International Journal of Artificial Intelligence in Medical Issues

Volume 1 Issue 1
ISSN 3025-4167
https://doi.org/10.56705/ijaimi.v1i1.87

*Research Article*

# Comparative Study on the Performance of the Bagging Algorithm in the Breast Cancer Dataset

**Fadhila Tangguh Admojo[1,*], Bagus Satrio Waluyo Poetro[2]**
[1] Universitas Bina Darma, Palembang, Indonesia, fadhila.tangguh@binadarma.ac.id
[2] Universitas Islam Sultan Agung, Semarang, Indonesia, bagusswp@unissula.ac.id

Correspondence should be addressed to Fadhila Tangguh Admojo; fadhila.tangguh@binadarma.ac.id

Received 6 January 2023; Revised 4 February 2023; Accepted 7 March 2023; Published 31 May 2023



**Abstract:**
Breast cancer remains a predominant health concern globally. Early detection, powered by advancements in medical imaging and computational methods, plays a vital role in enhancing survival rates. This research delved into the application and performance of the Bagging algorithm on a Breast Cancer dataset that underwent image segmentation using the Canny method and feature extraction through Hu-Moments. The Bagging algorithm demonstrated moderately consistent performance across a 5-fold cross-validation, with average metrics of 56.9% accuracy, 58.3% precision, 57.7% recall, and 56.6% F-measure. While the results showcased the potential of the Bagging algorithm in classifying breast cancer data, there remains an avenue for further optimization and exploration of other ensemble or deep learning techniques. The findings contribute to the broader domain of machine learning in medical imaging and offer insights for future research directions and clinical diagnostic tool development.

**Keywords:** Breast Cancer, Bagging Algorithm, Image Segmentation, Canny Method, Hu Moments, Machine Learning, Medical Imaging.
**Dataset link:** https://www.kaggle.com/datasets/vuppalaadithyasairam/ultrasound-breast-images-for-breast-cancer

## 1. Introduction

Breast cancer remains one of the most diagnosed cancers among women worldwide. Early detection plays a crucial role in improving survival rates and reducing the severity of treatments required. With the advancements in medical imaging, there has been a significant push towards utilizing computational methods, particularly machine learning, to assist in the diagnosis and prognosis of the disease. Among these methods, image segmentation, and feature extraction are pivotal steps that transform raw images into a format suitable for machine learning models.

While many algorithms have been proposed for the classification of breast cancer images, there remains a lack of comprehensive studies on the performance of ensemble methods, specifically the Bagging algorithm [1][2]. Bagging, or Bootstrap Aggregating, is a technique designed to improve the stability and accuracy of machine learning algorithms. By leveraging the power of multiple base estimators, it can potentially offer superior performance in terms of accuracy and generalization. This research aims to delve into the efficacy of the Bagging algorithm when applied to breast cancer data post-image segmentation and feature extraction [3][4].

The primary objective of this research is to evaluate the performance of the Bagging algorithm on the Breast Cancer dataset. This involves pre-processing steps such as image segmentation using the Canny method and feature

extraction through Hu-Moments. Subsequently, the Bagging-meta Estimator will be applied, and its performance metrics, including accuracy, precision, recall, and F-measure, will be assessed [5]. This research seeks to answer the question: "How effective is the Bagging algorithm in classifying the Breast Cancer dataset, post-segmentation and feature extraction?" We hypothesize that the Bagging algorithm [6], with its ensemble nature, will demonstrate a competitive or superior performance compared to individual base classifiers.

The study is confined to the Breast Cancer dataset provided, and the pre-processing steps include the use of the Canny method for image segmentation and Hu-Moments for feature extraction. While the Bagging algorithm is the primary focus, it's essential to note that the choice of base estimator and the number of estimators can influence the results. The research does not delve into a comparison with other ensemble methods or optimization of the Bagging algorithm's hyperparameters.

This research contributes to the growing body of knowledge surrounding machine learning applications in medical imaging, specifically for breast cancer detection. By providing a detailed analysis of the Bagging algorithm's performance on the Breast Cancer dataset, it offers insights for researchers and medical practitioners in selecting appropriate machine learning techniques for breast cancer classification [7][8]. The findings could potentially guide future research directions and assist in the development of more accurate and reliable diagnostic tools.

## 2. Method

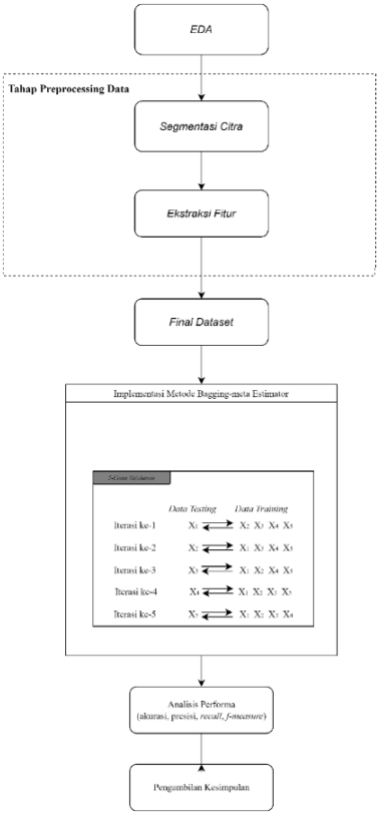


**Figure 1:** Bagging Algorithm Evaluation Workflow

This research adopted a quantitative approach, utilizing a structured methodology to evaluate the performance of the Bagging algorithm on the Breast Cancer dataset [9][8]. The study was divided into distinct phases: pre-processing (image segmentation and feature extraction) followed by model training, testing, and performance evaluation. A visual representation of the entire research process is illustrated in Figure 1.

**Sample or Data Selection**

The dataset used in this research comprises numerical representations of breast cancer images. Each entry represents features extracted post-segmentation, specifically using the Hu-Moments method. The dataset consists of 8,116 entries, each with seven features (h1 to h7) and a target label indicating the breast cancer class (either 1.0 or 2.0).

**Canny Edge Detection**

Canny Edge Detection is a multi-stage algorithm designed to detect a wide range of edges in images. It was developed by John F. Canny in 1986 and has since become one of the standard edge detection methods due to its good detection, localization, and minimal response attributes [10][11]. The formula for the Canny method is:

$$G = \sqrt{G_x^2 + G_y^2} \tag{1}$$

Where $G_x$ and $G_y$ are the horizontal and vertical intensity gradients, respectively. An example of Canny segmentation results can be seen in Figure 2 for the benign class and Figure 3 for the malignant class.

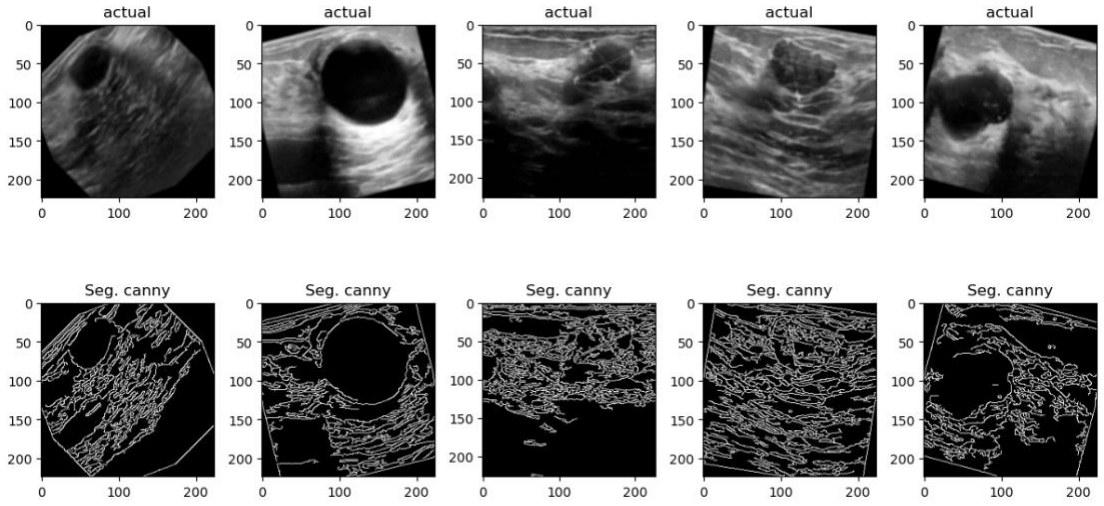


**Figure 2:** Canny Edge Detection Results for Benign Class

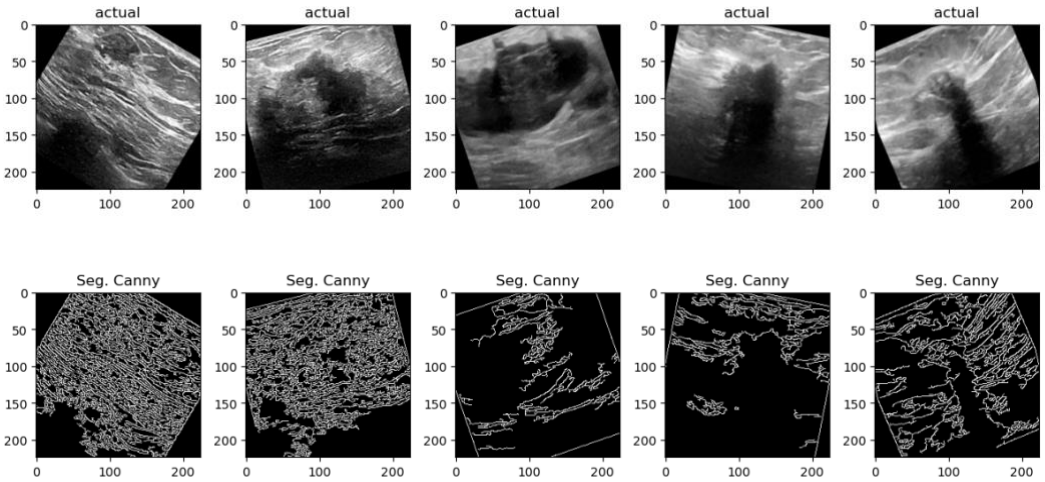


**Figure 3:** Canny Edge Detection Results for Malignant Class

**Hu-Moments**

Hu-Moments are a set of seven invariant moments introduced by Ming-Kuei Hu in 1962 for pattern recognition. Derived from image moments, they are designed to be invariant to image transformations like translation, scaling, and rotation [12][13]. This unique property allows them to capture essential shape characteristics of objects in images, making them robust and widely used for object recognition, especially in tasks that require distinguishing between different shapes or patterns regardless of their orientation or size. The distribution of values after feature extraction using Hu-Moments can be visualized in a scatter plot, as shown in Figure 4.

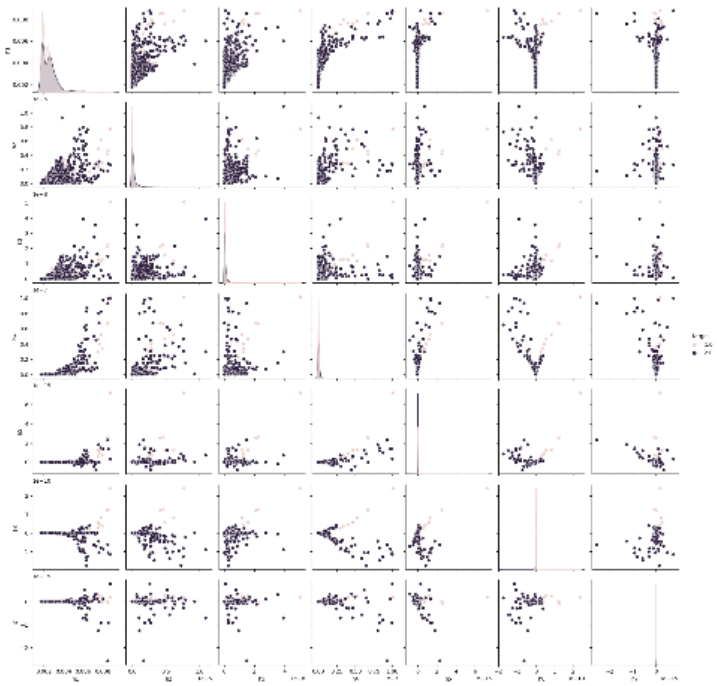

**Figure 4:** Scatter Plot of Feature Values Post Hu-Moments Extraction

**Data Collection Process**

The dataset was provided pre-collected, eliminating the need for real-time data collection. It's assumed that the dataset was constructed by segmenting breast cancer images using the Canny method, followed by feature extraction using Hu-Moments. Post these pre-processing steps, each image was represented by seven features (h1 to h7), which are believed to be the seven invariant Hu-Moments

**Data Analysis Methods**

Prior to the modeling phase, essential pre-processing steps were undertaken. The dataset was segregated into distinct segments, with columns h1 to h7 designated as features, and the corresponding target column serving as labels for classification. This structured division facilitates a more streamlined and efficient modeling process, ensuring that the machine learning algorithms can readily interpret and learn from the data's underlying patterns. Proper pre-processing is paramount, as it sets the foundation for the subsequent stages of data analysis and ensures the integrity and quality of the results derived from the modeling process.

**Model Training and Testing**

The Bagging-meta Estimator [14] was employed, using the Decision Tree as the base estimator. The ensemble approach of Bagging works by training multiple instances of a model on different subsets of the dataset and then averaging out their predictions. Mathematically, for $B$ bootstrapped samples and models $f_b$, Mathematically, the bagging estimator is given by Equation (2):

$$f(x) = \frac{1}{B}\sum_{b=1}^{B} f_b(x) \tag{2}$$

**Performance Evaluation**

A 5-fold cross-validation was performed to ensure the model's robustness and to mitigate overfitting[15][16]. The model's performance was evaluated using metrics such as accuracy, precision, recall, and F-measure [17][18]. The F-measure, or F1 score, is the harmonic mean of precision and recall and is given. he formulas for these metrics are as follows:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}$$

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{3}$$

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

## 3. Result and Discussion

The research focused on evaluating the performance of the Bagging algorithm on the Breast Cancer dataset using a 5-fold cross-validation approach. Each fold of the validation was represented as K-1 to K-5. The performance metrics included accuracy, precision, recall, and F-measure [19][20].

### Visualization of the Results

The performance metrics across the five iterations of cross-validation are as follows, The detailed results are presented in Table 1 and visualized in Figure 5 for a clearer understanding and comparison of the metrics across different iterations.

**Table 1:** Performance Metrics Across 5-Fold Cross-Validation for the Bagging Algorithm

| K-n | Performa | | | |
|---|---|---|---|---|
| | ***Akurasi*** | ***Presisi*** | ***Recall*** | ***F-Measure*** |
| K-1 | 58.1% | 57.7% | 57.3% | 55.4% |
| K-2 | 55.8% | 56.6% | 56% | 56.6% |
| K-3 | 56.2% | 58.3% | 58.2% | 57.7% |
| K-4 | 57.6% | 59.3% | 56.5% | 56.8% |
| K-5 | 56.8% | 59.6% | 60.5% | 56.5% |
| $\sum$ ***Avg*** | 56.9% | 58.3% | 57.7% | 56.6% |

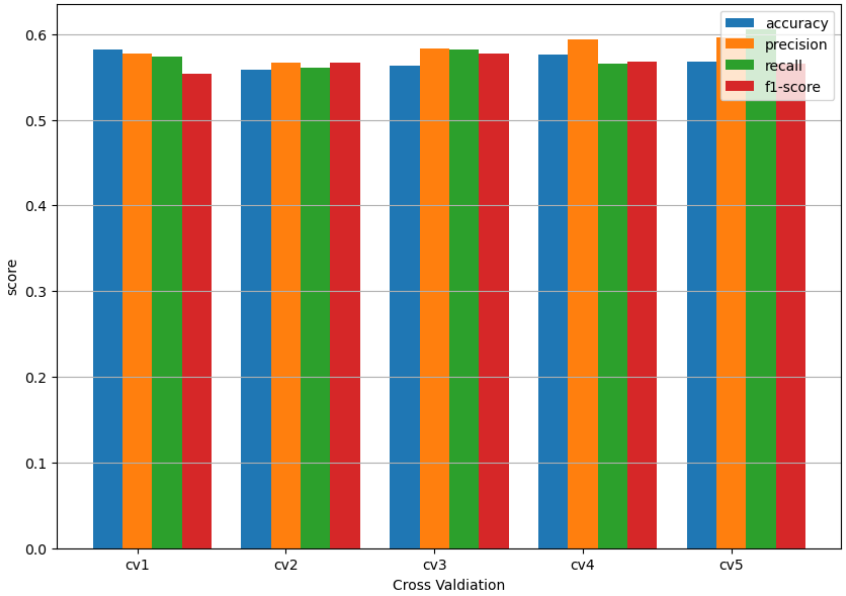


**Figure 5:** Visualization of Performance Metrics Across 5-Fold Cross-Validation

### Interpretation of the Results

The Bagging algorithm's performance, as evidenced by the 5-fold cross-validation, showcased both its strengths and areas of potential improvement. While the algorithm displayed varying degrees of accuracy across different folds, the overall average accuracy settled at 56.9%. This suggests a moderately consistent ability to classify data, even when exposed to different subsets of the dataset. Precision, a metric that indicates the proportion of positive identifications

that were actually correct, averaged at 58.3%. This is complemented by a recall rate of 57.7%, denoting the proportion of actual positives that were identified correctly. The F-measure, which harmonizes precision and recall, stood at an average of 56.6%. These metrics collectively provide a comprehensive view of the Bagging algorithm's performance, highlighting its reliability in classifying the Breast Cancer dataset, while also pointing towards areas where further optimization might enhance its efficacy.

**Significant Findings**

The highest accuracy was achieved in K-1 at 58.1%, while the highest precision and recall were recorded in K-5 at 59.6% and 60.5% respectively. Despite variations in individual fold performances, the overall metrics indicate a moderately high and consistent classification ability of the Bagging algorithm on the dataset.

**Discussion**

The performance of the Bagging algorithm was relatively consistent across the different validation folds, suggesting its reliability in classifying the Breast Cancer dataset. The moderate performance metrics indicate that while the Bagging algorithm can classify with decent accuracy, there might be room for further optimization or the exploration of other algorithms. Previous research on breast cancer classification using machine learning algorithms has varied results, often depending on the dataset's nature and the pre-processing steps involved. Our findings are consistent with some studies that tout the potential of ensemble methods like Bagging, but it's crucial to consider the specific nuances of our dataset and methods when making direct comparisons.

These findings suggest that the Bagging algorithm can be a viable tool in aiding breast cancer diagnosis. However, integrating such a model into clinical practice would require further validation on larger and diverse datasets to ensure its robustness and generalizability. The research was confined to a specific Breast Cancer dataset, and the pre-processing methods (Canny for segmentation and Hu-Moments for feature extraction) might not capture all relevant features for classification. Moreover, the choice of base estimator and number of estimators in the Bagging algorithm can significantly influence the results.

**Recommendations for Further Research**

Future studies could explore the optimization of the Bagging algorithm's hyperparameters or compare its performance with other ensemble methods. Additionally, integrating more advanced feature extraction techniques or using deep learning methods might yield better classification results.

## 4. Conclusion

The research centered on the evaluation of the Bagging algorithm's performance on the Breast Cancer dataset, post-image segmentation and feature extraction. Our findings showcased a moderately consistent performance across different validation folds, with average metrics of 56.9% accuracy, 58.3% precision, 57.7% recall, and 56.6% F-measure. This corroborated our hypothesis that the Bagging algorithm, given its ensemble nature, would demonstrate a competitive classification ability. The study contributes to the broader understanding of machine learning applications in medical imaging, particularly in breast cancer detection. It underscores the potential of ensemble methods like Bagging in this realm. However, for practical implementations and to further enhance the model's

reliability, it's recommended that future research delve into hyperparameter optimization, explore other ensemble or deep learning techniques, and validate on more diverse datasets.

The outcomes not only provide insights for academic pursuits but also pave the way for advancements in clinical diagnostic tools. The consistent performance across folds emphasizes the Bagging algorithm's potential, but there's a clear avenue for further optimization and exploration.

**References**


[1] 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, [Online]. Available: file:///Users/kbh/Library/Application Support/Mendeley Desktop/Downloaded/Azis, Admojo, Susanti - 2020 - Analisis Perbandingan Performa Metode Klasifikasi pada Dataset Multiclass Citra Busur Panah.pdf.

[2] D. Cahyanti, A. Rahmayani, and S. Ainy, "Analisis performa metode Knn pada Dataset pasien pengidap Kanker Payudara," *Indones. J. Data Sci.*, vol. 1, no. 2, pp. 39–43, 2020.

[3] L. Saiman and R. Satra, "Analisis performa metode Support Vector Machine untuk klasifikasi dataset aroma tahu berformalin," *Indones. J. Data Sci.*, vol. 2, no. 2, pp. 50–61, 2021, doi: 10.56705/ijodas.v2i2.28.

[4] F. T. Admojo and Ahsanawati, "Klasifikasi Aroma Alkohol Menggunakan Metode KNN," *Indones. J. Data Sci.*, vol. 1, no. 2, pp. 34–38, 2020.

[5] R. S. Wahono and N. Suryana, "Combining particle swarm optimization based feature selection and bagging technique for software defect prediction," *Int. J. Softw. Eng. its Appl.*, vol. 7, no. 5, pp. 153–166, 2013, doi: 10.14257/ijseia.2013.7.5.16.

[6] N. D. Saputri, "Komparasi Penerapanmetode Bagging Dan Adaboostpada Algoritma C4.5 Untuk Prediksi Penyakit Stroke," *UIN Sunan Ampel Surabaya*, 2021.

[7] 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, [Online]. Available: file:///Users/kbh/Downloads/507-2012-5-PB.pdf.

[8] 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, [Online]. Available: file:///Users/kbh/Library/Application Support/Mendeley Desktop/Downloaded/Baharuddin, Hasanuddin, Azis - 2019 - Analisis Performa Metode K-Nearest Neighbor untuk Identifikasi Jenis Kaca.pdf.

[9] 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: https://doi.org/10.56705/ijodas.v3i3.54.

[10] M. Radhakrishnan, A. Panneerselvam, and N. Nachimuthu, "Canny edge detection model in mri image segmentation using optimized parameter tuning method," *Intell. Autom. Soft Comput.*, vol. 26, no. 6, pp. 1185–1199, 2020, doi: 10.32604/iasc.2020.012069.

[11] E. A. Sekehravani, E. Babulak, and M. Masoodi, "Implementing canny edge detection algorithm for noisy image," *Bull. Electr. Eng. Informatics*, vol. 9, no. 4, pp. 1404–1410, 2020, doi: 10.11591/eei.v9i4.1837.

[12] A. Mustopa, H. M. Nawawi, S. Agustiani, and S. K. Wildah, “Feature Extraction With Forest Classifer To Predicate Covid 19 Based On Thorax X-Ray Results,” *Sistemasi*, vol. 11, no. 2, p. 515, 2022, doi: 10.32520/stmsi.v11i2.1966.

[13] G. Xie, B. Guo, Z. Huang, Y. Zheng, and Y. Yan, “Combination of Dominant Color Descriptor and Hu Moments in Consistent Zone for Content Based Image Retrieval,” *IEEE Access*, vol. 8, pp. 146284–146299, 2020, doi: 10.1109/ACCESS.2020.3015285.

[14] A. R. Arrahimi, M. K. Ihsan, D. Kartini, M. R. Faisal, and F. Indriani, “Teknik Bagging Dan Boosting Pada Algoritma CART Untuk Klasifikasi Masa Studi Mahasiswa,” *J. Sains dan Inform.*, vol. 5, no. 1, pp. 21–30, 2019, doi: 10.34128/jsi.v5i1.171.

[15] F. Tangguh and Y. Islami, “Analisis performa algoritma Stochastic Gradient Descent ( SGD ) dalam mengklasifikasi tahu berformalin,” *Indones. J. Data Sci.*, vol. 3, no. 1, pp. 1–8, 2022, doi: 10.56705/ijodas.v3i1.42.

[16] A. Maulida, “Penerapan Metode Klasifikasi K-Nearest Neigbor pada Dataset Penderita Penyakit Diabetes,” *Indones. J. Data Sci.*, vol. 1, no. 2, pp. 29–33, 2020.

[17] I. P. Putri, “Analisis Performa Metode K- Nearest Neighbor (KNN) dan Crossvalidation pada Data Penyakit Cardiovascular,” *Indones. J. Data Sci.*, vol. 2, no. 1, pp. 21–28, 2021, doi: 10.33096/ijodas.v2i1.25.

[18] A. Aisyah and S. Anraeni, “Analisis Penerapan Metode K-Nearest Neighbor (K-NN) pada Dataset Citra Penyakit Malaria,” *Indones. J. Data Sci.*, vol. 3, no. 1, pp. 17–29, 2022, doi: 10.56705/ijodas.v3i1.22.

[19] Ericha Apriliyani and Y. Salim, “Analisis performa metode klasifikasi Naïve Bayes Classifier pada Unbalanced Dataset,” *Indones. J. Data Sci.*, vol. 3, no. 2, pp. 47–54, 2022, doi: 10.56705/ijodas.v3i2.45.

[20] F. T. Admojo and S. R. Jabir, “Analisis performa metode Naïve Bayesh Classifier pada Electronic Nose dalam identifikasi formalin pada tahu,” *Indones. J. Data Sci.*, vol. 4, no. 1, pp. 1–16, 2023, doi: 10.56705/ijodas.v4i1.67.