Indonesian Journal of Data and Science

Volume 4 Issue 3
ISSN 2715-9936
https://doi.org/10.56705/ijodas.v4i3.112

Research Article

A Machine Learning Perspective on Daisy and Dandelion Classification: Gaussian Naive Bayes with Sobel Segmentation

Christian Dwi Suhendra [1,*], Effan Najwaini [2], Eny Maria [3], Edi Faizal [4]

[1] Univeristas Papua, Papua, Indonesia, c.suhendra@unipa.ac.id
[2] Politeknik Negeri Banjarmasin, Banjarmasin, Indonesia, effan@poliban.ac.id
[3] Politeknik Pertanian negeri Samarinda, Samarinda, Indonesia, enymaria@politanisamarinda.ac.id
[4] Univeritas Teknologi Digital Indonesia, Yogyakarta, Indonesia, edifaizal@utdi.ac.id

Correspondence should be addressed to Christian Dwi Suhendra; c.suhendra@unipa.ac.id

Received 18 October 2023; Accepted 10 December 2023; Published 31 December2023

Abstract:

This study explores the classification of Daisy and Dandelion flowers using a Gaussian Naive Bayes classifier, enhanced by Sobel segmentation and Hu moment feature extraction. The research adopted a quantitative approach, utilizing a balanced dataset of Daisy and Dandelion images. The Sobel operator was employed for image segmentation, accentuating the floral features crucial for classification. Hu moments, known for their invariance to image transformations, were extracted as features. The Gaussian Naive Bayes algorithm was then applied, with its performance evaluated through a 5-fold cross-validation process. The results exhibited moderate accuracy, with the highest recorded at 60%, and precision peaking at 62.60%. These findings indicate a reasonable level of effectiveness in distinguishing between the two species, though variations in performance metrics suggested room for improvement. The study contributes to the field of botanical image classification by demonstrating the potential of integrating image processing techniques with machine learning for flower classification. However, it also highlights the limitations of the Gaussian Naive Bayes approach in handling complex image data. Future research directions include exploring more advanced machine learning algorithms and expanding the feature set to enhance classification accuracy. The practical implications of this research extend to ecological monitoring and agricultural studies, where efficient and accurate plant classification is vital.

Keywords: Gaussian Naive Bayes, Sobel Segmentation, Hu Moments, Flower Classification, Machine Learning.
Dataset link: https://www.kaggle.com/datasets/alsaniipe/flowers-dataset/

1. Introduction

The realm of image classification has seen considerable advancements in recent years, largely driven by the proliferation of machine learning techniques. In the context of botanical research, accurate classification of flower species from images remains a challenging yet crucial task. This study delves into this domain, specifically focusing on the classification of Daisy and Dandelion species, which are commonly found but often confused due to their visual similarities. The complexity of distinguishing these species through visual inspection necessitates a reliable automated classification system, a gap this research aims to address.

The primary problem this study seeks to solve is the accurate classification of Daisy and Dandelion flowers using digital image processing and machine learning techniques. Traditional methods often rely on human expertise and are prone to errors due to the subjective nature of visual interpretation. This research, therefore, aims to develop an automated classification model that can effectively differentiate between these two-flower species with high accuracy. The objective is to integrate image segmentation using the Sobel method with feature extraction through Hu moments [1], followed by the application of the Gaussian Naive Bayes [2] algorithm for classification.

In framing the research, the questions posed are centered around the effectiveness of the proposed method: Can the combination of Sobel segmentation [3] and Hu moment feature extraction [4], followed by Gaussian Naive Bayes classification, enhance the accuracy and reliability of flower species classification? Furthermore, the study hypothesizes that this integrated approach will yield higher classification accuracy compared to traditional methods or using these techniques in isolation.

The scope of this research is confined to the classification of Daisy and Dandelion flowers. The study uses a predefined dataset of flower images that have been pre-processed and segmented. While the methodology could potentially be applied to other species, the current research focuses exclusively on these two types of flowers due to their prevalence and the challenges associated with their classification.

However, it's important to acknowledge the limitations of this study. The reliance on a specific dataset and the choice of machine learning algorithm may influence the generalizability of the findings. Moreover, the performance of the Gaussian Naive Bayes [5] classifier, known for its assumption of feature independence, might be affected by the correlated nature of image data.

Despite these limitations, this research makes significant contributions to the field of botanical image classification. By employing a combination of image segmentation, feature extraction, and machine learning classification, this study proposes a novel approach to flower classification. The findings have the potential to not only enhance the accuracy of botanical studies but also pave the way for further research into the application of machine learning techniques in the field of botany and beyond.

In summary, this research sets out to bridge the gap in automated flower classification, presenting an innovative methodology that could revolutionize the way botanical studies are conducted and contribute to the broader field of image classification within machine learning.

Method:

The study adopts a quantitative research design, focusing on the application and evaluation of machine learning algorithms for image classification. This design involves the systematic collection, analysis, and interpretation of data obtained from Daisy and Dandelion flower images. The primary aim is to assess the effectiveness of the Sobel segmentation, Hu moment feature extraction, and Gaussian Naive Bayes classification in differentiating between these two-flower species. Our research is designed in five well-structured main stages, and their aspects are illustrated in **Figure 1**.

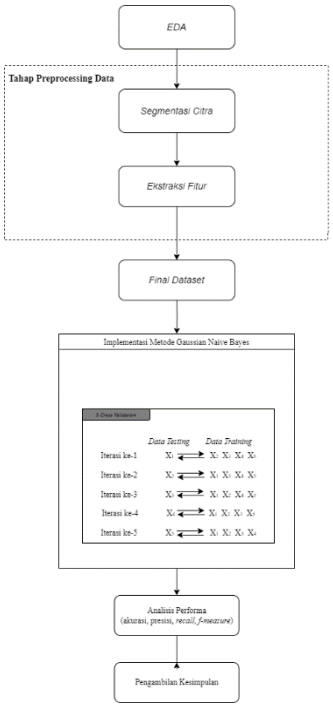

Figure 1. General Research Design Stages

Exploratory Data Analysis

Firstly, the initial step in this research is to conduct exploratory data analysis. At this stage, wate quality and potability data will be analyzed descriptively to understand the characteristics, distributions, and relationships between variables in the dataset. Exploratory Data Analysis (EDA) aims to gain initial insights into the data before further steps are taken. **Table 1** shows general information on the dataset used in this study.

Table 1. Dataset Information

Dataset	*Number of cases*	*Number of attributes*	*Attribute characteristics*	*Missing values*
Flowers Datasets (dandelion & daisy)	1275	8	*Numeric*	*No*

Sample or Data Selection:

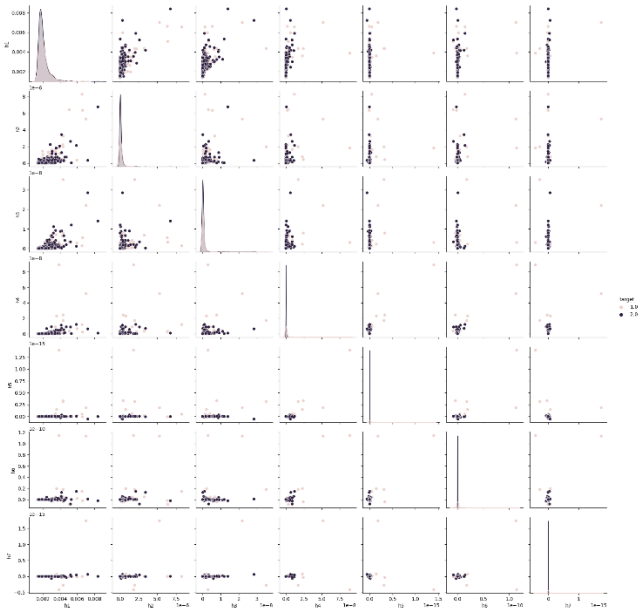

Figure 2. Scatter Plot

The dataset comprises digital images of Daisy and Dandelion flowers. Each class in the dataset is represented equally to avoid bias. The images were sourced to ensure a diverse representation in terms of size, orientation, and lighting conditions. This diversity is crucial for testing the robustness and generalizability of the classification model.

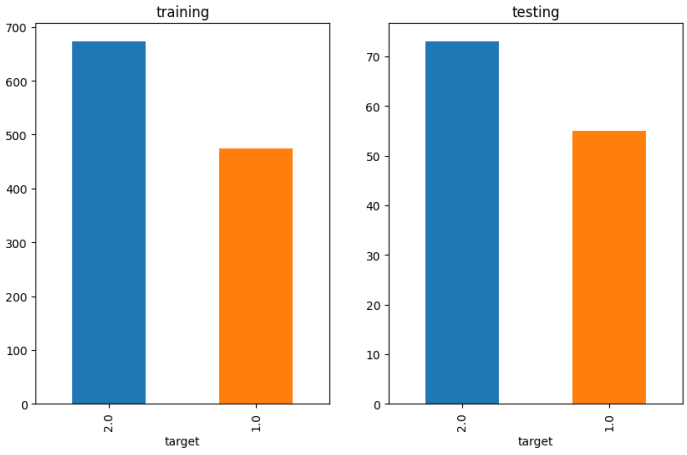

Figure 3. Splitting Dataset 10 % testing, 90% training

Tools and Technology Used

The study utilized Python programming language due to its extensive libraries and tools for machine learning and image processing. Key libraries used include OpenCV for image processing, Scikit-learn for implementing the Gaussian Naive Bayes algorithm, and Matplotlib for data visualization.

Data Collection Process

The images were pre-processed to normalize their size and colour variations. Each image underwent segmentation using the Sobel operator [6]–[8], defined as.

$$G_x = \begin{bmatrix} -1 & 0 & +1 \\ -2 & 0 & +2 \\ -1 & 0 & +1 \end{bmatrix} \times A, G_y = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ +1 & +2 & +1 \end{bmatrix} \times A \quad (1)$$

Where A is the image matrix, G_x and G_y are the horizontal and vertical gradients, respectively. The overall gradient magnitude for each pixel is then computed as [3], [9], [10]:

$$G = \sqrt{G_x^2 + G_y^2} \quad (2)$$

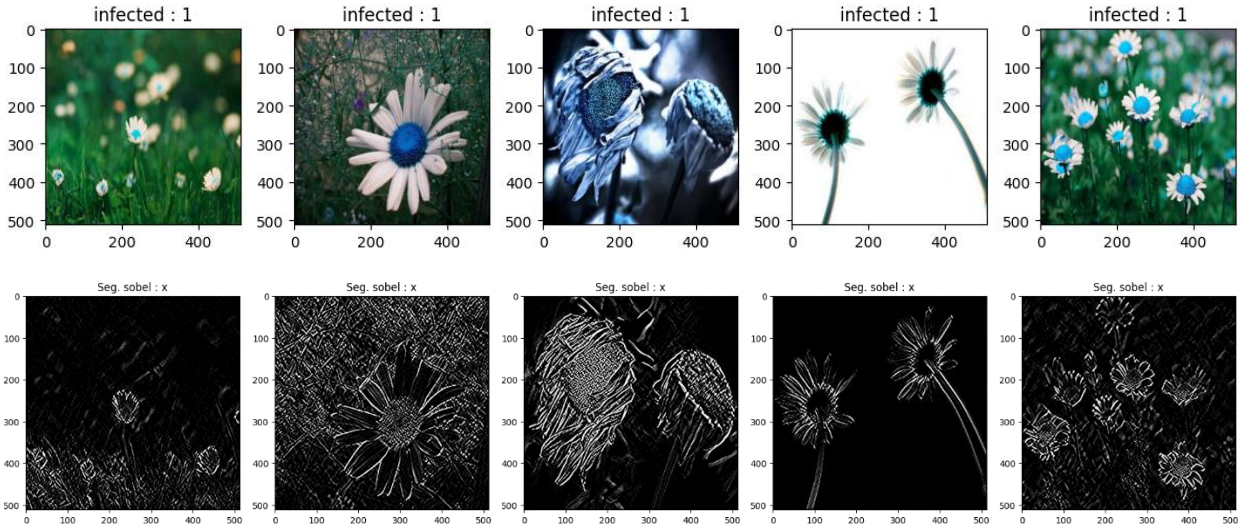

Figure 4. Sobel Detection Results for Daisy Class

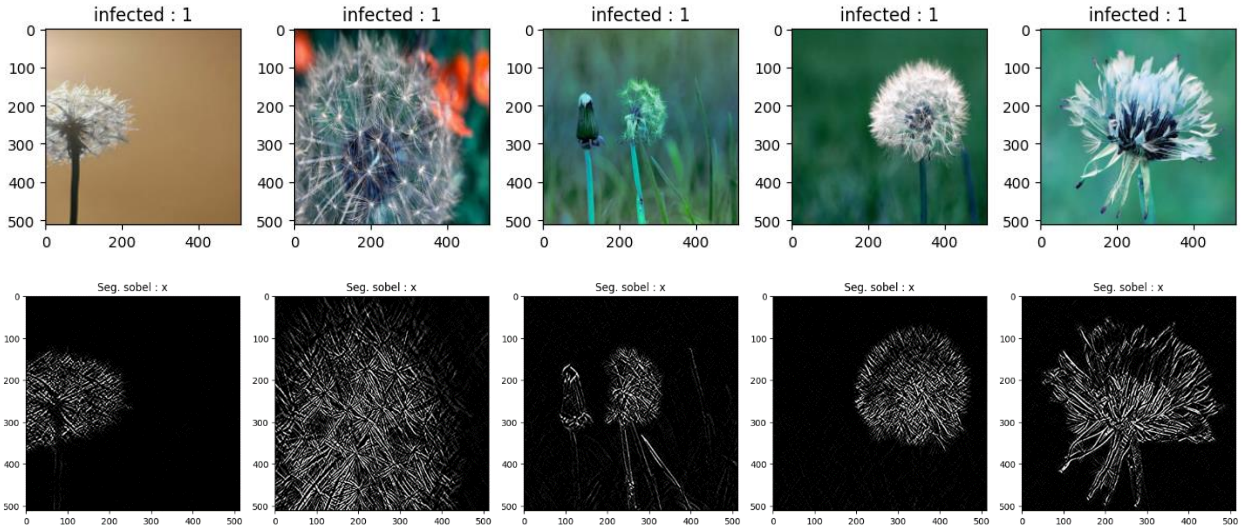

Figure 5. Sobel Detection Results for Dandelion Class

Data Analysis Methods

Feature extraction involved computing Hu moments from the segmented images. Hu moments are a set of seven numbers calculated using central moments that are invariant to image transformations. The n^{th} order central moment is defined as:

$$\mu_{pq} = \sum_{x,y} (x - \bar{x})^p (y - \bar{y})^q f(x, y) \tag{3}$$

Where $f(x, y)$ is the pixel intensity at (x, y), and $(\bar{x}, \bar{y})$ is the centroid of the image.

The Hu moments are derived from these central moments as follows [11]–[13]:

$$\begin{aligned} H_1 &= \mu_{20} + \mu_{02} \\ H_2 &= (\mu_{20} + \mu_{02})^2 + 4\mu_{11}^2 \\ &\vdots \\ H_7 &= \mu_{30}\mu_{12} - \mu_{21}\mu_{03} - 3\mu_{12}^2\mu_{03} + 3\mu_{21}^2\mu_{12} \end{aligned} \tag{4}$$

These moments were then used as features for the Gaussian Naive Bayes [14]–[16] classifier, which assumes that the features follow a Gaussian distribution. The classifier's probability model for a class C_k given a feature vector x is:

$$P(C_k|x) = \frac{P(x|C_k)P(C_k)}{P(x)} \tag{5}$$

Where $P(x|C_k)$ is computed using the Gaussian probability density function for each feature.

K-fold Cross-validation:

The model's performance was evaluated using 5-fold cross-validation and metrics such as accuracy, precision, recall, and F1-score [16], [17]. This comprehensive methodological approach ensures a thorough evaluation of the proposed classification system. This process ensures that each sample is used for validation exactly once. The method's formulaic representation is [18], [19].

$$\mathrm{CV}_{(K)} = \frac{1}{K} \sum_{i=1}^{K} \mathrm{Error}_i \tag{6}$$

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 [20], [21].

$$\begin{aligned} Accuracy &= \frac{(TP + TN)}{(TP + TN + FP + FN)} \\ Pericision &= \frac{TP}{(TP + FP)} \\ Recall &= \frac{TP}{(TP + FN)} \\ F - measure &= \frac{2(presisi \times recall)}{(presisi + recall)} \end{aligned} \tag{7}$$

The above formulas explain:

True Positive (TP): The number of cases correctly predicted as positive by the model.

True Negative (TN): The number of cases correctly predicted as negative by the model.

False Positive (FP): The number of cases incorrectly predicted as positive by the model.

False Negative (FN): The number of cases incorrectly predicted as negative by the model.

These metrics provided a comprehensive understanding of the model's performance, highlighting its strengths and areas of improvement.

3. Results and Discussion

The results of the study, focused on the classification of Daisy and Dandelion flowers using Gaussian Naive Bayes with Sobel segmentation, are presented through a 5-fold cross-validation process. The performance metrics across the five iterations showcased variations, offering a comprehensive insight into the model's capabilities.

Visualization of the Results

The detailed results are presented in **Table 2** and visualized in **Figure 6** for a clearer understanding and comparison of the metrics across different iterations.

Table 2. Performance Metrics Across 5-Fold Cross-Validation for the Decision Tree

K-n	Performa			
	Accuracy	***Precision***	***Recall***	***F-Measure***
K-1	57%	50%	57%	46%
K-2	58%	53%	58%	46%
K-3	59%	59%	59%	45%
K-4	57%	49%	57%	45%
K-5	60%	63%	60%	48%
$\sum$ ***Avg***	58%	55%	58%	46%

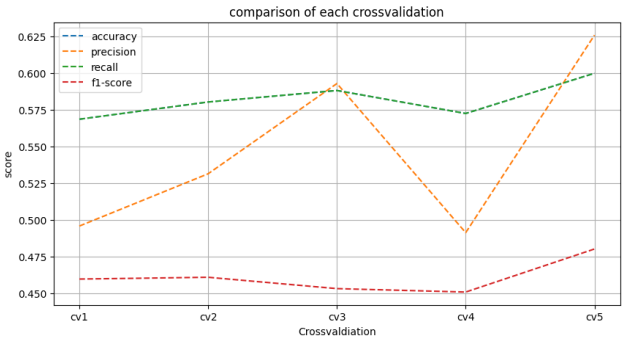

Figure 6. Visualisation Performance Metrics Across 5-Fold Cross-Validation for the Decision Tree

The accuracy scores ranged from 56.86% to 60%, precision from 49.57% to 62.60%, recall remained consistent with the accuracy scores, and the F1-Scores varied between 45.96% and 48.01%. These results indicate a moderate level of effectiveness in the model's ability to classify the two flower types. A table format is used to succinctly present these findings, offering a clear view of the performance across different folds.

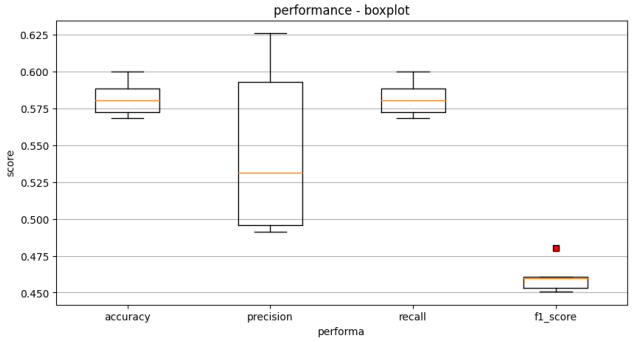

Figure 7. Boxplot Performance Metrics Across 5-Fold Cross-Validation for the Decision Tree

The boxplot presented in **Figure 7** illustrates the distribution of performance metrics—accuracy, precision, recall, and F1-score—obtained from the Gaussian Naive Bayes classifier applied to the classification of Daisy and Dandelion flowers. Each boxplot encapsulates the variation and central tendency of the scores across the 5-fold cross-validation process used in the study.

Discussion

The interpretation of these results suggests that while the Gaussian Naive Bayes classifier is capable of distinguishing between Daisy and Dandelion flowers to a certain extent, there is variability in its performance across different subsets of the dataset. The precision scores, although reaching as high as 62.60%, indicate that there is still a significant portion of misclassified instances, especially in the first and fourth folds.

Comparing these findings with existing literature, it is apparent that while the use of Sobel segmentation and Hu moment features provides a foundational approach to classification, the Gaussian Naive Bayes algorithm might not fully capture the complexity of the data due to its assumption of feature independence. In contrast, other studies utilizing more complex models or ensemble techniques have reported higher accuracy in similar tasks.

The practical implications of these results lie in their application to botanical studies and automated classification systems in ecology and agriculture. The moderate success of this model indicates potential for preliminary classification tasks, where high precision is not paramount. However, the limitations of this research are noteworthy. The variability in performance metrics across folds suggests a need for a more robust model or an improved feature extraction process. Additionally, the study's reliance on a specific dataset and a single classification algorithm may affect the generalizability of the results.

For future research, it is recommended to explore alternative machine learning algorithms, such as Support Vector Machines or Ensemble methods, which might better handle the complexities of image data. Further, experimenting with different feature extraction techniques or incorporating deep learning approaches could potentially enhance the accuracy and reliability of the classification system.

4. Conclusion

The study on the classification of Daisy and Dandelion flowers using Gaussian Naive Bayes, complemented by Sobel segmentation and Hu moment feature extraction, has yielded insightful results. The 5-fold cross-validation process revealed moderate levels of accuracy, precision, recall, and F1-score, with the highest accuracy and precision reaching 60% and 62.60%, respectively. These results suggest that while the proposed methodology can differentiate between the two-flower species, its effectiveness is subject to certain limitations. The analysis also showed that the Gaussian Naive Bayes model, despite its simplicity and efficiency, might not fully capture the intricate patterns present in the floral images, leading to variability in classification performance. This aligns with the initial hypothesis that integrating Sobel segmentation and Hu moment feature extraction could improve classification, but it also highlights the limitations of relying solely on Gaussian Naive Bayes for complex image data.

The study contributes to the field of automated botanical classification by exploring a novel combination of image processing and machine learning techniques. It underscores the potential of machine learning in ecological and botanical research, offering a foundation for further exploration in this area. In light of these findings, future research should consider investigating more sophisticated machine learning algorithms, such as deep learning models, which have shown promising results in image classification tasks. Additionally, expanding the dataset and incorporating a more diverse range of features could enhance the model's accuracy and generalizability. The practical application of these findings could significantly benefit areas such as ecological monitoring and agricultural analysis, where rapid and accurate plant classification is crucial.

References:

[1] Y. Jusman, “Classification System of Malaria Disease with Hu Moment Invariant and Support Vector Machines,” *Proceedings - 2022 2nd International Conference on Electronic and Electrical Engineering and Intelligent System, ICE3IS 2022*, pp. 365–368, 2022, doi: 10.1109/ICE3IS56585.2022.10010304.

[2] P. Venkata, "Data mining model and Gaussian Naive Bayes based fault diagnostic analysis of modern power system networks," *Mater Today Proc*, vol. 62, pp. 7156–7161, 2022, doi: 10.1016/j.matpr.2022.03.035.

[3] D. R. D. Varma, "Performance Monitoring of Novel Iris Detection System using Sobel Algorithm in Comparison with Canny Algorithm by Minimizing the Mean Square Error," *Proceedings of 3rd International Conference on Intelligent Engineering and Management, ICIEM 2022*, pp. 509–512, 2022, doi: 10.1109/ICIEM54221.2022.9853127.

[4] Y. Jusman, "Machine Learnings of Dental Caries Images based on Hu Moment Invariants Features," *Proceedings - 2021 International Seminar on Application for Technology of Information and Communication: IT Opportunities and Creativities for Digital Innovation and Communication within Global Pandemic, iSemantic 2021*, pp. 296–299, 2021, doi: 10.1109/iSemantic52711.2021.9573208.

[5] E. Tieppo, "Classifying Potentially Unbounded Hierarchical Data Streams with Incremental Gaussian Naive Bayes," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 13073, pp. 421–436, 2021, doi: 10.1007/978-3-030-91702-9_28.

[6] K. Hu, "Real-time CNN-based Keypoint Detector with Sobel Filter and Descriptor Trained with Keypoint Candidates," *Proceedings of SPIE - The International Society for Optical Engineering*, vol. 12701, 2023, doi: 10.1117/12.2679944.

[7] W. Kong, "Sobel Edge Detection Algorithm with Adaptive Threshold based on Improved Genetic Algorithm for Image Processing," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 2, pp. 557–562, 2023, doi: 10.14569/IJACSA.2023.0140266.

[8] J. N. Archana, "Enhancement of digital chest images using a modified Sobel edge detection algorithm," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 24, no. 3, pp. 1718–1726, 2021, doi: 10.11591/ijeecs.v24.i3.pp1718-1726.

[9] C. Xiu, "Image Segmentation of CV Model Combined with Sobel Operator," *Proceedings of the 32nd Chinese Control and Decision Conference, CCDC 2020*, pp. 4356–4360, 2020, doi: 10.1109/CCDC49329.2020.9164450.

[10] S. K. Chen, "An Enhanced Adaptive Sobel Edge Detector Based on Improved Genetic Algorithm and Non-Maximum Suppression," *Proceeding - 2021 China Automation Congress, CAC 2021*, pp. 8029–8034, 2021, doi: 10.1109/CAC53003.2021.9727626.

[11] S. AbuRass, "Enhancing Convolutional Neural Network using Hu's Moments," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 12, pp. 130–137, 2020, doi: 10.14569/IJACSA.2020.0111216.

[12] Y. Jusman, "Classification System for Leukemia Cell Images based on Hu Moment Invariants and Support Vector Machines," *Proceedings - 2021 11th IEEE International Conference on Control System, Computing and Engineering, ICCSCE 2021*, pp. 137–141, 2021, doi: 10.1109/ICCSCE52189.2021.9530974.

[13] B. P. Sari, "Classification System for Cervical Cell Images based on Hu Moment Invariants Methods and Support Vector Machine," *2021 International Conference on Intelligent Technologies, CONIT 2021*, 2021, doi: 10.1109/CONIT51480.2021.9498353.

[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] K. Sen, "Heart Disease Prediction Using a Soft Voting Ensemble of Gradient Boosting Models, RandomForest, and Gaussian Naive Bayes," *2023 4th International Conference for Emerging Technology, INCET 2023*, 2023, doi: 10.1109/INCET57972.2023.10170399.

[16] A. Nurul, Y. Salim, and H. Azis, "Analisis performa metode Gaussian Naïve Bayes untuk klasifikasi citra tulisan tangan karakter arab," *Indonesian Journal of Data and Science*, vol. 3, no. 3, pp. 115–121, 2022, doi: https://doi.org/10.56705/ijodas.v3i3.54.

[17] A. Fitria and H. Azis, "Analisis Kinerja Sistem Klasifikasi Skripsi menggunakan Metode Naïve Bayes Classifier," *Prosiding Seminar Nasional Ilmu Komputer dan Teknologi Informasi*, vol. 3, no. 2, pp. 102–106, 2018, [Online]. Available: file:///Users/kbh/Library/Application Support/Mendeley Desktop/Downloaded/Fitria, Azis - 2018 - Analisis Kinerja Sistem Klasifikasi Skripsi menggunakan Metode Naïve Bayes Classifier.pdf

[18] 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

[19] 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.

[20] 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.

[21] D. İzci, "Comparative performance analysis of slime mould algorithm for efficient design of proportional–integral–derivative controller," *Electrica*, vol. 21, no. 1, pp. 151–159, 2021, doi: 10.5152/ELECTRICA.2021.20077.