machine learning models. Sigma J Eng Nat Sci 2023;41:408–418. [CrossRef]

[17] Kahraman A, Ertürk E, Kiraz F. Evaluation of dry bean farming in Konya region and its importance for sustainable agriculture. Selcuk J Agric Food Sci 2023;37:252–258. [CrossRef]

[18] Katuuramu DN, Hart JP, Porch TG, Grusak MA, Glahn RP, Cichy KA. Genome–wide association analysis of nutritional composition–related traits and iron bioavailability in cooked dry beans (Phaseolus vulgaris L.). Mol Breed 2018;38:44. [CrossRef]