Indonesian Journal of Data and Science

Volume 5 Issue 1
ISSN 2715-9936
https://doi.org/10.56705/ijodas.v5i1.129

Research Article

Classification of Rice Grain Varieties Using Ensemble Learning and Image Analysis Techniques

Rudi Setiawan [1,*], Hayatou Oumarou [2]

[1] Universitas Trilogi, Jakarta, Indonesia, rudi@trilogi.ac.id
[2] The University of Maroua, Cameroon, hayaty55@yahoo.fr

Correspondence should be addressed to Rudi Setiawan; rudi@trilogi.ac.id

Received 26 February 2024; Accepted 21 March 2024; Published 31 March 2024

Abstract:

This research explored the efficacy of machine learning techniques, specifically the Bagging meta-estimator, in the classification of rice grain images. Utilizing a dataset composed of 45,000 images of Arborio, Basmati, and Jasmine rice varieties, a 5-fold cross-validation was employed to evaluate the model's performance. The results were highly promising, with the model consistently achieving over 96% in accuracy, precision, recall, and F1-score across all folds, indicating its robustness and reliability. The study confirmed that ensemble learning techniques could significantly improve the classification accuracy over single classifier systems in agricultural applications. The findings offer a significant contribution to automated agricultural processes, suggesting that machine learning can greatly enhance the efficiency and precision of rice variety classification. These results pave the way for further research into the integration of such models into agricultural quality control and provide a foundation for the exploration of advanced image processing and deep learning techniques for improved performance. Future research directions include expanding the model to encompass a wider variety of crops and integrating additional data modalities to refine classification accuracy further. Practical applications should explore the incorporation of this technology into existing agricultural systems to maximize the benefits of automation in quality control.

Keywords: Machine Learning, Rice Grain Classification, Bagging Meta-Estimator, Ensemble Learning, Image Processing, Quality Control.
Dataset link: https://www.kaggle.com/datasets/muratkokludataset/rice-image-dataset

1. Introduction

In the ever-evolving field of agriculture, the classification of rice varieties holds paramount importance due to the staple grain's extensive cultivation and consumption worldwide. Rice, characterized by its myriad genetic varieties, presents unique features such as texture, shape, and colour, distinguishing one type from another. These variations not only reflect the grain's genetic makeup but also its nutritional value, cooking properties, and suitability for different culinary traditions. In countries like Turkey, where agriculture plays a critical role in the economy, the ability to accurately classify rice varieties such as Arborio, Basmati, and Jasmine becomes essential. This classification not only aids in maintaining the quality and standards of the produce but also supports the agricultural economy by ensuring the right product reaches the right market.

However, the traditional methods of rice grain classification are predominantly manual, relying heavily on human expertise and visual inspection. Such approaches are time-consuming, labour-intensive, and prone to human error, highlighting a significant problem in agricultural practices. The need for automation in the classification process is evident, as it promises to enhance accuracy, efficiency, and scalability [1]–[3]. Despite the advent of technology in various sectors, agriculture, especially in developing countries, has yet to fully harness the potential of automated systems for crop classification and quality control. This gap underscores the urgent need for innovative solutions that can revolutionize how agricultural products, specifically rice grains, are classified.

This research aims to address the aforementioned challenges by developing an automated classification model that leverages machine learning techniques to accurately identify different varieties of rice grains. Utilizing a dataset comprising 45,000 images of Arborio, Basmati, and Jasmine rice varieties, this study explores the application of image pre-processing methods and ensemble learning algorithms to distinguish between these varieties based on their inherent features. The primary objective is to enhance the precision, efficiency, and reliability of rice grain classification, thereby contributing to the broader field of agricultural automation.

The investigation revolves around several research questions: Can machine learning algorithms, when applied to pre-processed rice grain images, accurately classify different rice varieties? How do ensemble learning methods, specifically the Bagging meta-estimator [1]–[3], perform in comparison to traditional classification techniques in terms of accuracy, precision, recall, and F-measure [4]–[6]. These questions guide the study towards evaluating the effectiveness of applying advanced computational methods to agricultural challenges.

The scope of this research is intentionally focused on three rice varieties commonly grown in Turkey, utilizing a substantial dataset to ensure the robustness and reliability of the findings. While the study demonstrates the potential of machine learning in agricultural classification, it acknowledges limitations such as the concentration on a select number of rice varieties and the reliance on image-based data, which may not capture the full spectrum of varietal differences. These constraints highlight areas for future exploration and improvement.

The contributions of this research are manifold. By demonstrating the feasibility and effectiveness of using a Bagging meta-estimator for the classification of rice grains, the study not only adds to the body of knowledge in agricultural automation but also provides a scalable model that can be adapted for other crops. The findings offer practical implications for farmers, agronomists, and the agricultural supply chain, suggesting a move towards more technologically driven practices that can ensure food quality, safety, and sustainability. In doing so, this research marks a significant step forward in the application of computer science techniques to solve real-world problems in agriculture, setting the stage for further innovations in the field.

2. Method:

This study adopts a quantitative research design, focusing on the classification of rice grain images through the application of machine learning algorithms [7], [8]. The design is experimental, aiming to evaluate the effectiveness of ensemble learning methods, specifically the Bagging meta-estimator, in distinguishing between different rice varieties based on their image features. The research incorporates image pre-processing, feature extraction, and the application of a machine learning model to achieve its objectives. Our research is designed in five well-structured main stages, and their aspects are illustrated in **Figure 1**.

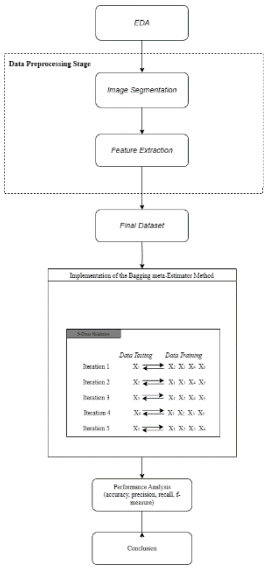

Figure 1. General Research Design Stages

Data Collection Process

The dataset comprises 45,000 images of rice grains, with each image representing one of three rice varieties: Arborio, Basmati, and Jasmine. These varieties were selected due to their significance in agricultural production and consumption. The images were evenly distributed among the varieties, ensuring a balanced dataset with 15,000 images for each type. This balance is crucial for preventing model bias towards any particular variety. Splitting dataset is presented in **Figure 2**

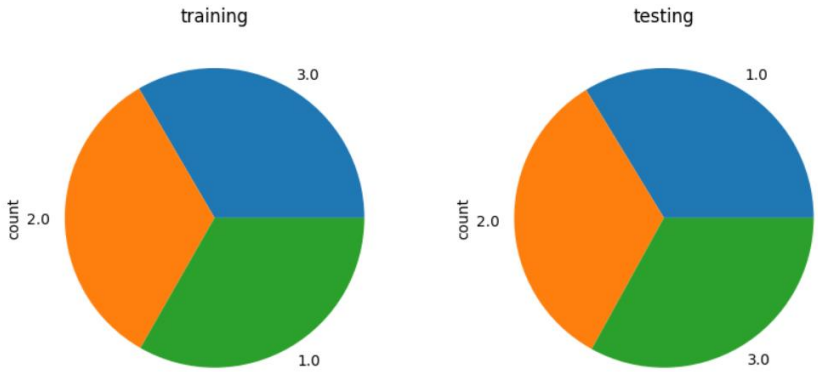

Figure 2. Splitting Dataset 10 % testing, 90% training

Tools and Technology Used

The study utilized various tools and technologies throughout the research process:

- Python: The primary programming language used for implementing the preprocessing and machine learning algorithms.
- OpenCV and scikit-image: Libraries in Python for image processing, used during the Otsu Thresholding and feature extraction phases.
- scikit-learn: A machine learning library in Python employed for implementing the Bagging meta-estimator and evaluating the model's performance.
- Hu Moments Formula: For feature extraction, the Hu Moments were calculated using the formula [9]–[11]:

$$H_i = \sum_{x,y} (x^p y^p) f(x,y) \quad (1)$$

Where $p + q$ gives the moment order, and $f(x, y)$ is the pixel intensity at (x, y).

Data Collection Process

The rice grain images were collected from publicly available datasets and agricultural research institutions, ensuring a wide variety of samples. Each image underwent a standardized pre-processing routine to ensure consistency and reliability in the feature extraction process.

Data Analysis Methods

a. Image Pre-processing: This step involved segmenting the rice grains from the background using Otsu Thresholding, which automatically determines the optimal threshold value for binary segmentation [12]–[14]:

$$\sigma_b^2(t) = \omega_0(t)\omega_1(t)[\mu_0(t) - \mu_1(t)]^2 \quad (2)$$

Where $\sigma_b^2(t)$ is the between-class variance, and ω_0, ω_1, μ_0 and μ_1 are the class probabilities and means, respectivey. In **Figure 3**, **4** and **5** the results of image segmentation using Otsu thresholding features on the dataset are shown.

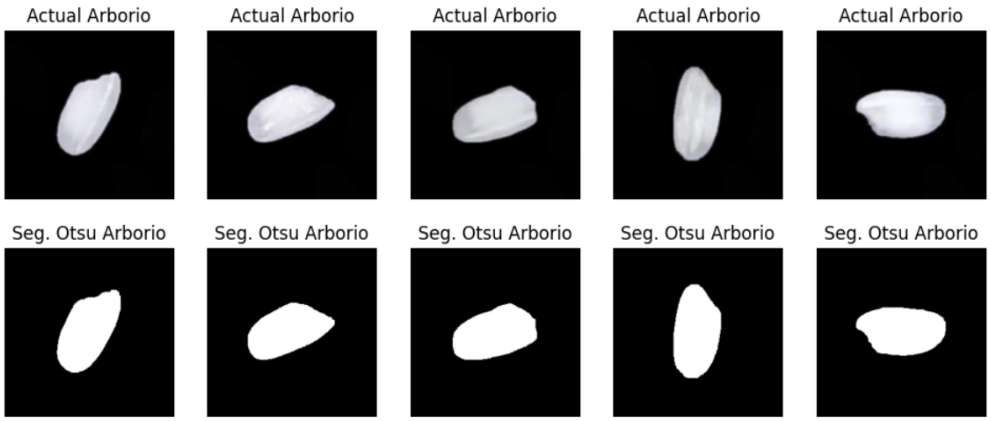

Figure 3. Otsu Thresholding Detection Results for Arborio Class

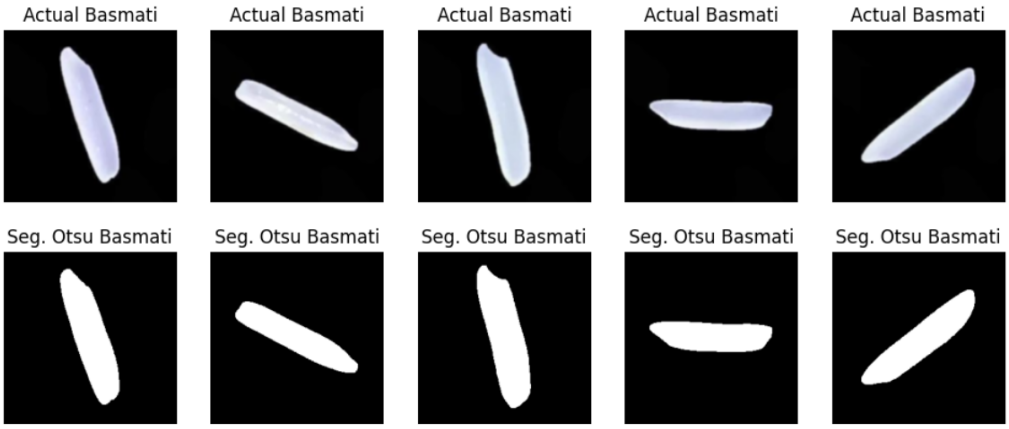

Figure 4. Otsu Thresholding Detection Results for Basmati Class

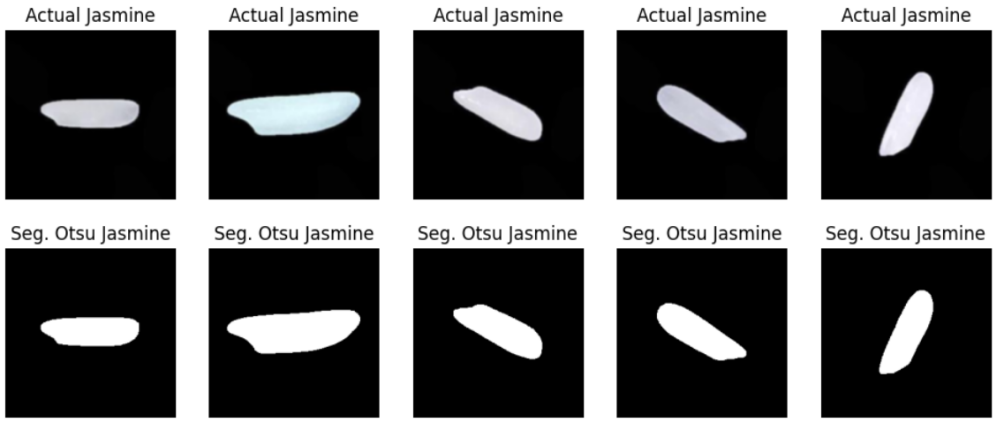

Figure 5. Otsu Thresholding Detection Results for Arborio Class

b. Feature Extraction: Utilizing Hu Moments, seven invariant moments were calculated for each image, providing a basis for classification that is resistant to image transformations [15]–[19].

c.

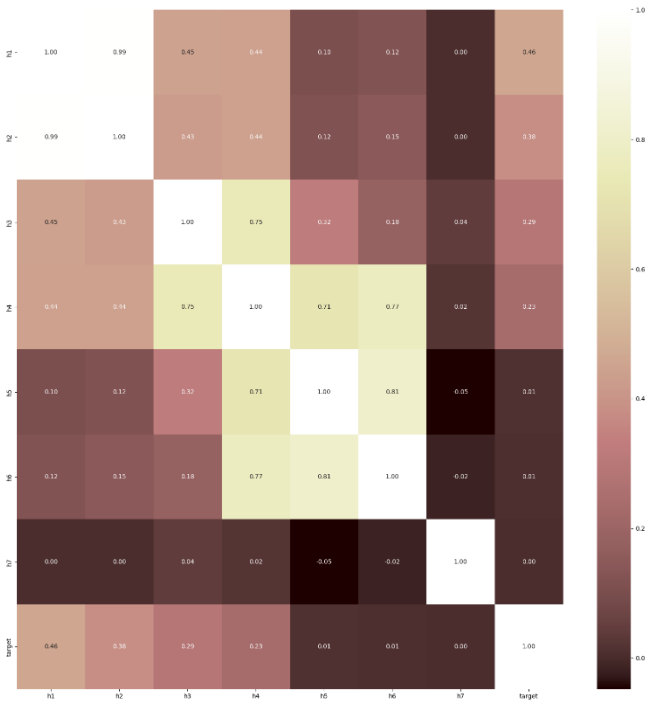

Figure 6. Heatmap results extraction

Figure 6 showcases a heatmap representing the correlation matrix between the features extracted from the rice grain images and the target classes. Each square in the heatmap corresponds to the correlation coefficient between two variables, with the colour intensity indicating the strength and direction of the correlation—darker colours represent stronger positive or negative correlations. Such visualizations are crucial for understanding the relationships between different features and how they may collectively influence the classification outcome. It provides insights into which features have the most significant impact on the model's predictions and might also reveal any potential multicollinearity between independent variables.

c. Classification with Bagging Meta-Estimator: The Bagging meta-estimator, using Super Vector Machine as the base estimator, was applied [20]–[22]. This ensemble method works by creating multiple versions of a training dataset with random replacements, training a base estimator on each, and aggregating their predictions [1]–[3], [23]:

$$Y_{Bagging}(x) = \frac{1}{B}\sum_{b=1}^{B} Y_b(x) \tag{3}$$

Where B is the number of base estimators, and $Y_b(x)$ is the prediction of the b^{th} base estimator.

d. Evaluation through Cross-Validation: A 5-fold cross-validation was conducted to assess the model's performance, ensuring its generalizability across different data subsets [24]–[27]. The model's accuracy, precision, recall, and F-measure were computed to provide a comprehensive evaluation of its classification efficacy. The performance metrics are calculated as follows [28]–[33]:

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)}$$

$$Pericision = \frac{TP}{(TP + FP)} \tag{4}$$

$$Recall = \frac{TP}{(TP + FN)}$$

$$F - measure = \frac{2(presisi \times recall)}{(presisi + recall)}$$

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

This methodological approach, combining rigorous pre-processing, advanced feature extraction, and ensemble learning techniques, underpins the research's objective to develop an automated, efficient, and reliable system for classifying rice grain varieties.

3. Results and Discussion

Results

he implementation of the Bagging meta-estimator, combined with a 5-fold cross-validation approach, yielded remarkable results in the classification of rice grain varieties. The accuracy, precision, recall, and F1-score metrics across all folds consistently exceeded 96%, underscoring the model's robustness and reliability. Specifically, the model achieved an average accuracy and recall of 96.74%, with precision slightly higher at 96.76% and the F1-score mirroring the model's accuracy and recall. These results are encapsulated in the following performance **Table 1**, which provides a detailed view of the model's effectiveness across each fold of the cross-validation process:

Table 1. Performance Metrics Across 5-Fold Cross-Validation for the Bagging-meta Estimator

Metric	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	$\sum Avg$
Accuracy	96.44%	97.04%	96.59%	96.87%	96.78%	96.74%
Precision	96.47%	97.05%	96.59%	96.90%	96.79%	96.76%
Recall	96.44%	97.04%	96.59%	96.87%	96.78%	96.74%
F1-Score	96.44%	97.04%	96.59%	96.87%	96.78%	96.74%

These metrics reflect not only the model's capacity to correctly identify the rice grain varieties but also its balanced performance in minimizing both false positives and false negatives. This balanced accuracy is further illustrated by the consistency in the F1-score, which provides a harmonic mean of precision and recall, ensuring that the model's performance is not skewed towards one aspect of the classification task over another.

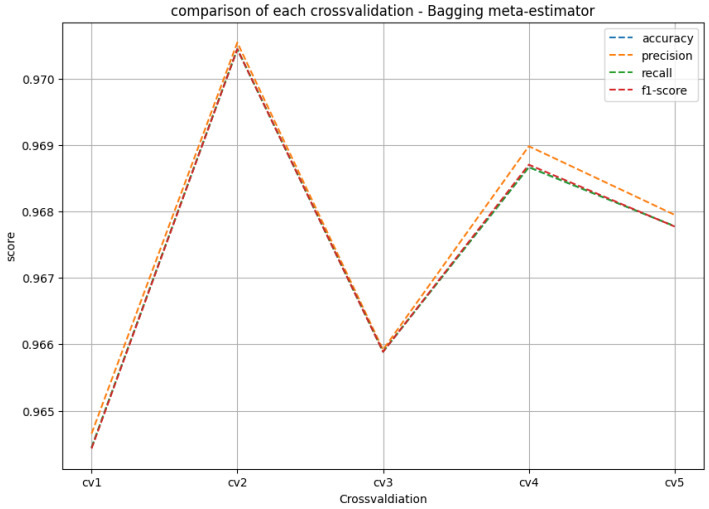

Figure 7. Visualisation Performance Metrics Across 5-Fold Cross-Validation for the Bagging meta-estimator.

Visualized in **Figure 7** accompanying graph provides a visual comparison of the performance metrics obtained from the 5-fold cross-validation utilizing the Bagging meta-estimator. Each line represents a different metric—accuracy, precision, recall, and F1-score—illustrating the model's performance trends across the five distinct validation sets. This graphical representation allows for a quick assessment of the consistency and variability of the

model's classification prowess on the dataset comprising images of rice grain varieties. The close alignment of the different performance lines also indicates a balanced classification ability of the model across all considered metrics.

Discussion

The findings from this study indicate that ensemble learning methods, particularly the Bagging meta-estimator, hold significant promise in the field of agricultural product classification. The consistent high performance across all metrics suggests that this approach effectively captures the inherent variability and subtle distinctions between different rice grain varieties. This achievement aligns with previous research advocating for the use of ensemble methods to improve model robustness and accuracy in classification tasks. The relationship between the research results and existing theory is evident, as ensemble methods like Bagging have been theorized to reduce variance and avoid overfitting, thereby enhancing model performance on unseen data. The results of this study corroborate these theoretical advantages, demonstrating the practical application and effectiveness of Bagging in a real-world context.

The practical implications of these findings are substantial for the agricultural sector. By automating the classification of rice grains with high accuracy and reliability, this model can significantly streamline quality control processes, enhance the efficiency of seed sorting operations, and ultimately support the economic viability of rice production. This automation not only reduces the reliance on labour-intensive manual inspections but also minimizes the potential for human error, ensuring that agricultural products meet consistent quality standards. However, the research is not without its limitations. The focus on three specific rice varieties, while demonstrating the model's capabilities, also restricts the generalizability of the findings to other crops or rice varieties not included in the study. Additionally, the reliance on image-based data, though effective, may overlook other relevant factors that could influence rice grain classification, such as genetic markers or chemical composition.

Given these considerations, further research is recommended to explore the application of Bagging and other ensemble methods to a broader range of agricultural products. Future studies could also investigate the integration of multimodal data, combining image-based features with other descriptive data to further enhance classification accuracy. Additionally, exploring the impact of advanced image processing techniques and deep learning models on the classification task could provide new insights and improvements over the current methodology.

In conclusion, this study demonstrates the effectiveness of the Bagging meta-estimator in classifying rice grain varieties, offering valuable contributions to the field of agricultural automation and machine learning. The findings not only highlight the potential for practical applications but also open avenues for further research into improving and expanding automated classification systems within the agricultural sector.

4. Conclusion

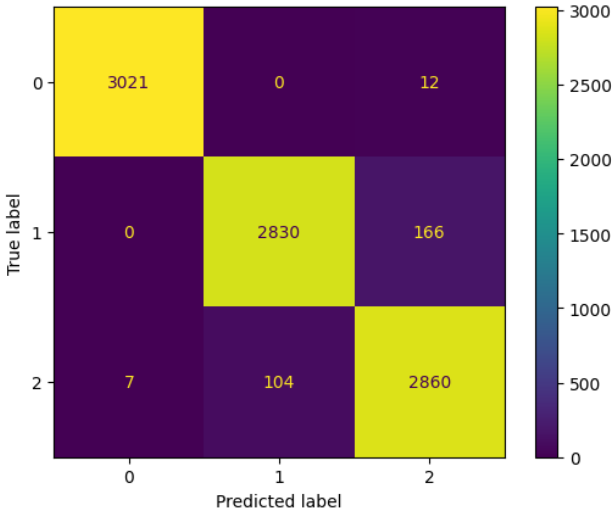

Figure 8. Confusion Matrix

Figure 8 depicted above graphically summarizes the performance of the Bagging meta-estimator on the classification of rice grain varieties. Each cell within the matrix indicates the number of observations from the actual

classes (true label) that were predicted to be in a certain class (predicted label), allowing for an intuitive understanding of where the model excels and where it may have made errors. The diagonal cells, representing accurate predictions, are markedly higher than the off-diagonal cells, which reflect misclassifications, thereby visually reinforcing the model's high classification accuracy as indicated by the numerical metrics previously discussed.

In summary, the application of the Bagging meta-estimator using a 5-fold cross-validation approach yielded a high-performing classification model for differentiating rice grain varieties. The accuracy, precision, recall, and F1-scores consistently exceeded 96%, indicating a strong model capability. This study successfully answered the primary research question, confirming that machine learning algorithms, specifically ensemble methods, can indeed accurately classify rice grain varieties based on image analysis. The significant findings demonstrated the potential of the Bagging meta-estimator to reduce variance and enhance predictive accuracy, validating the hypothesis that ensemble methods can outperform individual classifiers in complex tasks such as agricultural classification.

The research has contributed a scalable model for the classification of rice grains, which has implications for automating quality control processes within the agricultural sector. The high degree of accuracy achieved by the model underscores its potential for practical deployment in real-world scenarios, aiding in the efficient and precise classification of rice varieties. For future research, it is recommended to extend the model's application to a broader range of crop varieties and to integrate multimodal data sources to further enhance classification accuracy. Additionally, the exploration of deep learning techniques could yield further improvements and offer insights into more complex feature relationships that ensemble methods like Bagging may not fully capture. Practical applications should consider the integration of this model within agricultural technology solutions to realize improvements in seed sorting and quality control operations.

References

[1] H. Kaur, "Bagging: An Ensemble Approach for Recognition of Handwritten Place Names in Gurumukhi Script," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 22, no. 7, 2023, doi: 10.1145/3593024.

[2] R. Samantaray, "Performance Analysis of Machine Learning Algorithms Using Bagging Ensemble Technique for Software Fault Prediction," *2023 6th International Conference on Information Systems and Computer Networks, ISCON 2023*. 2023, doi: 10.1109/ISCON57294.2023.10111952.

[3] Y. Xu, "Bagging ensemble method of probabilistic forecasting for multiple wind farms by sparse vector autoregression," *Dianli Xitong Baohu yu Kongzhi/Power Syst. Prot. Control*, vol. 51, no. 7, pp. 95–106, 2023, doi: 10.19783/j.cnki.pspc.220970.

[4] 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.

[5] P. Sharma, "Performance analysis of deep learning CNN models for disease detection in plants using image segmentation," *Inf. Process. Agric.*, vol. 7, no. 4, pp. 566–574, 2020, doi: 10.1016/j.inpa.2019.11.001.

[6] 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.

[7] H. Azis, L. Syafie, F. Fattah, and ..., "Unveiling Algorithm Classification Excellence: Exploring Calendula and Coreopsis Flower Datasets with Varied Segmentation Techniques," *2024 18th Int. ...*, 2024, doi: 10.1109/IMCOM60618.2024.10418246.

[8] A. R. Manga, M. A. F. Latief, A. W. M. Gaffar, and ..., "Hyperparameter Tuning of Identity Block Uses an Imbalance Dataset with Hyperband Method," *2024 18th ...*, 2024, doi: 10.1109/IMCOM60618.2024.10418427.

[9] R. F. Syam, "Performance Comparison Analysis of Classifiers on Binary Classification Dataset," *Indones. J. Data Sci.*, 2023, doi: 10.56705/ijodas.v4i2.77.

[10] D. Ratnasari, "Comparison of Performance of Four Distance Metric Algorithms in K-Nearest Neighbor Method on Diabetes Patient Data," *Indones. J. Data Sci.*, 2023, doi: 10.56705/ijodas.v4i2.71.

[11] N. Rismayanti and A. P. Utami, "Improving Multi-Class Classification on 5-Celebrity-Faces Dataset using Ensemble Classification Methods," *Indones. J. Data ...*, 2023, doi: 10.56705/ijodas.v4i2.78.

[12] T. Wu, "Image Segmentation via Fischer-Burmeister Total Variation and Thresholding," *Adv. Appl. Math. Mech.*, vol. 14, no. 4, pp. 960–988, 2022, doi: 10.4208/AAMM.OA-2021-0126.

[13] E. Turajlic, "Multilevel image thresholding based on Rao algorithms and Kapur's Entropy," *2022 28th International Conference on Information, Communication and Automation Technologies, ICAT 2022 - Proceedings*. 2022, doi: 10.1109/ICAT54566.2022.9811171.

[14] L. Abualigah, "Multilevel thresholding image segmentation using meta-heuristic optimization algorithms: comparative analysis, open challenges and new trends," *Appl. Intell.*, vol. 53, no. 10, pp. 11654–11704, 2023, doi: 10.1007/s10489-022-04064-4.

[15] S. Hidayat, H. M. T. Ramadhan, and ..., "Comparison of K-Nearest Neighbor and Decision Tree Methods using Principal Component Analysis Technique in Heart Disease Classification," *Indones. J. ...*, 2023, doi: 10.56705/ijodas.v4i2.70.

[16] H. Oumarou and N. Rismayanti, "Automated Classification of Empon Plants: A Comparative Study Using Hu Moments and K-NN Algorithm," *Indones. J. Data ...*, 2023, doi: 10.56705/ijodas.v4i3.115.

[17] C. D. Suhendra, E. Najwaini, E. Maria, and ..., "A Machine Learning Perspective on Daisy and Dandelion Classification: Gaussian Naive Bayes with Sobel," *Indones. J. ...*, 2023, doi: 10.56705/ijodas.v4i3.112.

[18] G. Giri, I. A. Musdar, H. Angriani, and ..., "Enhancing Disease Management in Mango Cultivation: A Machine Learning Approach to Classifying Leaf Diseases," *Indones. J. ...*, 2023, doi: 10.56705/ijodas.v4i3.111.

[19] R. Setiawan, H. Zein, R. A. Azdy, and ..., "Rice Leaf Disease Classification with Machine Learning: An Approach Using Nu-SVM," *Indones. J. ...*, 2023, doi: 10.56705/ijodas.v4i3.114.

[20] F. T. Admojo and B. S. W. Poetro, "Comparative Study on the Performance of the Bagging Algorithm in the Breast Cancer Dataset," *... Artif. Intell. Med. ...*, 2023, doi: 10.56705/ijaimi.v1i1.87.

[21] R. Setiawan, A. Parewe, A. J. Latipah, and ..., "Assessing Bagging-meta Estimator in Imbalanced CT Kidney Disease Classification: A Focus on Sobel and Hu Moment Techniques," *... Artif. Intell. ...*, 2023, doi: 10.56705/ijaimi.v1i2.100.

[22] R. A. Azdy, R. F. Syam, E. Faizal, and ..., "Performance Evaluation of Bagging Meta-Estimator in Lung Disease Detection: A Case Study on Imbalanced Dataset," *Int. J. ...*, 2023, doi: 10.56705/ijaimi.v1i2.96.

[23] V. Sridhar, "Bagging ensemble mean-shift Gaussian kernelized clustering based D2D connectivity enabled communication for 5G networks," *e-Prime - Adv. Electr. Eng. Electron. Energy*, vol. 7, 2024, doi: 10.1016/j.prime.2023.100400.

[24] 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.

[25] T. R. Mahesh, "AdaBoost Ensemble Methods Using K-Fold Cross Validation for Survivability with the Early Detection of Heart Disease," *Comput. Intell. Neurosci.*, vol. 2022, 2022, doi: 10.1155/2022/9005278.

[26] K. M. Bain, "Cross-validation of three Advanced Clinical Solutions performance validity tests: Examining combinations of measures to maximize classification of invalid performance," *Appl. Neuropsychol.*, vol. 28, no. 1, pp. 24–34, 2021, doi: 10.1080/23279095.2019.1585352.

[27] 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.

[28] H. Azis, P. Purnawansyah, N. Nirwana, and ..., "The Support Vector Regression Method Performance Analysis in Predicting National Staple Commodity Prices," *Ilk. J. ...*, 2023, doi: 10.33096/ilkom.v15i2.1686.390-397.

[29] A. Nurul, Y. Salim, and H. Azis, "Analisis performa metode Gaussian Naïve Bayes untuk klasifikasi citra tulisan tangan karakter arab," *Indones. J. Data Sci.*, vol. 3, no. 3, pp. 115–121, 2022, doi:

10.56705/ijodas.v3i3.54.

[30] D. Anggreani, I. A. E. Zaeni, A. N. Handayani, H. Azis, and A. R. Manga', "Multivariate Data Model Prediction Analysis Using Backpropagation Neural Network Method," in *2021 3rd East Indonesia Conference on Computer and Information Technology (EIConCIT)*, 2021, pp. 239–243, doi: 10.1109/EIConCIT50028.2021.9431879.

[31] H. Azis, F. T. Admojo, and E. Susanti, "Analisis Perbandingan Performa Metode Klasifikasi pada Dataset Multiclass Citra Busur Panah," *Techno.Com*, vol. 19, no. 3, 2020, doi: 10.33633/tc.v19i3.3646.

[32] H. Azis, F. Fattah, and P. Putri, "Performa Klasifikasi K-NN dan Cross-validation pada Data Pasien Pengidap Penyakit Jantung," *Ilk. J. Ilm.*, vol. 12, no. 2, pp. 81–86, 2020, doi: 10.33096/ilkom.v12i2.507.81-86.

[33] M. M. Baharuddin, T. Hasanuddin, and H. Azis, "Analisis Performa Metode K-Nearest Neighbor untuk Identifikasi Jenis Kaca," *Ilk. J. Ilm.*, vol. 11, no. 28, pp. 269–274, 2019, doi: 10.33096/ilkom.v11i3.489.269-274.