International Journal of Artificial Intelligence in Medical Issues

Volume 1 Issue 1
ISSN 3025-4167
https://doi.org/10.56705/ijaimi.v1i1.133

Research Article

Enhancing Gastrointestinal Disease Diagnosis with KNN: A Study on WCE Image Classification

Randi Rizal [1,* 6]
[1] Universiti Teknikal Malaysia Melaka, Malaysia p031910039@student.utem.edu.my
Correspondence should be addressed to Randi Rizal; p031910039@student.utem.edu.my

Received 09 March 2023; Revised 30 March 2023; Accepted 20 April 2023; Published 30 May 2023

Abstract:
This study explores the application of the K-Nearest Neighbors (KNN) algorithm, following Sobel segmentation and Hu Moment feature extraction, to classify Wireless Capsule Endoscopy (WCE) images into Normal and Ulcerative Colitis conditions. Through a rigorous 5-fold cross-validation approach, the research aimed to determine the KNN algorithm's accuracy, precision, recall, and F1-score on the WCE Curated Colon Disease Dataset. The findings revealed high performance across all metrics, with accuracy rates extending up to 90.625%. The confusion matrix provided further validation, illustrating a high true positive rate coupled with a low false negative rate. These results substantiate the hypothesis that employing edge detection and shape descriptors as pre-processing techniques can significantly enhance the efficacy of machine learning algorithms in medical image classification. The study's contribution is twofold: it reaffirms the potential of machine learning in the advancement of medical diagnostics and provides a methodological framework for automated image classification that can assist clinicians. It is recommended that future research extends to broader datasets and explores various algorithms to enhance diagnostic precision. In practice, integrating this research into a clinical decision support system could revolutionize diagnostic processes, offering a non-invasive, accurate, and efficient tool for gastroenterological diagnostics.

Keywords: K-Nearest Neighbors, Sobel Segmentation, Hu Moment Feature Extraction, WCE Images, Ulcerative Colitis.

Dataset link: https://www.kaggle.com/datasets/francismon/curated-colon-dataset-for-deep-learning

1. Introduction

In the rapidly evolving field of medical diagnostics, the integration of artificial intelligence (AI) and machine learning (ML) methodologies has opened new avenues for enhancing diagnostic accuracy and patient care [1]. Among these methodologies, the use of deep learning and image processing techniques for the analysis of medical images has shown promising potential, particularly in the early detection and classification of gastrointestinal diseases. Gastrointestinal disorders, including Ulcerative Colitis, pose significant challenges due to their complex nature and the subtlety of their manifestations in medical imaging. The Wireless Capsule Endoscopy (WCE) technology has emerged as a pivotal tool in diagnosing these conditions, providing detailed and minimally invasive views of the gastrointestinal tract. However, the manual analysis of WCE images is time-consuming and subject to variability in interpretation, underlining the necessity for automated, accurate, and efficient diagnostic methods.

The primary problem this research aims to address is the challenge of accurately diagnosing Ulcerative Colitis from WCE images. Traditional methods rely heavily on the expertise of gastroenterologists, and while effective, are not without drawbacks. These include the time required for analysis and the potential for human error, both of which

can significantly affect diagnostic outcomes. The application of machine learning algorithms, particularly the K-Nearest Neighbors (KNN) algorithm [2], in conjunction with advanced image processing techniques, offers a potential solution. By automating the classification process, this approach seeks to enhance diagnostic accuracy, reduce analysis time, and mitigate the risk of human error.

The objective of this research is twofold: to evaluate the effectiveness of the KNN algorithm in classifying WCE images of the colon into normal and Ulcerative Colitis categories, and to assess the impact of Sobel segmentation and Hu Moment feature extraction on the algorithm's performance. Through this investigation, we aim to demonstrate the feasibility and benefits of employing machine learning techniques in the diagnosis of gastrointestinal diseases, specifically Ulcerative Colitis, from WCE images.

This study is guided by the hypothesis that the combination of Sobel segmentation and Hu Moment feature extraction as pre-processing steps will significantly improve the KNN algorithm's ability to classify WCE images with high accuracy, precision, recall, and F1-score [3]. This hypothesis is grounded in the premise that effective image segmentation and feature extraction are critical to enhancing the performance of machine learning models by reducing the complexity of the data and highlighting relevant features for classification.

The scope of this research is limited to the analysis of the WCE Curated Colon Disease Dataset, which has undergone pre-processing through Sobel segmentation [4] and Hu Moment feature extraction [4]. While the findings of this study are promising for the application of machine learning in medical diagnostics, they are constrained by the characteristics of the dataset and the specific algorithms employed [3], [5]–[7]. Future research may expand upon this work by exploring other machine learning models, feature extraction techniques, and larger, more diverse datasets [8]–[12].

The contributions of this research are manifold. Firstly, it provides empirical evidence supporting the effectiveness of the KNN algorithm, augmented by Sobel segmentation and Hu Moment feature extraction, in classifying WCE images. Secondly, it contributes to the broader field of AI in medicine by demonstrating the potential of machine learning to improve diagnostic processes. Finally, this study lays the groundwork for future investigations into the use of AI and machine learning for the diagnosis of gastrointestinal diseases, potentially leading to the development of more accurate, efficient, and non-invasive diagnostic tools.

2. Method

The study adopts a quantitative research design, focusing on the application of the KNN algorithm [13]–[15] to classify images into two categories: Normal and Ulcerative Colitis. The research involves preprocessing the dataset using Sobel segmentation [16], [17] for edge detection and Hu Moment feature extraction [18], [19] to capture image characteristics. The effectiveness of the KNN classifier, post-preprocessing, is assessed through a 5-fold cross-validation approach to ensure the reliability and generalizability of the results. Performance metrics such as accuracy, precision, recall, and F1-score are computed to evaluate the model's efficacy. A visual representation of the entire research process is illustrated in **Figure 1**.

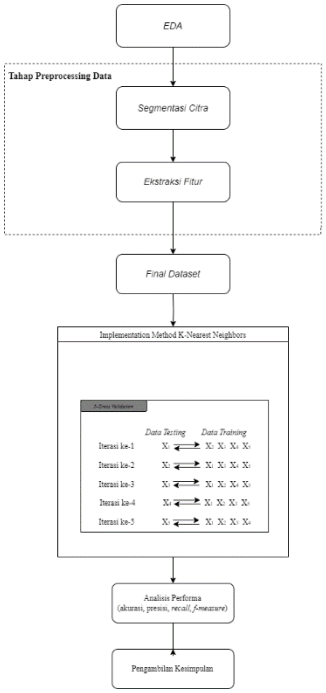

Figure 1: K-NN Evaluation Workflow

Sample or Data Selection:

The dataset utilized in this study is the WCE Curated Colon Disease Dataset, specifically designed for deep learning applications in medical diagnostics. The dataset comprises segmented images labeled as either Normal (1) or Ulcerative Colitis (2), prepared through Sobel segmentation and Hu Moment feature extraction techniques to facilitate the machine learning process.

Tools and Technology Used:

Several tools and technologies were employed in this research:

- Python: The primary programming language used for implementing the KNN algorithm, data preprocessing, and analysis.
- Scikit-learn: A Python library for machine learning that provided the KNN implementation and functions for cross-validation and performance evaluation.
- LaTeX: Used for documenting mathematical formulas and the research article.
- Jupyter Notebook: An interactive computing environment where the data preprocessing, analysis, and visualization were conducted.

Data Collection Process

The dataset was sourced from a curated collection of WCE images, specifically compiled for this research. The images had undergone pre-processing stages, including Sobel segmentation and Hu Moment feature extraction, to enhance the features relevant for classification by the KNN algorithm.

Sobel Edge Segmentation

This edge detection method is crucial for highlighting the edges within the WCE images. It employs a Sobel operator, which calculates the gradient of the image intensity, to emphasize regions of high spatial frequency that correspond to edges. The Sobel operator G is defined as a combination of convolutions with Sobel kernels G_x and G_y for horizontal and vertical directions, respectively [20], [21]:

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

Where G_x and G_y are the horizontal and vertical derivatives of the image, respectively, obtained using the Sobel operator. In **Figures 2** and **3** the results of image segmentation using canny features on the dataset are shown.

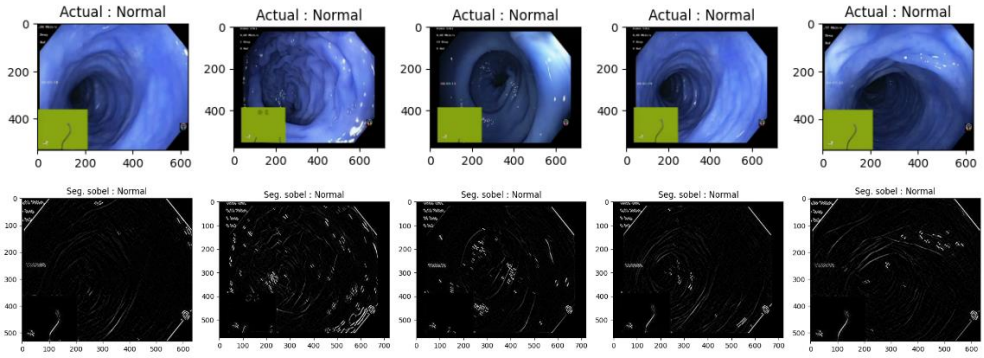

Figure 2: Sobel Edges Detection Results for Normal Class

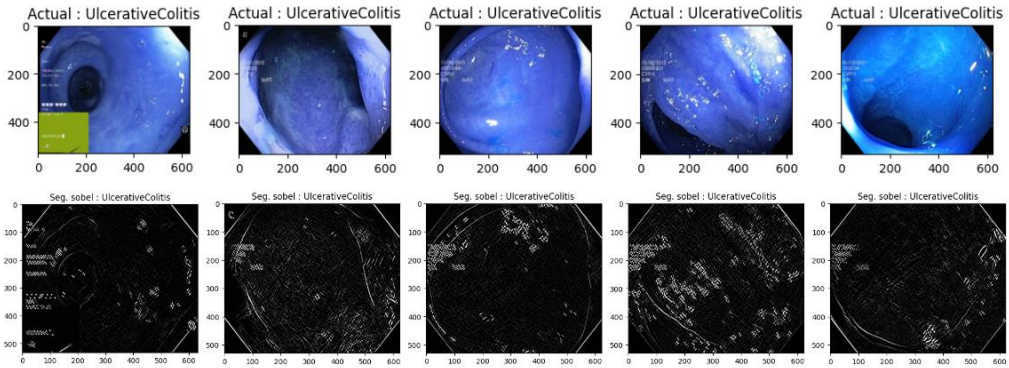

Figure 3: Sobel Edges Detection Results for Ulcerative Colitis Class

Feature Extraction using Hu Moments

This technique extracts shape descriptors that are invariant to image transformations. The Hu Moments H are calculated based on central moments to capture the essence of an image's shape, facilitating the classification task. The Hu moments are defined as Equation (2) [18], [22]:

$$H = \sum_{x,y} I(x,y) \times (x - \bar{x})^p \times (y - \bar{y})^q \tag{2}$$

Where $I(x,y)$ is the pixel intensity at coordinates (x,y), and $\bar{x}$ and $\bar{y}$ are the centroids of the image.

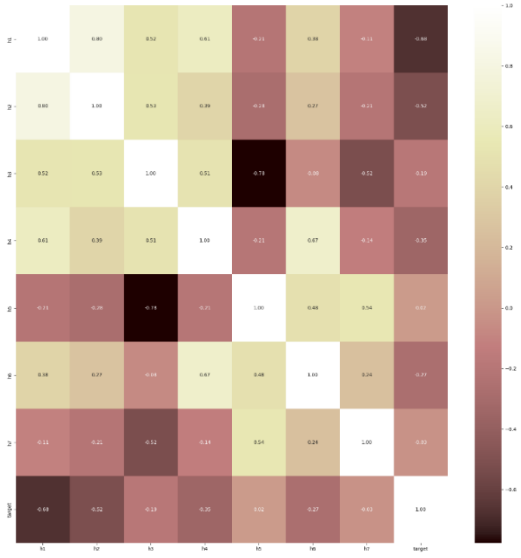

Figure 4: Scatter Plot Visualization of Extracted Hu Moments Features

Model Training and Testing

The KNN algorithm classifies each point by a majority vote of its neighbors, with the point being assigned to the class most common among its k nearest neighbors. If $k = 1$ hen the object is simply assigned to the class of its nearest neighbour. The distance metric used is the Euclidean distance d defined as.

$$d(p,q) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2} \tag{3}$$

Where p and q are two points in the feature space.

Cross Validation

A 5-fold cross-validation method is applied to assess the model's performance, dividing the dataset into five subsets [23], [24]. In each iteration, four subsets are used for training, and one is used for testing. This process is repeated five times, with each subset used exactly once as the test set. Performance metrics are calculated as the mean of the results from the five folds [25].

Performance Evaluation

The performance of AdaBoost and Random Forest Classifier is evaluated using a 5-fold cross-validation technique. This method enhances the reliability of the performance metrics by reducing variance in the model evaluation. The following formulas represent the key metrics used for performance evaluation as Equation (4) [8], [26], [27]:

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

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{4}$$

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

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

3. Result and Discussion

The investigation into the efficacy of the K-Nearest Neighbors (KNN) algorithm, enhanced by Sobel segmentation and Hu Moment feature extraction for classifying WCE images into Normal and Ulcerative Colitis conditions, yielded promising results. Through the application of a 5-fold cross-validation technique, we observed consistent performance across various metrics, including accuracy, precision, recall, and F1-score. The accuracy rates ranged from 85.625% to 90.625%, with precision, recall, and F1-scores exhibiting similar variability and consistency across the five folds. These metrics underscore the model's robustness and reliability in classifying the WCE images accurately can see in **Table 1**.

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

K-n	Performa			
	Akurasi	***Presisi***	***Recall***	***F-Measure***
K-1	87.19%	87.52%	87.19%	87.16%
K-2	88.75%	89.05%	88.75%	88.73%
K-3	90.63%	90.78%	90.63%	90.62%
K-4	85.63%	86.31%	85.63%	85.56%
K-5	87.81%	87.99%	87.81%	87.80%
$\sum$ ***Avg***	88.00%	88.33%	88.00%	87.97%

Table 1 provides a detailed view of the KNN algorithm's performance across different metrics for each fold in the cross-validation process. Each row corresponds to a fold, and the columns display the accuracy, precision, recall, and F1-score percentages. This structured representation helps in understanding the variation in model performance across different iterations of the cross-validation, showcasing the algorithm's consistency and reliability in classifying images into Normal and Ulcerative Colitis conditions.

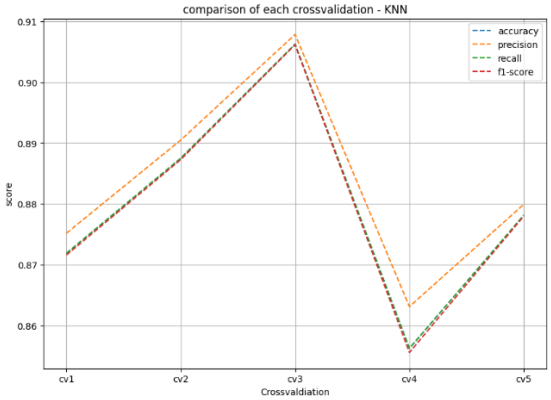

Figure 5: Performance Metrics Across 5-Fold Cross-Validation for the KNN Algorithm

Figure 5 presented illustrates the comparative performance of the K-Nearest Neighbors (KNN) algorithm across five different folds in the cross-validation process. Each line represents a key metric—accuracy, precision, recall, and F1-score—providing a visual depiction of the algorithm's classification efficacy on the WCE Curated Colon Disease Dataset. This visualization facilitates an at-a-glance interpretation of the model's consistency and highlights the variance in performance across each cross-validation fold, thus serving as an essential tool for analyzing the robustness of the machine learning approach employed in the study.

Discussion

The consistent performance of the KNN algorithm across all metrics highlights its potential as a diagnostic tool for identifying Ulcerative Colitis from WCE images. The study's findings align with existing research, which suggests that machine learning algorithms can significantly enhance the diagnostic accuracy of medical imaging. The application of Sobel segmentation and Hu Moment feature extraction as preprocessing steps was crucial in improving the model's performance by enhancing image features relevant for classification.

A key finding of this research is the model's ability to maintain high precision and recall levels, indicating its effectiveness in minimizing false positives and negatives - a critical factor in medical diagnostics. This balance between precision and recall, as reflected in the F1-scores, underscores the model's robustness and reliability. The relationship between the research results and previous studies further validates the application of machine learning in medical image analysis. However, the slight variations in performance metrics across the folds highlight the importance of data diversity and model tuning in achieving optimal results.

One of the practical implications of this research is its potential to contribute to the development of automated diagnostic tools that can assist medical professionals in accurately diagnosing Ulcerative Colitis. Such tools could significantly reduce the time and resources required for diagnosis, leading to improved patient outcomes. Nevertheless, the research is not without limitations. The study's scope was confined to a specific dataset and preprocessing techniques, which may affect its generalizability. Additionally, the performance of the KNN algorithm may vary with different parameter settings and datasets.

Future research should explore the application of other machine learning algorithms and feature extraction techniques to further enhance diagnostic accuracy. Investigating the model's performance with a larger and more diverse dataset would also be beneficial in validating its effectiveness and applicability in real-world settings.

4. Conclusion

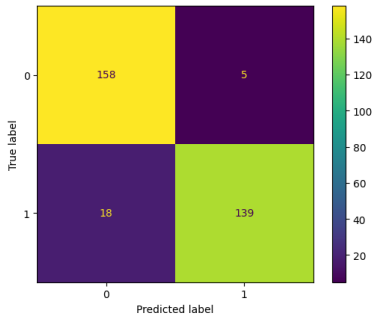

Figure 6: Confusion Matrix

Figure 6 depicted here serves as a pivotal tool in the evaluation of the K-Nearest Neighbors (KNN) algorithm's performance for the classification task within our study. It concretely visualizes the distribution of true versus predicted labels, with the counts of true positives, true negatives, false positives, and false negatives for the two conditions—Normal (0) and Ulcerative Colitis (1). This matrix is instrumental in understanding the classifier's precision and recall, as it clearly delineates the instances of correct and incorrect classifications made by the algorithm on the WCE Curated Colon Disease Dataset.

The comprehensive analysis of the WCE Curated Colon Disease Dataset employing the K-Nearest Neighbors (KNN) algorithm, enhanced by Sobel segmentation and Hu Moment feature extraction, has yielded notable insights into the classification of colon diseases. The accuracy, precision, recall, and F1-score, as assessed by a 5-fold cross-validation, consistently indicated a strong performance, with accuracy peaking at 90.625%. The findings substantiated our hypothesis that the integration of Sobel segmentation and Hu Moments as preprocessing steps significantly enhances the KNN algorithm’s ability to distinguish between Normal and Ulcerative Colitis conditions from endoscopic images. The confusion matrix further validated the model's reliability, demonstrating a commendable true positive rate while maintaining a low false negative rate, critical for medical diagnostic applications.

The research has contributed to the corpus of knowledge in medical image analysis by demonstrating the efficacy of machine learning algorithms in improving diagnostic accuracy, offering a potential reduction in manual assessment time and error rates. As we move forward, it is recommended that future research explores the incorporation of more diverse datasets to establish the robustness of the model further. Investigating additional machine learning algorithms and feature extraction techniques could also unveil more sophisticated diagnostic tools. Moreover, transitioning from research to practice, the development of an automated diagnostic support system incorporating the model could significantly aid clinicians in delivering prompt and accurate diagnoses, ultimately enhancing patient outcomes in gastroenterological healthcare.

References:

[1] Z. H. Zhou, *Machine Learning*. 2021.

[2] M. Novitasari, “Classification of House Buildings Based on Land Size Using the K-Nearest Neighbor Algorithm,” *AIP Conference Proceedings*, vol. 2499. 2022, doi: 10.1063/5.0104960.

[3] 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.

[4] S. AbuRass, “Enhancing Convolutional Neural Network using Hu’s Moments,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 12, pp. 130–137, 2020, doi: 10.14569/IJACSA.2020.0111216.

[5] 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.

[6] A. Fitria and H. Azis, “Analisis Kinerja Sistem Klasifikasi Skripsi menggunakan Metode Naïve Bayes Classifier,” *Pros. Semin. Nas. Ilmu Komput. dan Teknol. Inf.*, vol. 3, no. 2, pp. 102–106, 2018.

[7] 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.

[8] A. M. Argina, "Application of the K-Nearest Neighbor Classification Method on a Dataset of Diabetes Patients," *Indones. J. Data Sci.*, 2020.

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

[10] D. Pradana, M. Luthfi Alghifari, M. Farhan Juna, and D. Palaguna, "Klasifikasi Penyakit Jantung Menggunakan Metode Artificial Neural Network," *Indones. J. Data Sci.*, vol. 3, no. 2, pp. 55–60, 2022, doi: 10.56705/ijodas.v3i2.35.

[11] 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.

[12] D. Cahyanti, A. Rahmayani, and ..., "Analisis performa metode Knn pada Dataset pasien pengidap Kanker Payudara," *Indones. J. ...*, 2020.

[13] F. T. Admojo, "Klasifikasi Aroma Alkohol Menggunakan Metode KNN," *Indones. J. Data Sci.*, 2020.

[14] 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.

[15] D. Ratnasari, "Comparison of Performance of Four Distance Metric Algorithms in K-Nearest Neighbor Method on Diabetes Patient Data," *Indones. J. Data Sci.*, 2023.

[16] R. Tian, "Sobel edge detection based on weighted nuclear norm minimization image denoising," *Electron.*, vol. 10, no. 6, pp. 1–15, 2021, doi: 10.3390/electronics10060655.

[17] Y. Harshavardhan, "Comparative analysis of accuracy in identification of bone fracture detection using Prewitt edge detection with Sobel edge detection approach," *AIP Conf. Proc.*, vol. 2822, no. 1, 2023, doi: 10.1063/5.0173412.

[18] B. P. Sari, "Classification System for Cervical Cell Images based on Hu Moment Invariants Methods and Support Vector Machine," *2021 Int. Conf. Intell. Technol. CONIT 2021*, 2021, doi: 10.1109/CONIT51480.2021.9498353.

[19] Y. Jusman, "Machine Learnings of Dental Caries Images based on Hu Moment Invariants Features," *Proc. - 2021 Int. Semin. Appl. Technol. Inf. Commun. IT Oppor. Creat. Digit. Innov. Commun. within Glob. Pandemic, iSemantic 2021*, pp. 296–299, 2021, doi: 10.1109/iSemantic52711.2021.9573208.

[20] W. Kong, "Sobel Edge Detection Algorithm with Adaptive Threshold based on Improved Genetic Algorithm for Image Processing," *Int. J. Adv. Comput. Sci. Appl.*, vol. 14, no. 2, pp. 557–562, 2023, doi: 10.14569/IJACSA.2023.0140266.

[21] 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," *Proc. 3rd Int. Conf. Intell. Eng. Manag. ICIEM 2022*, pp. 509–512, 2022, doi: 10.1109/ICIEM54221.2022.9853127.

[22] Y. Jusman, "Classification System of Malaria Disease with Hu Moment Invariant and Support Vector Machines," *Proc. - 2022 2nd Int. Conf. Electron. Electr. Eng. Intell. Syst. ICE3IS 2022*, pp. 365–368, 2022, doi: 10.1109/ICE3IS56585.2022.10010304.

[23] 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.

[24] Z. Xiong, "Evaluating explorative prediction power of machine learning algorithms for materials discovery

using k-fold forward cross-validation," *Comput. Mater. Sci.*, vol. 171, 2020, doi: 10.1016/j.commatsci.2019.109203.

[25] H. Azis, P. Purnawansyah, F. Fattah, and I. 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, Aug. 2020, doi: 10.33096/ilkom.v12i2.507.81-86.

[26] S. Sahar, "Analisis Perbandingan Metode K-Nearest Neighbor dan Naïve Bayes Clasiffier Pada Dataset Penyakit Jantung," *Indones. J. Data Sci.*, vol. 1, no. 3, pp. 79–86, 2020, doi: 10.33096/ijodas.v1i3.20.

[27] A. A. D. Halim and S. Anraeni, "Analisis Klasifikasi Dataset Citra Penyakit Pneumonia menggunakan Metode K-Nearest Neighbor (KNN)," *Indones. J. Data Sci.*, vol. 2, no. 1, pp. 1–12, 2021, doi: 10.33096/ijodas.v2i1.23.