International Journal of Artificial Intelligence in Medical Issues

Volume 2 Issue 1
ISSN 3025-4167
https://doi.org/10.56705/ijaimi.v2i1.138

Research Article

Segmentation and Feature Extraction for Malaria Detection in Blood Smears

Nurul Rismayanti [1,*]
[1] Universitas Negeri Malang, nurul.rismayanti.2305348@students.um.ac.id
Correspondence should be addressed to Nurul Rismayanti; nurul.rismayanti.2305348@students.um.ac.id

Received 24 March 2024; Revised 08 April 2024; Accepted 30 April 2024; Published 31 May 2024

Abstract:
Malaria remains a critical global health challenge, particularly in tropical and subtropical regions. Early and accurate diagnosis is essential for effective treatment and control. Traditional methods of malaria diagnosis, such as microscopic examination of blood smears, are time-consuming and prone to human error. This study aims to develop an automated system for malaria detection using machine learning techniques, specifically a decision tree classifier. The dataset, sourced from the National Institutes of Health (NIH), comprises 27,558 blood smear images equally divided into Normal and Malaria classes. The preprocessing steps included segmentation using the Canny edge detector and feature extraction using Hu Moments, followed by data normalization to ensure a mean of 0 and variance of 1. The decision tree classifier was trained and evaluated using 5-fold cross-validation, yielding an average accuracy of 77.32%, precision of 77.31%, recall of 77.37%, and F1-Score of 77.48%. These results demonstrate the model's robustness and effectiveness in differentiating between malaria-infected and uninfected images. The study confirms the viability of using Hu Moments for feature extraction and highlights the decision tree classifier's suitability for this task. The proposed method has significant implications for automated malaria diagnosis, potentially improving diagnostic accuracy and efficiency in clinical settings. Future research should validate these findings on diverse datasets, explore advanced classification techniques, and integrate real-time image acquisition to enhance practical applicability. The integration of such automated systems in healthcare can revolutionize malaria diagnosis, especially in resource-limited settings.

Keywords: Malaria Detection, Machine Learning, Decision Tree, Hu Moments, Image Processing.

Dataset link: https://www.kaggle.com/datasets/iarunava/cell-images-for-detecting-malaria

1. Introduction

through the bites of infected Anopheles mosquitoes. Despite significant advances in medicine and public health, malaria remains a major public health challenge, particularly in tropical and subtropical regions. The disease is responsible for substantial morbidity and mortality, affecting millions of people worldwide each year. Accurate and timely diagnosis is crucial for effective treatment and control of malaria. Traditional diagnostic methods, such as microscopic examination of blood smears, require skilled personnel and are time-consuming, often leading to delays in diagnosis and treatment. Consequently, there is a growing need for automated diagnostic systems that can provide rapid and accurate malaria detection.

The problem to be solved in this research is the inefficiency and potential inaccuracy of manual malaria diagnosis. Microscopic examination, while considered the gold standard, is labor-intensive and subject to human error. Misdiagnosis can result in inappropriate treatment, contributing to drug resistance and increased morbidity and mortality. Automated systems using machine learning techniques offer a promising solution by providing consistent

and objective analysis. This study aims to develop a machine learning-based approach to detect malaria from blood smear images, leveraging image processing and classification techniques to enhance diagnostic accuracy and efficiency [1], [2].

The primary objective of this research is to create an automated system that accurately distinguishes between malaria-infected and uninfected blood smear images. By implementing advanced image processing techniques and machine learning algorithms, we seek to improve the speed and accuracy of malaria diagnosis. Specifically, the study focuses on pre-processing the images using the Canny edge detector for segmentation and extracting features using Hu Moments. The decision tree (DT) classifier will be employed to categorize the images into two classes: Normal and Malaria. Performance metrics such as accuracy, precision, recall, and F1-measure will be used to evaluate the effectiveness of the proposed method [3]–[5].

Several research questions guide this study. Can segmentation and feature extraction techniques effectively differentiate between Normal and Malaria images? How accurately can a decision tree classifier diagnose malaria using the extracted features? What are the advantages and limitations of using decision trees for this type of classification problem? Additionally, how does the proposed automated system compare with traditional diagnostic methods in terms of accuracy and efficiency? Addressing these questions will help determine the feasibility and reliability of the machine learning approach in real-world clinical settings.

The scope of this research is confined to the use of a publicly available dataset of blood smear images from the National Institutes of Health (NIH). The dataset consists of 27,558 images, equally divided between Normal and Malaria classes, ensuring a balanced dataset. This study does not involve real-time image acquisition or integration into clinical practice, focusing instead on developing and validating the machine learning model [6]–[8]. One limitation of the research is the dependency on the quality and representativeness of the dataset, which may affect the generalizability of the findings to other populations and clinical settings.

This research contributes to the field by providing a scalable and efficient method for malaria diagnosis using machine learning techniques. The automated system developed in this study has the potential to reduce the burden on healthcare professionals, minimize diagnostic errors, and expedite treatment decisions. By demonstrating the feasibility of using decision trees for malaria detection, this study paves the way for further research and development of more sophisticated and integrated diagnostic tools. Ultimately, the findings could lead to improved malaria control and management, particularly in resource-limited settings where the disease is most prevalent.

2. Method

The research design of this study is based on a quantitative approach, utilizing machine learning techniques for image classification. The objective is to develop an automated system for detecting malaria from blood smear images using a decision tree classifier. The research follows a structured methodology that includes data preprocessing, feature extraction, model training, and evaluation. The segmentation process involves the use of the Canny edge detector to enhance the boundaries in the images, while feature extraction is performed using Hu Moments, a set of shape descriptors. The data is then normalized to ensure that the features have a mean of 0 and variance of 1. The

performance of the classifier is evaluated using accuracy, precision, recall, and F1-measure. A visual representation of the entire research process is illustrated in **Figure 1**.

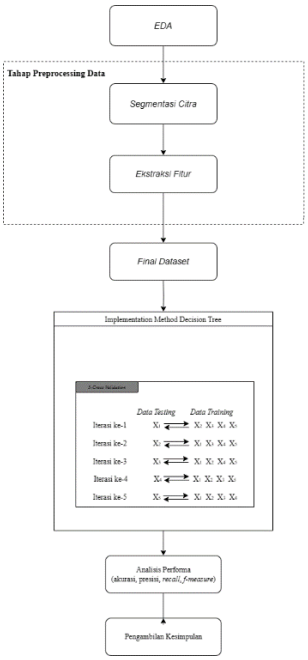

Figure 1: Decision Tree Evaluation Workflow

Sample or Data Selection:

The dataset used in this research is obtained from the National Institutes of Health (NIH) repository, consisting of 27,558 images categorized into two classes: Normal and Malaria. This dataset is balanced, with an equal number of images in each class, which helps in mitigating any bias during the training and evaluation of the model.

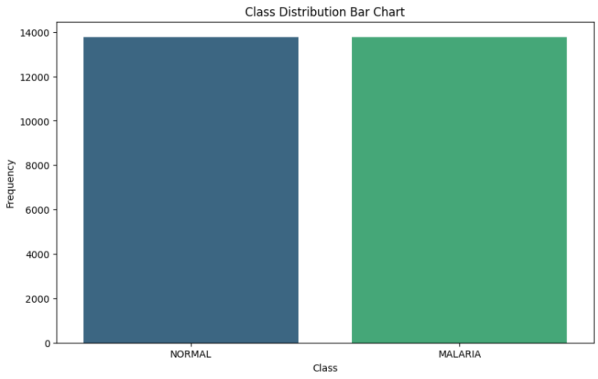

Figure 2: Class Distribution

The data is split into training and testing sets in an 80:20 ratio, ensuring that both sets are representative of the overall dataset. This split allows the model to be trained on a substantial amount of data while keeping enough data aside for robust testing and validation.

To visualize the data and provide an overview of the pre-processing steps, several plots and diagrams will be presented, including segmentation results, scatter plots of Hu Moments, boxplots, histograms, and correlation heatmaps. These visualizations will help in understanding the distribution and relationships of the features used in the classification process.

Tools and Technology Used:

The implementation of the research methodology is carried out using Python, a widely-used programming language for machine learning and data science. The following libraries are utilized:

- OpenCV for image processing and segmentation.
- Scikit-learn for machine learning algorithms and evaluation metrics.
- Matplotlib and Seaborn for data visualization.
- NumPy and Pandas for data manipulation and analysis.

These tools provide a comprehensive environment for developing and validating the machine learning model.

Data Collection Process

The data collection process involves downloading the blood smear image dataset from the NIH repository. The images are then pre-processed to enhance their quality and extract relevant features. The pre-processing steps include resizing the images to a standard size, applying the Canny edge detector for segmentation, and extracting Hu Moments for feature representation [9]–[11]. **Figure 3** and **4** to show the effect of the Canny edge detector on the images.

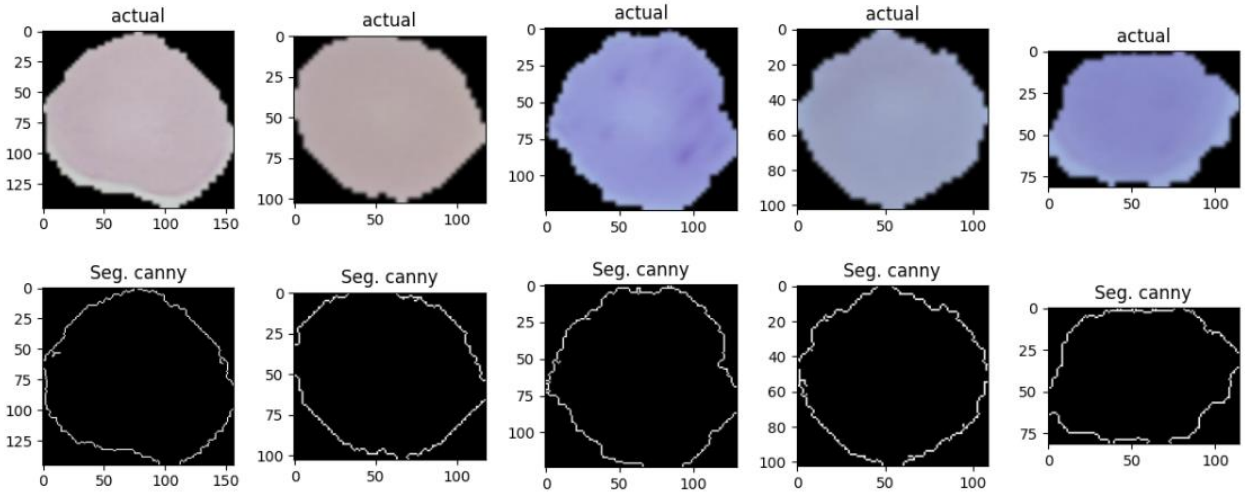

Figure 3: Normal Class

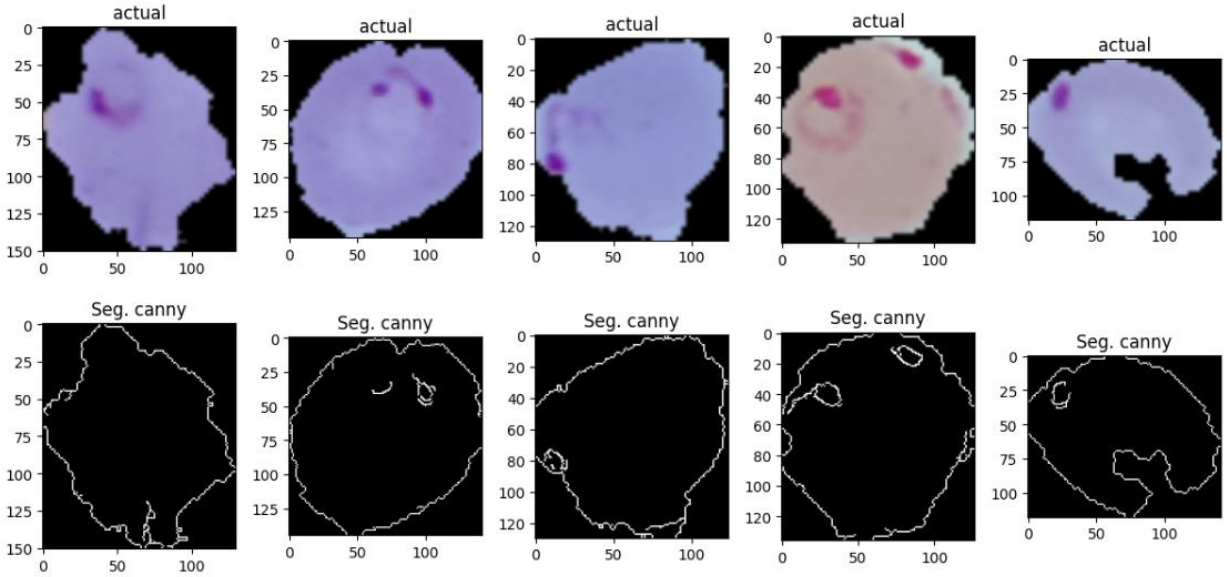

Figure 4: Malaria Class

The Hu Moments are mathematical descriptors that provide a compact representation of the shape characteristics of the segmented images [12], [13]. These moments are invariant to image transformations such as translation, scale, and rotation.

The formula for Hu Moments is given by:

$$Hu = [n_{20} + n_{02}, (n_{20} + n_{02})^2 + 4n_{11}^2, (n_{30} - 3n_{12})^2 + (3n_{21} - n_{03})^2, (n_{30} + n_{12})^2 + (n_{21} + n_{03})^2] \quad (1)$$

Where n_{ij} are the normalized central moments. **Figure 5** to visualize the relationships between different features, **Figure 6** to understand the distribution and variability of the features and **Figure 7** to identify the correlations between different Hu Moments.

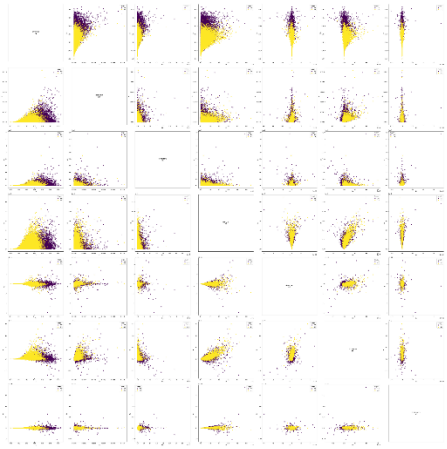

Figure 5: Scatter plots of all combinations of Hu Moments

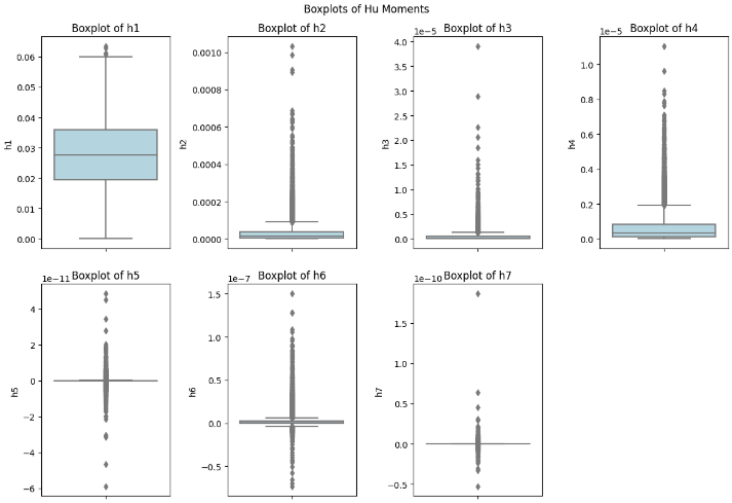

Figure 6: Boxplots and histograms of Hu Moments

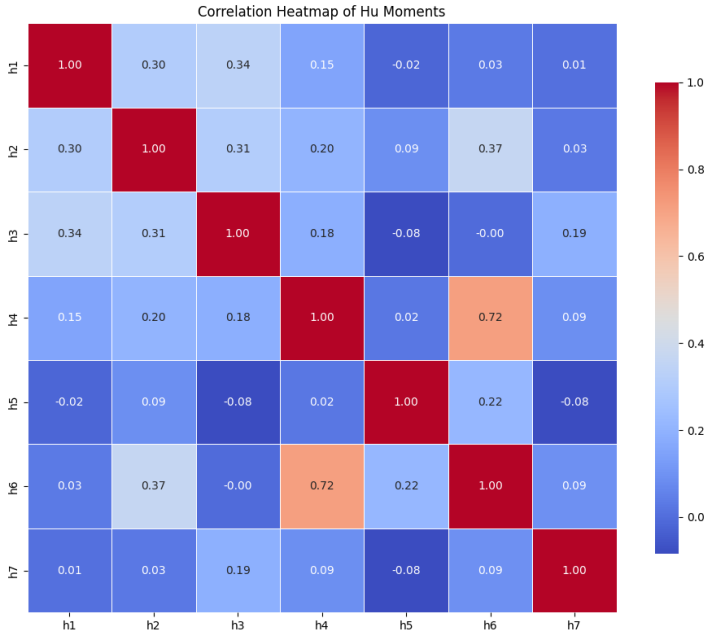

Figure 7: Correlation Heatmap

Data Analysis Methods

The data analysis involves training a decision tree classifier on the pre-processed and feature-extracted data. The decision tree algorithm is chosen for its simplicity and interpretability. A decision tree is a flowchart-like structure where each internal node represents a decision based on an attribute [14]–[16], each branch represents the outcome of a decision, and each leaf node represents a class label. **Figure 8** to compare the distributions of features between classes

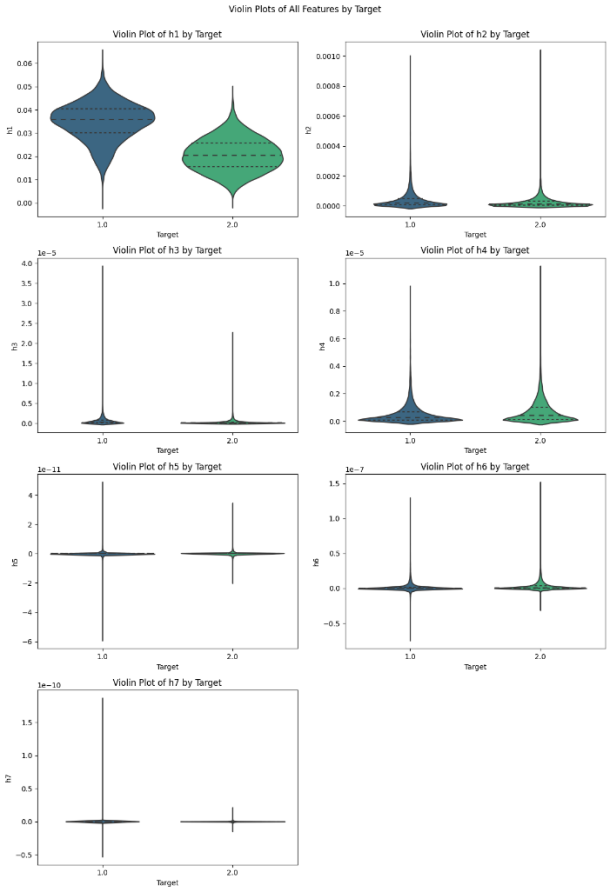

Figure 8: Violin Plots and Boxen Plots

The decision tree model recursively splits the data into subsets based on the value of the attributes, aiming to maximize the separation between classes. The decision tree algorithm uses a criterion to determine the best split at each node [17]. Common criteria include Gini impurity and entropy. The formula for Gini impurity is given by [18]–[20]:

$$Gini(D) = 1 - \sum_{i=1}^{n} p_i^2 \tag{1}$$

Where p_i is the probability of an element being classified into class i.

The entropy criterion, used in information gain, is given by:

$$Entrophy(D) = -\sum_{i=1}^{n} p_i \log_2(p_i) \tag{1}$$

Information gain, which measures the reduction in entropy from a split, is given by:

$$IG(D, A) = Entrophy(D) - \sum_{v \in Values(A)} \frac{|D_v|}{|D|} Entrophy(D_v) \tag{1}$$

Where D is the dataset, A is an attribute, $Values(A)$ are the possible values of A and D_v is the subset of D where A has value v.

The classifier is trained on the training set and evaluated on the test set using the following performance metrics [21]–[23]:

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

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

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \quad (3)$$

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

where TP, TN, FP, and FN represent the counts of true positives, true negatives, false positives, and false negatives, respectively.

3. Result and Discussion

The dataset used in this study consisted of 27,558 blood smear images, balanced across two classes: Normal and Malaria. The images were pre-processed using the Canny edge detector for segmentation, followed by feature extraction using Hu Moments. The features were normalized to have a mean of 0 and variance of 1. The decision tree classifier was then trained and evaluated using 5-fold cross-validation. The performance metrics were calculated for each fold, and the results were averaged to obtain the overall performance.

Table 1 summarizes the performance of the decision tree classifier across the five folds

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

K-n	**Performa**			
	Accuracy	***Precision***	***Recall***	***F-Measure***
K-1	77.65%	77.56%	77.41%	77.77%
K-2	76.94%	77.23%	77.23%	77.23%
K-3	78.08%	78.31%	77.98%	78.19%
K-4	76.61%	77.41%	77.07%	76.89%
K-5	77.29%	77.03%	77.18%	77.32%
$\sum$ ***Avg***	77.32%	77.31%	77.37%	77.48%

To visualize the performance of the decision tree classifier, graphical representations such as performance plots and a confusion matrix will be displayed. These visualizations will provide a clear understanding of how well the classifier distinguishes between the Normal and Malaria classes.

The performance plots illustrate the accuracy, precision, recall, and F1-Score across the five folds, while the confusion matrix provides a detailed view of the classifier's predictions against the actual class labels.

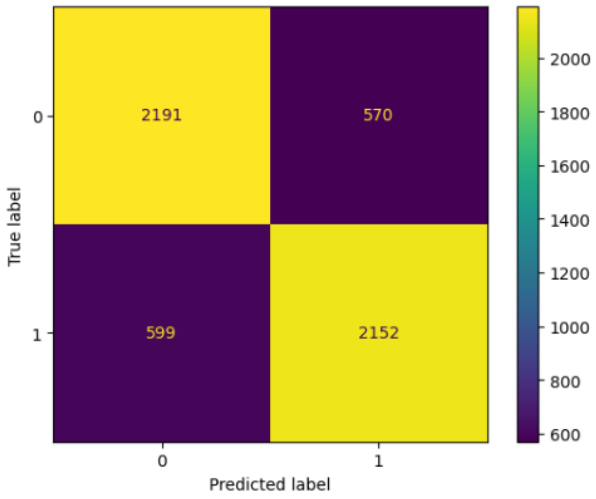

Figure 3: Confusion Matrix of the Decision Tree

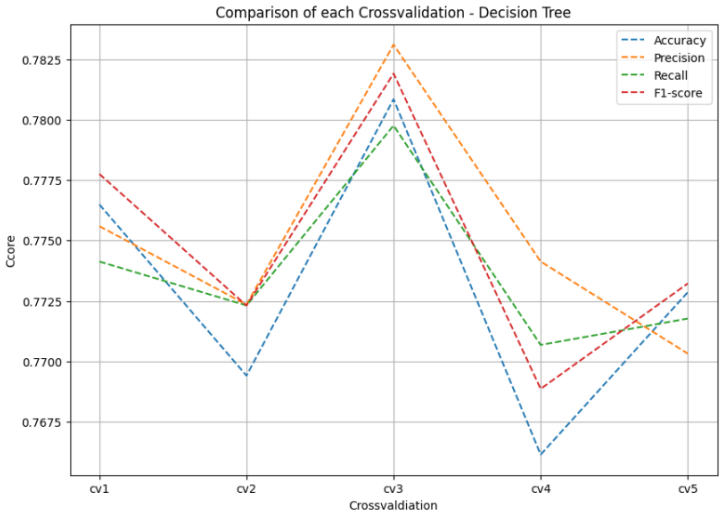

Figure 4: Performance Comparison of Each Cross-validation Fold

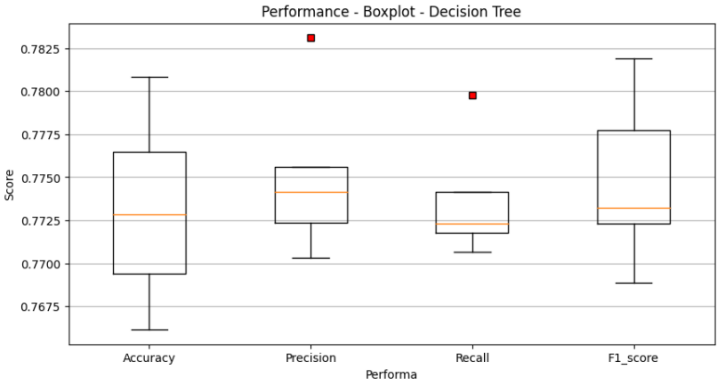

Figure 5: Performance Boxplot of the Decision Tree

Discussion

The data processing results indicate that the decision tree classifier achieved a consistent performance across all five folds, with an average accuracy of 77.32%. The precision, recall, and F1-Score were also consistently high, demonstrating the model's ability to effectively distinguish between malaria-infected and uninfected blood smear images. The segmentation using the Canny edge detector and feature extraction with Hu Moments proved effective in capturing the relevant characteristics of the images.The significant findings of this study are the robust performance metrics achieved by the decision tree classifier. The high precision and recall values indicate that the model is both accurate in its predictions and reliable in identifying true positive cases of malaria. These results suggest that the decision tree classifier can be a valuable tool for automated malaria diagnosis, potentially reducing the workload on healthcare professionals and increasing the speed and accuracy of diagnosis.

In comparison with previous research, the use of Hu Moments and decision tree classifiers in this study aligns with existing findings that shape-based features and tree-based classifiers can be highly effective for image classification tasks. The results reinforce the potential of machine learning techniques in medical image analysis, particularly for diseases like malaria that require rapid and accurate diagnosis. The practical implications of the research are significant. An automated system based on the proposed methodology could be deployed in clinical settings, providing a scalable solution for malaria diagnosis. This would be particularly beneficial in resource-limited regions where access to skilled laboratory personnel is limited.

However, the study has several limitations. The reliance on a single dataset means that the findings may not generalize to other datasets or real-world clinical settings without further validation. Additionally, the decision tree classifier, while interpretable, may not capture complex relationships between features as effectively as more advanced models such as random forests or deep neural networks. Future research should explore the use of more sophisticated models and validate the findings on a broader range of datasets. Real-time image acquisition and integration into clinical workflows should also be investigated to assess the practical feasibility of the proposed system. Further studies could also explore the combination of different feature extraction methods to enhance the robustness and accuracy of the classification.

4. Conclusion

In summary, this study successfully developed an automated system for malaria detection using blood smear images by leveraging the decision tree classifier. The preprocessing steps involved segmenting the images with the Canny edge detector and extracting Hu Moments as features, followed by normalizing the data. The classifier demonstrated consistent performance across five-fold cross-validation, achieving an average accuracy of 77.32%, precision of 77.31%, recall of 77.37%, and F1-Score of 77.48%. These results indicate the model's robustness and reliability in distinguishing between malaria-infected and uninfected images.

The research answered key questions regarding the effectiveness of segmentation and feature extraction techniques, and the decision tree classifier's diagnostic accuracy. The findings confirmed that Hu Moments are suitable features for this classification task, and the decision tree provides a simple yet effective model. The study contributes to the field by providing a scalable and efficient method for automated malaria diagnosis, which can potentially be

integrated into clinical practice to improve diagnostic accuracy and reduce workload on healthcare professionals. Future research should focus on validating the model on diverse datasets, exploring more advanced classification techniques, and integrating real-time image acquisition to enhance the practical applicability of the proposed system.

References:

[1] J. Trivedi, "Canny edge detection based real-time intelligent parking management system," *Sci. J. Silesian Univ. Technol. Ser. Transp.*, vol. 106, pp. 197–208, 2020, doi: 10.20858/sjsutst.2020.106.17.

[2] B. Iqbal, "Canny edge detection and Hough transform for high resolution video streams using Hadoop and Spark," *Cluster Comput.*, vol. 23, no. 1, pp. 397–408, 2020, doi: 10.1007/s10586-019-02929-x.

[3] 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, [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10418246/.

[4] H. Azis, D. Widyawati, and ..., "Prediksi potensi donatur menggunakan model Logistic Regression," *Indones. J. ...*, 2023, [Online]. Available: https://jurnal.yoctobrain.org/index.php/ijodas/article/view/64.

[5] H. Azis, P. Purnawansyah, N. Nirwana, and ..., "The Support Vector Regression Method Performance Analysis in Predicting National Staple Commodity Prices," *Ilk. J. ...*, 2023, [Online]. Available: https://jurnal.fikom.umi.ac.id/index.php/ILKOM/article/view/1686.

[6] 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.

[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, [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.

[8] H. Darwis, H. Azis, and Y. Salim, "Backpropagation Neural Network with Combination of Activation Functions for Inbound Traffic Prediction," *Knowl. Eng. Data Sci.*, vol. 4, no. 1, pp. 14–28, 2021, [Online]. Available: file:///Users/kbh/Downloads/20008-70582-1-PB.pdf.

[9] S. T. H. Kieu, "COVID-19 Detection Using Integration of Deep Learning Classifiers and Contrast-Enhanced Canny Edge Detected X-Ray Images," *IT Prof.*, vol. 23, no. 4, pp. 51–56, 2021, doi: 10.1109/MITP.2021.3052205.

[10] M. R. Sharan, "Classification of Medicinal Leaf by Using Canny Edge Detection and SVM Classifier," *2022 Int. Conf. Futur. Technol. INCOFT 2022*, 2022, doi: 10.1109/INCOFT55651.2022.10094461.

[11] S. K. T. Hwa, "Tuberculosis detection using deep learning and contrast-enhanced canny edge detected X-Ray images," *IAES Int. J. Artif. Intell.*, vol. 9, no. 4, pp. 713–720, 2020, doi: 10.11591/ijai.v9.i4.pp713-720.

[12] 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, [Online]. Available: https://jurnal.yoctobrain.org/index.php/ijodas/article/view/112.

[13] H. Oumarou and N. Rismayanti, "Automated Classification of Empon Plants: A Comparative Study Using Hu Moments and K-NN Algorithm," *Indones. J. Data ...*, 2023, [Online]. Available: https://jurnal.yoctobrain.org/index.php/ijodas/article/view/115.

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

[15] J. A. D. de Jesus Ferreira, "Decision tree classifiers for unmanned aircraft configuration selection," *Aircr. Eng. Aerosp. Technol.*, vol. 93, no. 6, pp. 1122–1132, 2021, doi: 10.1108/AEAT-03-2021-0074.

[16] I. A. P. Banlawe, "Decision Tree Learning Algorithm and Naïve Bayes Classifier Algorithm Comparative Classification for Mango Pulp Weevil Mating Activity," *2021 IEEE Int. Conf. Autom. Control Intell. Syst. I2CACIS 2021 - Proc.*, pp. 317–322, 2021, doi: 10.1109/I2CACIS52118.2021.9495863.

[17] D. Widyawati, A. Faradibah, and ..., "Comparison Analysis of Classification Model Performance in Lung Cancer Prediction Using Decision Tree, Naive Bayes, and Support Vector Machine," *Indones. J. ...*, 2023, [Online]. Available: https://jurnal.yoctobrain.org/index.php/ijodas/article/view/76.

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

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

[20] R. Rohan, "Classification of cardiac arrhythmia diseases from obstructive sleep apnea signals using decision tree classifier," *Int. J. Comput. Inf. Syst. Ind. Manag. Appl.*, vol. 12, pp. 248–264, 2020.

[21] N. Rismayanti, A. Naswin, U. Zaky, M. Zakariyah, and D. A. Purnamasari, "Evaluating Thresholding-Based Segmentation and Humoment Feature Extraction in Acute Lymphoblastic Leukemia Classification using Gaussian Naive Bayes," *Int. J. Artif. Intell. Med. Issues*, vol. 1, no. 2, 2023.

[22] A. Sinra, B. S. W. Poetro, H. Angriani, H. Zein, and ..., "Optimizing Neurodegenerative Disease Classification with Canny Segmentation and Voting Classifier: An Imbalanced Dataset Study," *... Artif. Intell. ...*, 2023, [Online]. Available: https://www.jurnal.yoctobrain.org/index.php/ijaimi/article/view/97.

[23] 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, [Online]. Available: https://www.jurnal.yoctobrain.org/index.php/ijaimi/article/view/87.