D-5930
[1-8]

RESEARCH ARTICLE Agricultural Science Digest, Volume Issue : ()

Stacking Approach for Classification of Multiclass Dry Beans

Arpita Nath Boruah[1], Mrinal Goswami[1]

10.18805/ag.D-5930

ABSTRACT

Background: The most frequently cultivated edible legume crop in the world is Dry Beans (DB), which exhibit high genetic diversity. The impact of seed quality on crop production is undeniable. Seed classification is essential for production and marketing to provide building blocks for sustainable agricultural systems. With regard to big production numbers, traditional methods for classifying seed quality have flaws, including complicated processes, poor precision and sluggish inspection. Rapid and high throughput solutions are provided by automatic categorization algorithms based on machine learning and computer vision. Modern automatic classification models have made significant strides, yet there is still room for improvement by adding new methods. Since crop production is in the form of population rather than a single variation, the main goal of this study is to offer a technique for getting homogeneous dry bean variants. Although numerous intelligent models have been presented, most rely on a single classifier, which makes them unable to handle noisy and unbalanced data and can cause overfitting.

Methods: To reduce bias and variance and avoid overfitting a single classifier-based model, this study provides an ensemble-based prediction model combining pertinent attributes and a simple stacking ensemble technique, Xtreme Stacking Prediction of Dry Beams (X-SPDB). The forecast is made using the proposed X-SPDB, which incorporates several assumptions.

Result: Comparisons are made between the proposed X-SPDB's performance and simple Decision Tree, SVM, Random Forest, Naive Bayes, SVM, Logistic Regression and SVM with XGBoost.

Key words: Dry bean seeds, Feature selection, Machine learning.

INTRODUCTION

Dry beans (*Phaseolus vulgaris* L.), one of the most significant legumes in the world, are valued for their high nutritional content (vitamins, proteins, minerals and carbs), widespread consumption and large production areas (Long *et al*., 2019; Suarez-Martinez *et al*., 2016). DB has various characteristics that significantly contribute to agricultural sustainability in the context of diverse cropping systems. For any agricultural endeavor to thrive, utilizing superior seeds that can yield uniform and robust plants on time is crucial. Seed quality is determined by genetic, physiological, hygienic and physical characteristics. In 2020, the world's total DB production and harvested area were 27.5 million metric tonnes and 34.8 million hectares, respectively. While the area harvested increased by 36% during the same period, DB production has climbed by around 60% since 1990 (FAO, 2022). According to Vandemark *et al*. (2017), dried beans cultivated in the United States show an average on-farm output growth of 12.9 kg/ha per year between 1909 and 2012. The leading cause of these increases is selection for plant type, disease resistance and insect resistance. Siddiq *et al*. (2022) state that beans are essential for food security and avoiding malnutrition. Amazingly, 300 million individuals worldwide eat beans in their yearly meals.

The qualities of seeds have a significant impact on crop productivity in the agricultural industry. Various computer tools are available to assess the quality of agricultural and food products. But the majority of them are carried out by traditional means. For instance, manually determining the type of dry beans requires a knowledgeable person and a significant amount of time and it relies on human comprehension when categorizing seeds. Classifying the variety of seeds is challenging because they look similar manually. Without specialized machinery or automated software procedures, it is practically difficult for a human operator to understand or manage such seeds. The method of identifying seeds takes time and is subject to different interpretations. From a practical perspective, the situation becomes increasingly challenging regarding business and technical concerns. In particular, the color of various species of dry beans might vary and this information is not included in the geometric data. Therefore, creating an automated system for quickly identifying and categorizing seed traits is economically necessary and technically imperative. Automatic methods are greatly needed in the agricultural industry, which opens up new applications for computer vision techniques, including image categorization, identification of patterns and image splitting (Li *et al*., 2018; Zheng *et al*., 2019; Lu *et al*., 2022). Combining image processing tools and artificial

[1]Faculty of Engineering, Assam Down Town University, Panikhaiti, Guwahati-781 026, Assam, India.

Corresponding Author: Mrinal Goswami,Faculty of Engineering, Assam Down Town University, Panikhaiti, Guwahati-781 026, Assam, India. Email: mrinal.g@adtu.in

How to cite this article: Boruah, A.N. and Goswami, M. (2024). Stacking Approach for Classification of Multiclass Dry Beans. Agricultural Science Digest. DO: 10.18805/ag.D-5930.

Submitted: 14-10-2023 **Accepted:** 27-05-2024 **Online:** 31-07-2024

intelligence (AI) approaches has significantly improved the ability to analyze images and extract valuable information. Traditional seed evaluation methods can be time-consuming and subjective. However, image processing tools and AI algorithms can automate and enhance seed evaluation (Jiao *et al*., 2019; Liu *et al*., 2020; Oguine *et al*., 2022).

However, these models only use one classifier, which has certain drawbacks, such as overfitting and biases in the case of large datasets. They cannot provide accurate predictions despite a decreased error rate. As a result, this study suggests an ensemble-based Dry Beans prediction model called Xtreme Stacking Prediction of Dry Beams (X-SPDB), which uses the ensemble method because it is more stable and better predictable than a single classifier and reduces model bias, variance and overfitting while increasing predictive accuracy (Polikar, 2006; 13. Sagi and Rokach, 2018; Dong *et al*., 2020). The proposed model X-SPDB eliminates unnecessary features by utilizing the feature selection approach. When the dataset′s heatmap is studied in X-SPDB, it is clear that only a few features are connected with the target; as a result, the most pertinent features are chosen using Sequential Floating Forward Selection (SFFS).

The paper is structured as a literature survey that reviews the research and work performed on the dry bean, its prediction and its importance. Methodology: Gives an elaborate description of the model design for prediction. The result section discusses and analyzes the result and finally, the conclusion draws the work summary.

Literature survey

Several research, like those by Khilari *et al*. (2022) and Gupta and Vanmathi (2021), predict the quality of wine using machine learning algorithms. The random forest (RF) model successfully predicted wine quality in 92% and 80.9% of the cases in the two trials, respectively. Kayastha *et al*. (2024) have reviewed the fundamentals of precision agriculture, utilizing sophisticated technologies like GPS, sensors and data analytics to enhance resource efficiency and boost crop production. It emphasizes incorporating sustainable methods within precision agriculture frameworks, stressing the significance of environmental monitoring, soil vitality and biodiversity preservation. Additionally, it underscores the synergy between advanced agricultural technologies and eco-friendly farming approaches, outlining a trajectory for the agricultural sector toward sustainable and resilient nutritional security. Rajendra Prasad *et al*. (2024) have provided an overview of how Indian seed regulations have contributed to the growth of the Indian seed industry and the effects of the COVID-19 pandemic on the seed sector. Using common techniques like linear discriminant analysis (LDA), RF and support vector machine (SVM), De Medeiros *et al*. (2020) classified soybean seeds and seedlings according to appearance and physiological capacity; K-nearest neighbors (KNN) and Naive Bayes (NB) classifiers were used by Khatri *et al*. (2022) to classify the seeds of three varieties of wheat; Shingade *et al*. (2022) investigated the ability of the RF classifier to anticipate sustainable agricultural yield for a specific year; Li *et al*. (2020) employed the upgraded ILEWSM method for the visual detection of external flaws and internal quality of apple fruits using the Otsu segmentation approach and the normalized spectral ratio. Various machine learning algorithms were employed by Gupta and Vanmathi (2021) to predict wine quality and the RF model displayed the best performance, obtaining approximately 76.4% for white wine prediction and 73.3% for red wine prediction. Although the technology to categorize bean seed species was initially developed a few years ago, machine learning (ML) and artificial intelligence (AI) are now frequently utilized in research to identify dry bean seed species. Klc *et al*.'s (2007) computer vision system (CVS), which considers the samples' dimensions and color amounts, was developed for the quality control of the beans. An artificial neural network (ANN) was used to determine the hue of the beans. The samples were divided into five categories following the standards the system and the experts set. ANN was examined in 371 samples. Venora *et al*. (2009) recommended utilizing KS-400, a for-profit image analysis package, to perform a linear discriminant analysis (LDA) approach for categorizing six Italian landrace bean varieties. The experiments involved assessing traits such as the size, shape, color and texture of the grains and the results were remarkable, achieving an impressive success rate of 99.56%. Further experiments on fifteen Italian traditional landraces of beans were done by Venora *et al*. (2009) in their follow-up study, with a success percentage of 98.49%. For the Turkish Standards Institutes to define common dry bean varieties with physically similar traits but no distinctive color, Koklu *et al*. (2020) have developed an artificial intelligence-based CVS. Many machine learning methods, including kNN, SVM, MLP and DT, have been 10-fold cross-validated and compared to the model classification. 92.52%, 93.13%, 87.92% and 91.73%, respectively, were the correct classification rates for DT, SVM, kNN and MLP. Due to the cultivation of multiple populations with various genotypes, the finished products will contain seeds from several species. Oliveira *et al*. (2021) divided fermented cocoa beans into four groups using a quick and trustworthy computer vision system. Predictive traits were taken from the beans and used to identify the samples. Employing digital red, green and blue (RGB) images, they recommended employing RF to assess the quality of fermented beans as a cut test. Khan *et al*. (2023) presented a methodology that considered the removal of outliers, class balancing using adaptive synthetic and then the procedure to determine the classifier with the best performance. Aggarwal *et al*. (2022) have put forward research that facilitates providing farmers with IT-enabled solutions by employing data analytics on gathered information. It utilizes a web application designed to monitor soil fertility and offer recommendations to farmers regarding the most suitable crop(s) for cultivation in their specific

geographical region. Macuacua *et al*. (2023) developed a system for the classification of varieties of seeds automatically using different combinations of data techniques. Kim *et al*. (2024) have demonstrated that AI-powered irrigation systems outperform traditional irrigation methods by delivering significant cost savings, enhancing crop yields and promoting water conservation. They've indicated that this study represents a landmark in integrating AI into precision agriculture, paving the path for a more sustainable and productive future in legume farming. Setyaningrum *et al*. (2024) have employed a complete randomized block design featuring a single factor: fertilizer type, comprising seven levels. These levels included inorganic fertilizers (Urea 50 kg/ha, SP36 100 kg/ha and KCl 100 kg/ha), Indigofera tinctoria compost, corncob compost, peanut green manure, chicken manure, goat manure and cow manure (applied at a rate of 5 tons/ha), with each treatment replicated three times.

MATERIALS AND METHODS

The X-SPDB that has been proposed is divided into two phases: feature selection and stacking ensemble approach. To choose the most critical features from the initial set of features, feature selection is done and a stacking ensemble method is used for the classification. Fig 1 displays the conceptual diagram of the proposed X-SPDB.

Feature selection

Choosing a subset of essential attributes to incorporate into a model is a process known as feature selection. Determining the crucial attributes for classifying DB involves analyzing feature importance through tree-based classifiers and evaluating the correlation among features using a heatmap generated from a feature correlation matrix.

Stacking ensemble technique

The Base-Models and Meta-Model significantly impact Stacking performance. Base models should be selected so that they provide various predictions about the situation at hand and are very successful in resolving it. The architecture of a stacking model is given in Fig 2. As a result, Random Forest and XGBoost are included in the proposed X-SPDB due to their robustness and diversity of assumptions used in prediction. The meta-model smoothly interprets the predictions generated by the underlying models. As a result, linear models, such as logistic regression for classification tasks and regression tasks, are frequently utilized as the meta-model. So, for BD prediction, the suggested X-SPDB uses logistic regression.

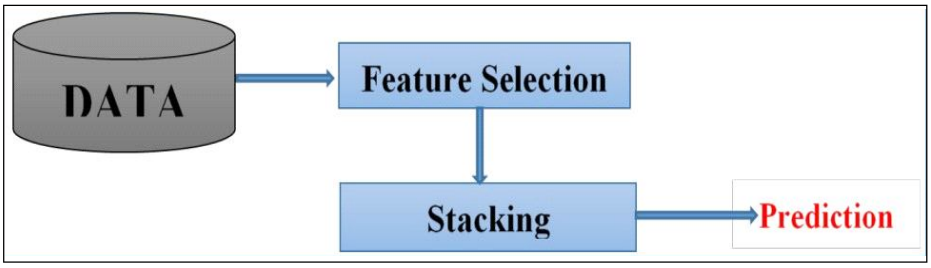

Fig 1: Structure of the proposed X-SPDB.

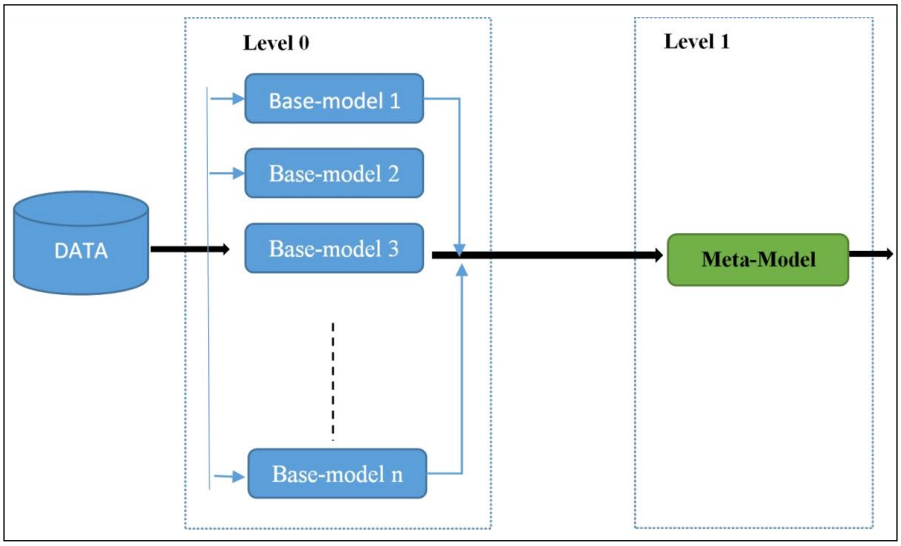

Fig 2: Stacking model architecture.

The stacking ensemble technique's schematic representation is shown in Fig 3.

RESULTS AND DISCUSSION

The experimentation uses the PYTHON 3.8 version in a Windows environment. The UCI machine learning repository, which contains 13611 samples with 16 features, served as the source of the dataset for DB (Koklu and Ozkan, 2020). There are 7 different classes of DB denoted by names: "Seker," "Barbunya," "Bombay," "Cali," "Dermosan," "Horoz," and "Sira."

In the first phase of the proposed approach X-SPDB, the features are first scored and only 4 of the 16 characteristics are found to be potentially important for the PD classification using a correlation matrix and heatmap. Thus, the four pertinent traits are chosen using the SFFS. Fig 4 and 5 display the top 10 features graphically and a heatmap of the connected features.

The top 4 relevant features are:

1. Shapefactor3
2. Shapefactor1
3. MajorAxisLength
4. AspectRatio

If the correlation coefficient is higher than 0.75, the characteristic with the lower feature score is deleted from the pair since it is deemed redundant. Under this, the features Compactness, Shape Factor 2 MinorAxisLength, ConvexArea, Area and EquivDiameter have a correlation of more than 75% than that of Shapefactor3, Shapefactor1, MajorAxisLength, AspectRatio. They are removed, proposed X-SPDB's second stage of processing the DB dataset with the four features for DB prediction. Table 1

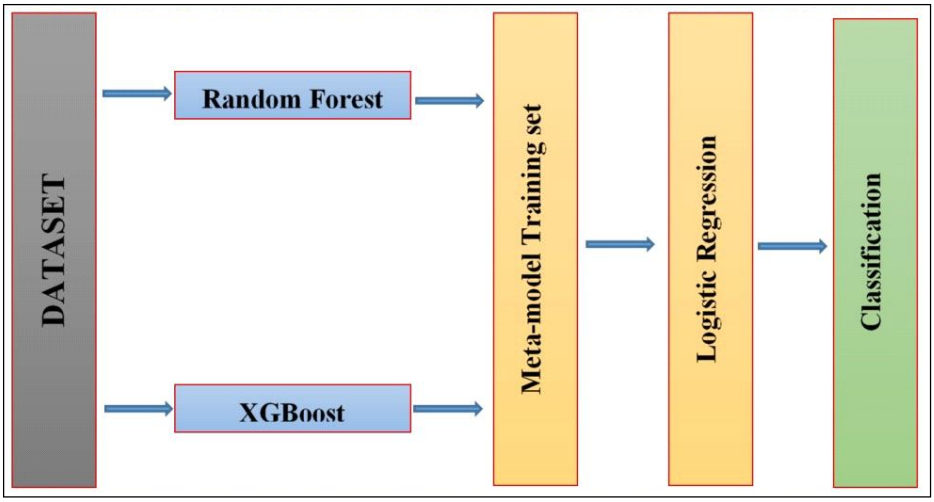

Fig 3: Stacking ensemble technique schematic diagram.

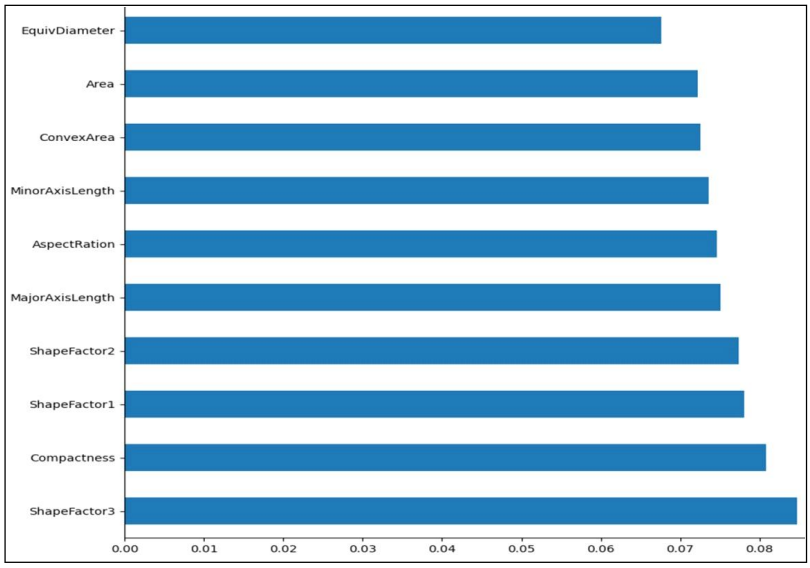

Fig 4: Top 10 features.

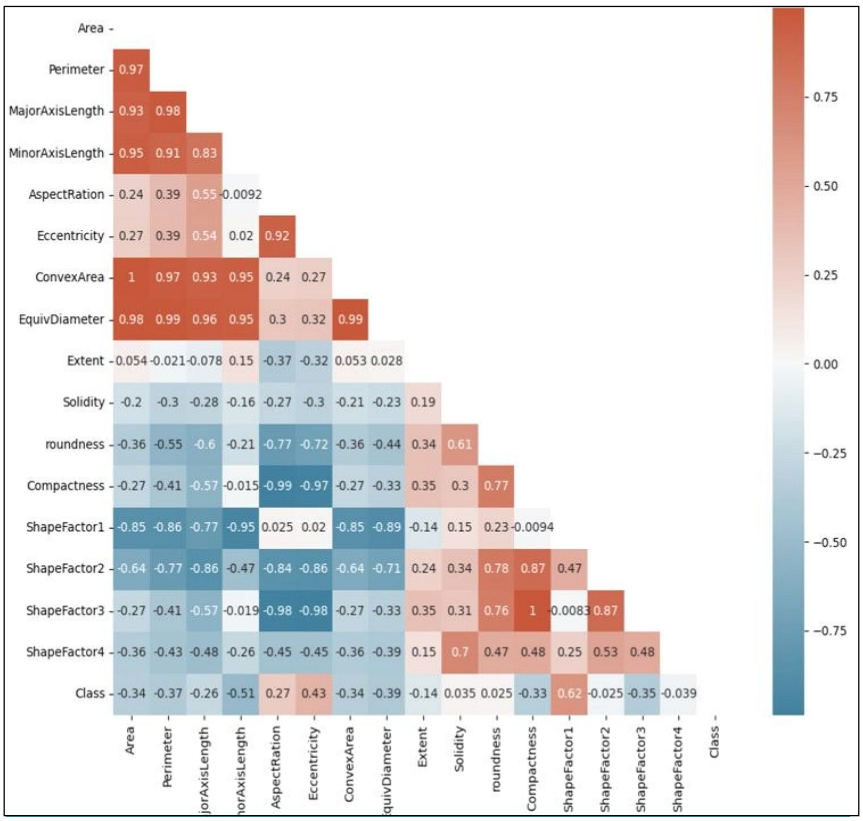

Fig 5: Heatmap of the features.

Table 1: Performance comparison.

Methods	Accuracy	Precision	Recall	F1 score
SVM	81.07	61.17	66.7	63.82
Logistic Regression	87.61	75.62	69.07	72.19
Naïve Bayes	81.71	70.20	63.4	66.63
Random Forest	92.24	72.76	67.87	70.23
XGBoost	92.41	73.30	76.60	74.913
Simple DT	84.48	67.10	77.30	71.84
X-SPDB	97.84	94.60	73.08	82.46

compares the proposed X-SPDB′s classification accuracy and precision, recall and F1 measures to that of simple DT, Logistic Regression, SVM, Naive Bayes, Random Forest and SVM.

By preprocessing the database using the SFFS feature selection method, the proposed method X-SPDB outperforms logistic regression, SVM, Naive Bayes, Random Forest, XGBoost and simple DT while enhancing the performances of the stacking ensemble method. The SFFS eliminates the redundant and irrelevant characteristics that reduce a classifier′s prediction power and effectiveness to improve performance. Therefore, the suggested approach X-SPDB has a more significant performance than the other classifiers, which are considered by combining the benefits of the feature selection and stacking ensemble methods. For easier visualization and comprehension, Fig 6 provides a graphical depiction of the performance comparison of all the categorization methods considered.

The feature selection step is eliminated for future performance analysis of the proposed X-SPDB and the model is referred to as X-SPDB-1. The effectiveness of X-SPDB-1 is thus evaluated in comparison to logistic regression, SVM, Naive Bayes, Random Forest, XGBoost and simple DT. Every classifier being compared is trained using every feature of the DB dataset. The comparison of accuracy is shown in Table 2. Table 2 shows that technique X-SPDB-1 has a higher classification accuracy than all other approaches that were considered for comparison. All single classifier-based models, including logistic regression, SVM, Naive Bayes, Random Forest, XGBoost and simple DT. As a result, these models cannot manage noisy and unbalanced data, leading to overfitting and decreased

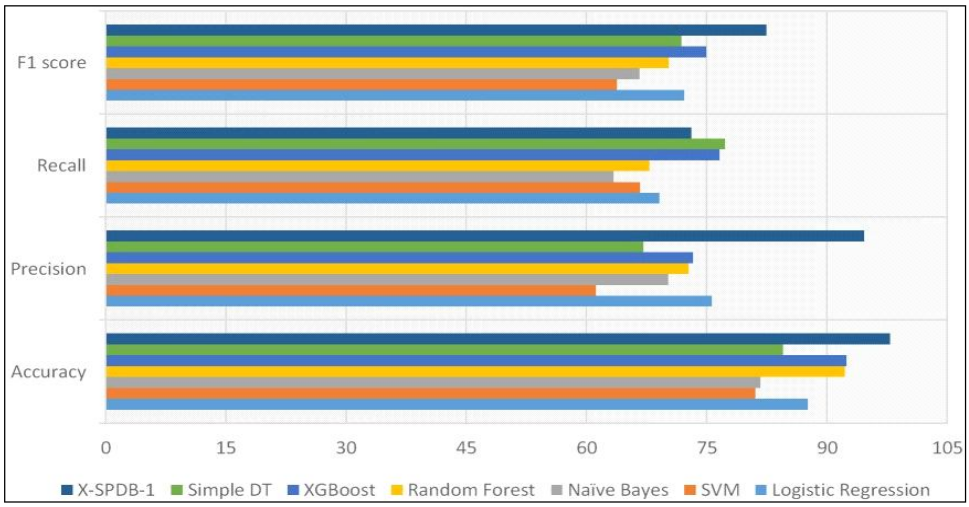

Fig 6: Graphical representation of performance analysis.

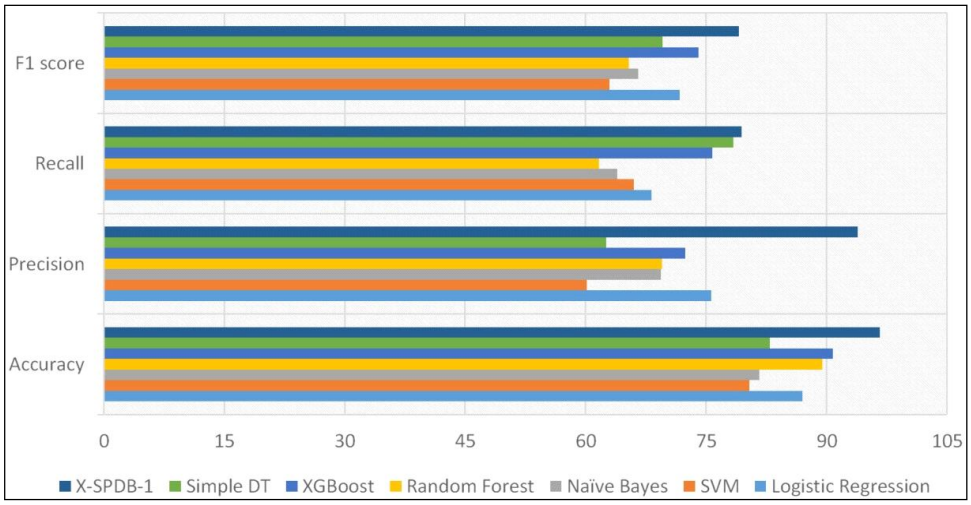

Fig 7: Graphical analysis of X-SPDB-1 performance.

Table 2: Performance comparison of X-SPDB-1.

Methods	Accuracy	Precision	Recall	F1 score
SVM	80.4	60.17	66.0	62.95
Logistic Regression	87.00	75.6	68.2	71.70
Naïve Bayes	81.6	69.40	63.9	66.53
Random Forest	89.47	69.54	61.64	65.35
XGBoost	90.75	72.43	75.76	74.05
Simple DT	82.91	62.57	78.38	69.58
X-SPDB-1	96.6	93.9	79.4	79.07

forecast accuracy. At the same time, X-SPDB-1 is a Stacking Ensemble, a meta-learning technique that eliminates the drawback of classifier-based models and minimizes variance, producing valuable results. Fig 7 provides a graphical depiction of the accuracy comparison of all the categorization methods considered for easier visualization and comprehension. For additional analysis of X-SPDB-1, F1-score, recall and precision are also used. The comparisons of the performance measures are shown in Table 2. Table 2 shows that X-SPDB-1 also performs better in precision, recall and F1-score in addition to accuracy.

CONCLUSION

The research presented here suggests the classification of DB using an ensemble-based model called X-SPDB. To effectively detect DB, the suggested X-SPDB first deleted

the unwanted and redundant attributes using the SFFS feature selection technique before processing the preprocessed data using the simple stacking ensemble technique. The SFFS feature selection technique eliminates the extraneous features that could harm X-SPDB performance. Additionally, by lowering variance and overfitting issues, the simple stacking ensemble technique solves the drawbacks of the single classifier models. Based on the findings of the experiments, the X-SPDB model demonstrates a high level of accuracy in predicting DB. The results indicate that the X-SPDB model outperforms logistic regression, SVM, Naive Bayes, Random Forest, XGBoost and simple DT models regarding accuracy, precision, recall and F1-score measures. Therefore, it can be concluded that the X-SPDB system is crucial for identifying DB. However, to ensure that only relevant data is considered, evaluating the performance of X-SPDB using different feature selection methods is advisable.

Conflict of interest

We wish to confirm that there are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome.

REFERENCES

Aggarwal A., Sharma D. (2022). IoT-based Recommender Engine for Yielding Better Crops. Bhartiya Krishi Anusandhan Patrika. 37(4): 363-368. doi: 10.18805/BKAP540.

Dong, X., Yu, Z., Cao, W., Shi, Y., and Ma, Q. (2020). A survey on ensemble learning. Frontiers of Computer Science. 14(2): 241-258.

FAO. (2022). Crop Production and Trade Data. Retrieved from http:// www.fao.org/faostat/en/#data.

Gupta M., Vanmathi C. (2021). A study and analysis of machine learning techniques in predicting wine quality, Int. J. Recent Technol. Eng. (IJRTE) 10(1): 2277-3878

Jiao, L., Zhang, F., Liu, F., Yang, S., Li, L., Feng, Z., and Qu, R. (2019). A survey of deep learning-based object detection. IEEE Access. 7: 128837-128868.

Kayastha, S., Behera, A., Sahoo, J.P., Mahapatra, M. (2024). Growing Green: Sustainable Agriculture Meets Precision Farming: A Review . Bhartiya Krishi Anusandhan Patrika. 38(4): 349-355. doi: 10.18805/BKAP697.

Khan, M.S., Nath, T.D., Hossain, M.M., Mukherjee, A., Hasnath, H.B., Meem, T.M., Khan, U. (2023). Comparison of multiclass classification techniques using dry bean dataset. International Journal of Cognitive Computing in Engineering. 4: 6-20.

Khatri, A., Agrawal, S., Chatterjee, J.M. (2022). Wheat seed classification: Utilizing ensemble machine learning approach, Sci. Program. 2626868.

Khilari, N., Hadawale, P., Shaikh, H., Kolase, S. (2022) Analysis of Machine Learning Algorithm to Predict Wine Quality, 2, Computer Engineering, Jaihind College of Engineering, Pune, Maharashtra, India pp. 231-236.

Kýlýç K., Boyacý IH., Koksel H., Kusmenoglu I. (2007) A classification system for beans using computer vision system and artificial neural networks. Journal of Food Engineering. 78(3): 897-90.

Kim, T.H., AlZubi, A.A. (2024) AI-Enhanced Precision Irrigation in Legume Farming: Optimizing Water Use Efficiency. Legume Research. doi: 10.18805/LRF-791.

Koklu, M., Ozkan, I.A. (2020) Multiclass classification of dry beans using computer vision and machine learning techniques. Computers and Electronics in Agriculture 174 105507.

Li, L., Peng Y., Yang, C., Li, Y. (2020) Optical sensing system for detection of the internal and external quality attributes of apples, Postharvest Biol. Technol. 162: 111101.

Li, J., Chen, L. and Huang, W. (2018). Detection of early bruises on peaches (*Amygdalus persica* L.) using hyperspectral imaging coupled with improved watershed segmentation algorithm. Postharvest Biology and Technology. 135: 104-113.

Liu, L., Ouyang, W., Wang, X., Fieguth, P., Chen, J., Liu, X. and Pietikäinen, M. (2020). Deep learning for generic object detection: A survey. International Journal of Computer Vision. 128(2): 261-318.

Long, Y., Bassett, A., Cichy, K., Thompson, A. and Morris, D. (2019). Bean split ratio for dry bean canning quality and variety analysis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops.

Lu, Y., Young, S., Wang, H. and Wijewardane, N. (2022). Robust plant segmentation of color images based on image contrast optimization. Computers and Electronics in Agriculture. 193: 106711.

Macuacua, J.C., Centeno, J.A.S., Amisse, C. (2023) Data mining approach for dry bean seeds classification. Smart Agricultural Technology 5.

Medeiros, A.D. de, Capobiango, N.P., da Silva, J.M., da Silva L.J, da Silva, C.B., dos Santos Dias, D.C.F. (2020) Interactive machine learning for soybean seed and seedling quality classification. Sci. Rep. 10 11267

Oguine, K.J., Oguine, O.C., and Bisallah, H.I. (2022). YOLO v3: Visual and real-time object detection model for smart surveillance systems. Computer Vision and Pattern Recognition.

Oliveira, M.M., Cerqueira, B.V., Barbon, S., Barbin, D.F. (2021) Classification of fermented cocoa beans (cut test) using computer vision. Journal of Food Composition and Analysis 97:103771.

Polikar, R. (2006). Ensemble-based systems in decision making. IEEE Circuits and Systems Magazine. 6(3): 21-45. Prasad Rajendra S., Rajatha K.D., Surabhi V.K., Rani Uma K. (2024). Indian Seed Legislation and Effect of Covid-19 on Seed Industry: A Review . Agricultural Reviews. 45(1): 89-95. doi: 10.18805/ag.R-2340.

Sagi, O. and Rokach, L. (2018). Ensemble learning: A survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 8(4).

Setyaningrum, D., Budiastuti, M.T.S., Purnomo, D., Sudadi (2024). role of organic fertilizer types on nutrient absorption and soybean yield in Teak-based agroforestry systems. Agricultural Science Digest. 44(1): 35-40. doi: 10.18805/ag.DF-574.

Shingade, S.D., Mudhalwadkar, R.P., Masal, K.M. (2022) Random Forest Machine Learning Classifier for Seed Recommendation. Proceedings of the International Conference on Edge Computing and Applications (ICECAA 2022) IEEE Xplore Part Number: CFP22BV8-ART; 978-1-6654-8232-5.

Siddiq, M., Uebersax, M. A. and Siddiq, F. (2022). Global Production, Trade, Processing and Nutritional Profile of Dry Beans and Other Pulses. In: Dry Beans and Pulses: [M. Siddiq and M.A. Uebersax (Eds.)], Production, Processing and Nutrition (2nd ed., pp. 1-28). John Wiley and Sons.

Suárez-Martínez, S.E., Ferriz-Martínez, R.A., Campos-Vega, R., EltonPuente, J.E., de la Torre Carbot, K., and García-Gasca, T. (2016). Bean seeds: Leading nutraceutical source for human health. CyTA Journal of Food. 14(1): 131-137.

Vandemark, G., Brick, M. A., Kelly, J. D., Osorno, J. M. and Urrea, C. A. (2017). Yield gains in dry beans in the U.S. USDA-ARS/UNL Faculty Publication.1781.

Venora, G., Grillo, O., Ravalli, C., Cremonini, R. (2009) Identification of Italian landraces of bean (*Phaseolus vulgaris* L.) using an image analysis system. Scientia Horticulturae. 121(4): 410-418.

Zheng, Q., Huang, W., Cui, X., Dong, Y., Shi, Y., Ma, H., and Liu, L. (2019). Identification of wheat yellow rust using optimal three-band spectral indices in different growth stages. Sensors. 19(1): 35.