Indonesian Journal of Data and Science

Volume 5 Issue 1
ISSN 2715-9936
https://doi.org/10.56705/ijodas.v5i1.124

Research Article

Evaluating the Performance of Voting Classifier in Multiclass Classification of Dry Bean Varieties

I Putu Adi Pratama [1,*]; Ery Setiyawan Jullev Atmadji [2]; Dwi Amalia Purnamasar [3]; Edi Faizal [4]

[1] UHN IGB Sugriwa Denpasar, Bali, Indonesia, putuadi@uhnsugriwa.ac.id
[2] Politeknik Negeri Jember, Jember, Indonesia, ery@polije.ac.id
[3] Politeknik Negeri Batam, Batam, Indonesia, dwiamalia@polibatam.ac.id
[4] Universitas Teknilogi Digital Indonesia, Yogyakarta, Indonesia, edifaizal@utdi.ac.id

Correspondence should be addressed to I Putu Adi Pratama; putuadi@uhnsugriwa.ac.id

Received 15 January 2023; Accepted 28 February 2023; Published 31 March 2024

Abstract:

This study explores the application of a voting classifier, integrating Decision Tree, Logistic Regression, and Gaussian Naive Bayes models, for the multiclass classification of dry bean varieties. Utilizing a dataset comprising 13,611 images of dry bean grains, captured through a high-resolution computer vision system, we extracted 16 features to train and test the classifier. Through a rigorous 5-fold cross-validation process, we assessed the model's performance, focusing on accuracy, precision, recall, and F1-score metrics. The results demonstrated significant variability in the classifier's performance across different data subsets, with accuracy rates fluctuating between 31.23% and 96.73%. This variability highlights the classifier's potential under specific conditions while also indicating areas for improvement. The research contributes to the agricultural informatics field by showcasing the effectiveness and challenges of using ensemble learning methods for crop variety classification, a crucial task for enhancing agricultural productivity and food security. Recommendations for future research include exploring additional features to improve model generalization, extending the dataset for broader applicability, and comparing the voting classifier's performance with other ensemble methods or advanced machine learning models. This study underscores the importance of machine learning in advancing agricultural classification tasks, paving the way for more efficient and accurate crop sorting and grading processes.

Keywords: Voting Classifier, Dry Bean Classification, Machine Learning, Agricultural Informatics, Ensemble Learning.
Dataset link: https://www.kaggle.com/datasets/nimapourmoradi/dry-bean-dataset-classification

1. Introduction

In the realm of agricultural technology and food security, the classification of crop varieties plays a pivotal role in supporting the efficiency of production and distribution systems. Among various crops, dry beans stand out due to their nutritional value and significance in the global food supply chain. The distinct varieties of dry beans, each with unique characteristics and applications, necessitate accurate and efficient classification methods. Traditionally, this process has relied heavily on manual inspection and basic mechanical sorting techniques, which are not only labor-intensive but also prone to errors. The advent of computer vision and machine learning technologies has opened new avenues for addressing these challenges, offering the potential for automated, high-precision classification systems.

However, the classification of dry bean varieties using computer vision presents its own set of challenges. The subtle differences in shape, size, and colour among the varieties require sophisticated feature extraction and classification techniques to achieve high accuracy [1], [2]. Moreover, the effectiveness of machine learning models in such applications is heavily dependent on the selection of appropriate algorithms and the integration of their predictions [3]. This research seeks to solve the problem of accurately classifying seven different registered varieties of dry beans by employing a voting classifier [4] that combines the strengths of Decision Tree [5], Logistic Regression [6], and Gaussian Naive Bayes classifiers [7].

The primary objective of this study is to evaluate the performance of the voting classifier in the multiclass classification of dry bean varieties, focusing on metrics such as accuracy, precision, recall, and F-measure [8]–[10]. We aim to investigate whether the synergistic combination of different machine learning algorithms can outperform individual classifiers in this context. This leads to the research question of how effective a voting classifier is in distinguishing among multiple dry bean varieties based on a comprehensive set of features extracted through a computer vision system.

The scope of this research is limited to the analysis of image data from 13,611 grains of seven different dry bean varieties, with the classification model being trained and tested on this dataset. While the findings of this study have the potential to significantly advance the field of agricultural technology, it is important to acknowledge the limitations inherent in working with a fixed dataset and the specific choice of classifiers.

This research contributes to the body of knowledge in agricultural informatics by demonstrating the application of a voting classifier for the multiclass classification of crop varieties, specifically dry beans. It not only provides insights into the effectiveness of combining multiple machine learning algorithms for classification tasks but also lays the groundwork for future studies to explore the integration of more advanced algorithms and the application of similar methodologies to other crops. Through this study, we aim to enhance the technological capabilities in the agricultural sector, promoting greater efficiency and accuracy in the classification of crop varieties, thereby supporting food security and sustainability efforts worldwide.

2. Method:

The study adopts an experimental research design, focusing on the application of machine learning algorithms for the classification of dry bean varieties. The core of the research involves the development and evaluation of a voting classifier that integrates the predictions of Decision Tree [11], [12], Logistic Regression [13], and Gaussian Naive Bayes [14] classifiers. The effectiveness of this ensemble method is assessed through cross-validation and performance metrics such as accuracy, precision, recall, and F-measure [15]. Our research is designed in five well-structured main stages, and their aspects are illustrated in **Figure 1**.

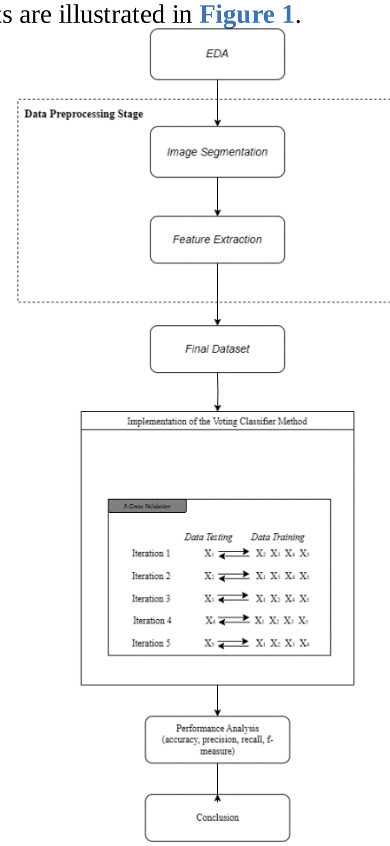

Figure 1. General Research Design Stages

Data Collection Process:

The dataset consists of 13,611 images of dry bean grains, representing seven different registered varieties. These images were captured using a high-resolution camera as part of a computer vision system designed for this purpose. Each image in the dataset is associated with 16 features, including 12 dimensions and 4 shape forms, extracted to facilitate the classification process. Splitting dataset can show in **Figure 2**.

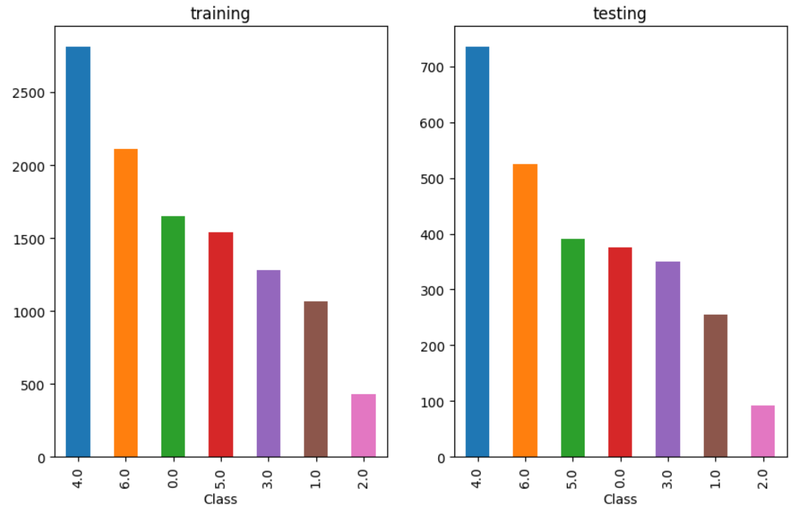

Figure 2. Splitting Dataset 20 % testing, 80% training

Tools and Technology Used

The research employs Python as the primary programming language, utilizing libraries such as NumPy for numerical computations, pandas for data manipulation, scikit-learn for machine learning algorithms, and Matplotlib for visualization. The computer vision tasks, including image segmentation and feature extraction, were performed using the OpenCV library.

Data Collection Process

The data collection process involved capturing images of dry bean grains using a high-resolution camera. Following image acquisition, the images were subjected to pre-processing steps including segmentation to isolate the beans from the background, and feature extraction to obtain measurable characteristics of each bean. The features extracted include geometric properties such as area, perimeter, major and minor axis lengths, aspect ratio, and others relevant to the classification of bean varieties.

Data Analysis Methods

Classification Algorithm: Voting Classifier

The data analysis encompasses several key steps, starting with data pre-processing to prepare the dataset for machine learning algorithms. This includes encoding categorical variables and normalizing numerical features. The core of the analysis involves constructing a voting classifier that combines the outputs of Decision Tree [11], Logistic Regression [16], and Gaussian Naive Bayes [17] classifiers based on a majority voting scheme. The formula for calculating the majority vote V for a given sample x is given by [18], [19].

$$V(x) = mode\{C_1(x), C_2(x), C_3(x)\} \tag{1}$$

Where C_1, C_2, C_3 represent the predictions made by Decision Tree, Logistic Regression, and Gaussian Naive Bayes classifiers, respectively, and mode refers to the most frequent prediction.

K-fold Cross-validation:

5-fold cross-validation was employed to assess the model's performance, ensuring the reliability of the results. The dataset is divided into five subsets; in each iteration, four subsets are used for training, and the remaining one is used for testing [20]–[23]. The process is repeated five times, with each subset used exactly once as the test set. The performance metrics are averaged over the five folds [24]–[26]. The formula for calculating accuracy in each fold is.

$$\mathrm{CV}_{(K)} = \frac{1}{K}\sum_{i=1}^{K} \mathrm{Error}_i \tag{2}$$

Performance Comparison Analysis

Post-validation, the model's performance was assessed using metrics such as accuracy, precision, recall, and F-measure. Their respective formulae are [27]–[30].

$$\begin{gathered} Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)} \\ Pericision = \frac{TP}{(TP+FP)} \\ Recall = \frac{TP}{(TP+FN)} \\ F-measure \ = \frac{2(presisi \times recall)}{(presisi + recall)} \end{gathered} \tag{3}$$

Where TP, TN, FP and FN represent the numbers of true positives, true negatives, false positives, and false negatives, respectively.

3. Results and Discussion

Results

The investigation into the efficacy of a voting classifier for the multiclass classification of dry bean varieties yielded insightful results. Across the 5-fold cross-validation process, the performance metrics—accuracy, precision, recall, and F1-score—varied significantly, indicating the classifier's varying levels of success in correctly identifying the seven different registered dry bean varieties. The highest accuracy observed was 96.73% in the third fold, showcasing the model's potential under optimal conditions. Conversely, the lowest accuracy recorded was 31.23% in the fifth fold, highlighting challenges in certain scenarios. Precision, recall, and F1-scores mirrored this variability, with their peak performances also in the third fold, signifying a balanced prediction capability that aligns precision and recall effectively. The detailed results are presented in **Table 1** and visualized in **Figure 3** for a clearer understanding and comparison of the metrics across different iterations.

Table 1. Performance Metrics Across 5-Fold Cross-Validation for the Voting Classifier

K-n	**Metrics**			
	Accuracy	***Precision***	***Recall***	***F-Measure***
K-1	34.67%	32.41%	34.63%	24.39%
K-2	88.43%	89.56%	88.80%	88.98%
K-3	96.73%	96.50%	96.51%	96.60%
K-4	89.38%	90.30%	89.09%	90.26%
K-5	31.23%	48.52%	31.19%	25.55%
$\sum \boldsymbol{Avg}$	68.09%	71.46%	68.04%	65.16%

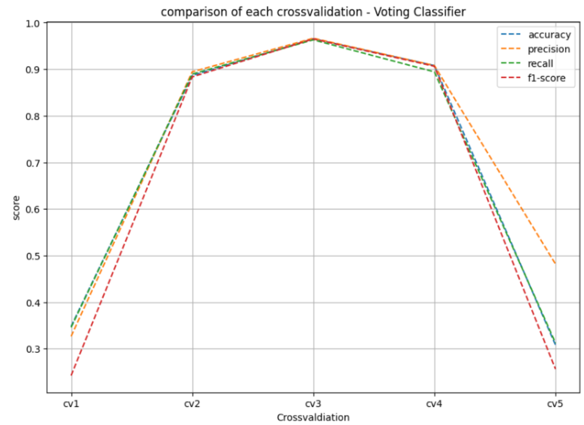

Figure 3. Visualisation Performance Metrics Across 5-Fold Cross-Validation for the Voting Classifier

The analysis of these metrics suggests that while the voting classifier has a strong foundational capability for classifying dry bean varieties, its performance is highly dependent on the specific data subset being evaluated. This variability is critical in understanding the classifier's application scope and areas requiring refinement.

Discussion

The results of this study reveal the complex nature of employing machine learning models for agricultural classification tasks. The significant variance in performance metrics across different folds suggests that the classifier's effectiveness is contingent upon the dataset's diversity and the distribution of features within. The peak performance in the third fold indicates that under certain conditions, the voting classifier can achieve high levels of accuracy and balance between precision and recall. However, the stark contrast in performance in other folds points to potential overfitting to specific features or underrepresentation of certain bean varieties in the training data.

Comparing these findings with previous research, it's evident that machine learning, and specifically ensemble methods like voting classifiers, hold promise for agricultural applications. Previous studies have also reported the effectiveness of ensemble methods in improving classification outcomes by leveraging the strengths of multiple learning algorithms. However, the performance variability observed in this study underscores the importance of diverse and representative training data to enhance model robustness.

The practical implications of these results are significant for the agricultural sector, particularly in automating and improving the efficiency of crop sorting and classification processes. Implementing such a classifier could reduce manual labor, increase processing speed, and potentially lead to more accurate market grading of dry bean varieties. Nevertheless, this research is not without limitations. The variability in model performance across folds indicates a sensitivity to data distribution, which could limit its practical application without further refinement. Moreover, the reliance on a fixed set of features extracted from the image data may not capture the full complexity of bean variety characteristics.

For future research, it's recommended to explore the inclusion of additional, perhaps more nuanced, features that could improve the classifier's ability to distinguish between closely related bean varieties. Further, investigating the model's performance on a larger, more diverse dataset could provide insights into enhancing its generalizability. Finally, comparing the voting classifier's performance with other ensemble methods or advanced machine learning models could identify pathways to further optimize classification accuracy and reliability in agricultural applications.

4. Conclusion

This research embarked on evaluating the efficacy of a voting classifier in the multiclass classification of dry bean varieties, employing a methodology that combined Decision Tree, Logistic Regression, and Gaussian Naive Bayes classifiers. The investigation revealed notable variability in model performance across different folds of cross-validation, with accuracy rates ranging from as high as 96.73% to as low as 31.23%. This variability underscored the model's potential in certain conditions while highlighting areas needing improvement in others. Through the lens of precision, recall, and F1-scores, the study further emphasized the importance of balancing these metrics to achieve a robust classification model. The discussion illuminated the critical role of data diversity and feature distribution in influencing the classifier's effectiveness, aligning with previous research that supports the promise of ensemble methods in agricultural applications.

The findings answer the research question affirmatively; a voting classifier can effectively classify dry bean varieties, particularly under optimal conditions highlighted by the highest achieved metrics. This research contributes to the agricultural informatics field by demonstrating the applicability of ensemble learning methods in crop variety classification, an area with significant implications for food security and agricultural efficiency. Given the observed variability in performance, it is recommended that future research explore the incorporation of additional features and the utilization of larger, more diverse datasets to enhance model generalization. Moreover, comparative studies with other ensemble methods or advanced machine learning models may uncover opportunities to further optimize the classification process. In practice, integrating such improved models into agricultural processing can lead to significant advancements in crop sorting, grading, and overall management systems, driving efficiency and productivity in the sector.

References:

[1] M. A. Febriantono, “Classification of multiclass imbalanced data using cost-sensitive decision tree c5.0,” *IAES Int. J. Artif. Intell.*, vol. 9, no. 1, pp. 65–72, 2020, doi: 10.11591/ijai.v9.i1.pp65-72.

[2] R. Panneerselvam, “Multi-Class Skin Cancer Classification Using a Hybrid Dynamic Salp Swarm Algorithm and Weighted Extreme Learning Machines with Transfer Learning,” *Acta Inform. Pragensia*, vol. 12, no. 1, pp. 141–159, 2023, doi: 10.18267/j.aip.211.

[3] E. Saad, “Predicting death risk analysis in fully vaccinated people using novel extreme regression-voting classifier,” *Digit. Heal.*, vol. 8, 2022, doi: 10.1177/20552076221109530.

[4] S. K. Jha, “Breast Cancer Prediction Using Voting Classifier Model,” *AI-Centric Model. Anal. Concepts, Technol. Appl.*, pp. 132–144, 2023, doi: 10.1201/9781003400110-8.

[5] M. Aqib, “Classification of Edge Applications using Decision Tree, K-NN, & SVM Classifier,” *2022 IEEE Students Conf. Eng. Syst. SCES 2022*, 2022, doi: 10.1109/SCES55490.2022.9887690.

[6] C. Xi, “Effectiveness of Newmark-based sampling strategy for coseismic landslide susceptibility mapping using deep learning, support vector machine, and logistic regression,” *Bull. Eng. Geol. Environ.*, vol. 81, no. 5, 2022, doi: 10.1007/s10064-022-02664-5.

[7] M. V Anand, “Gaussian Naïve Bayes Algorithm: A Reliable Technique Involved in the Assortment of the Segregation in Cancer,” *Mob. Inf. Syst.*, vol. 2022, 2022, doi: 10.1155/2022/2436946.

[8] B. Cao, “Performance analysis and comparison of PoW, PoS and DAG based blockchains,” *Digit. Commun. Networks*, vol. 6, no. 4, pp. 480–485, 2020, doi: 10.1016/j.dcan.2019.12.001.

[9] S. Rahman, “Performance analysis of boosting classifiers in recognizing activities of daily living,” *Int. J. Environ. Res. Public Health*, vol. 17, no. 3, 2020, doi: 10.3390/ijerph17031082.

[10] A. A. Ewees, “Performance analysis of Chaotic Multi-Verse Harris Hawks Optimization: A case study on solving engineering problems,” *Eng. Appl. Artif. Intell.*, vol. 88, 2020, doi: 10.1016/j.engappai.2019.103370.

[11] A. Anitha, “Disease prediction and knowledge extraction in banana crop cultivation using decision tree classifiers,” *Int. J. Bus. Intell. Data Min.*, vol. 20, no. 1, pp. 107–120, 2022, doi: 10.1504/IJBIDM.2022.119957.

[12] T. R. Sahoo, “Decision tree classifier based on topological characteristics of subgraph for the mining of protein complexes from large scale PPI networks,” *Comput. Biol. Chem.*, vol. 106, 2023, doi: 10.1016/j.compbiolchem.2023.107935.

[13] F. Huang, “Logistic Regression Fitting of Rainfall-Induced Landslide Occurrence Probability and Continuous Landslide Hazard Prediction Modelling,” *Diqiu Kexue - Zhongguo Dizhi Daxue Xuebao/Earth Sci. - J. China Univ. Geosci.*, vol. 47, no. 12, pp. 4609–4628, 2022, doi: 10.3799/dqkx.2021.164.

[14] S. Naiem, “Enhancing the Efficiency of Gaussian Naïve Bayes Machine Learning Classifier in the Detection

of DDOS in Cloud Computing," *IEEE Access*, vol. 11, pp. 124597–124608, 2023, doi: 10.1109/ACCESS.2023.3328951.

[15] S. W. Sharshir, "Performance enhancement of stepped double slope solar still by using nanoparticles and linen wicks: Energy, exergy and economic analysis," *Appl. Therm. Eng.*, vol. 174, 2020, doi: 10.1016/j.applthermaleng.2020.115278.

[16] Z. Zhao, "Logistic Regression Analysis of Risk Factors and Improvement of Clinical Treatment of Traumatic Arthritis after Total Hip Arthroplasty (THA) in the Treatment of Acetabular Fractures," *Comput. Math. Methods Med.*, vol. 2022, 2022, doi: 10.1155/2022/7891007.

[17] I. F. Hanbal, "Classifying Wastes Using Random Forests, Gaussian Naïve Bayes, Support Vector Machine and Multilayer Perceptron," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 803, no. 1, 2020, doi: 10.1088/1757-899X/803/1/012017.

[18] R. Kakkar, R. Gupta, M. S. Obaidiat, N. K. Jadav, and S. Tanwar, "Majority Voting-based Consensus Mechanism for UAVs Decision Making in Battlefield," in *2023 International Conference on Computer, Information and Telecommunication Systems (CITS)*, Jul. 2023, pp. 01–05, doi: 10.1109/CITS58301.2023.10188747.

[19] D. Widyawati, A. Faradibah, and ..., "Comparison Analysis of Classification Model Performance in Lung Cancer Prediction Using Decision Tree, Naive Bayes, and Support Vector Machine," *Indones. J. ...*, 2023, doi: 10.56705/ijodas.v4i2.76.

[20] J. Zhang, "Multi-class object detection using faster R-CNN and estimation of shaking locations for automated shake-and-catch apple harvesting," *Comput. Electron. Agric.*, vol. 173, 2020, doi: 10.1016/j.compag.2020.105384.

[21] Y. Nie, "Deep Melanoma classification with K-Fold Cross-Validation for Process optimization," *IEEE Med. Meas. Appl. MeMeA 2020 - Conf. Proc.*, 2020, doi: 10.1109/MeMeA49120.2020.9137222.

[22] M. H. D. M. Ribeiro, "Ensemble approach based on bagging, boosting and stacking for short-term prediction in agribusiness time series," *Appl. Soft Comput. J.*, vol. 86, 2020, doi: 10.1016/j.asoc.2019.105837.

[23] S. Ortiz-Toquero, "Classification of Keratoconus Based on Anterior Corneal High-order Aberrations: A Cross-validation Study," *Optom. Vis. Sci.*, vol. 97, no. 3, pp. 169–177, 2020, doi: 10.1097/OPX.0000000000001489.

[24] O. Karal, "Performance comparison of different kernel functions in SVM for different k value in k-fold cross-validation," *Proc. - 2020 Innov. Intell. Syst. Appl. Conf. ASYU 2020*, 2020, doi: 10.1109/ASYU50717.2020.9259880.

[25] T. A. Reist, "Cross validation of aerodynamic shape optimization methodologies for aircraft wing-body optimization," *AIAA J.*, vol. 58, no. 6, pp. 2581–2595, 2020, doi: 10.2514/1.J059091.

[26] M. Rafało, "Cross validation methods: Analysis based on diagnostics of thyroid cancer metastasis," *ICT Express*, vol. 8, no. 2, pp. 183–188, 2022, doi: 10.1016/j.icte.2021.05.001.

[27] M. A. A. Walid, "Adapted Deep Ensemble Learning-Based Voting Classifier for Osteosarcoma Cancer Classification," *Diagnostics*, vol. 13, no. 19, 2023, doi: 10.3390/diagnostics13193155.

[28] S. Kumari, D. Kumar, and M. Mittal, "An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier," *Int. J. Cogn. Comput. Eng.*, vol. 2, pp. 40–46, Jun. 2021, doi: 10.1016/j.ijcce.2021.01.001.

[29] K. Sen, "Heart Disease Prediction Using a Soft Voting Ensemble of Gradient Boosting Models, RandomForest, and Gaussian Naive Bayes," *2023 4th Int. Conf. Emerg. Technol. INCET 2023*, 2023, doi: 10.1109/INCET57972.2023.10170399.

[30] H. Tella, "Bagging and Voting Deep Learning Ensemble Methods for Binary Classifications of Solar Panel Cells Defects," *2023 20th International Multi-Conference on Systems, Signals and Devices, SSD 2023*. pp. 104–108, 2023, doi: 10.1109/SSD58187.2023.10411247.