International Journal of Artificial Intelligence in Medical Issues

Volume 2 Issue 2
ISSN 3025-4167
https://doi.org/10.56705/ijaimi.v2i2.155

Research Article

Classification of Skin Diseases using Decision Tree Algorithm on an Imbalanced Dataset

Nurul Rismayanti [1*]; Sitti Fatimah Azzahrah [2]
[1] Universitas Negeri Malang, nurulrismayanti.labfik@umi.ac.id
[1] Universitas Muhammadiyah Pare-pare, 220280052sittifatimahazzahrad@gmail.com
Correspondence should be addressed to Nurul Rismayanti; nurulrismayanti.labfik@umi.ac.id

Received 18 September 2024; Revised 22 September 2024; Accepted 10 October 2024; Published 30 November 2024

Abstract:
Skin infections caused by pathogens such as bacteria and fungi are common and can lead to serious health complications if not properly managed. Accurate classification of these infections is crucial for effective treatment and management. This study focuses on classifying two skin diseases, Chickenpox and Shingles, using a Decision Tree algorithm applied to an imbalanced dataset sourced from Kaggle. The dataset, which is imbalanced by nature, was split into training (80%) and testing (20%) subsets. Pre-processing involved segmentation using Thresholding to isolate regions of interest and feature extraction using Hu Moments to capture shape characteristics of the lesions. The dataset was scaled to ensure that all features had a mean of 0 and variance of 1. The classifier's performance was evaluated using 5-fold cross-validation, yielding a mean accuracy of 66.06%, with precision, recall, and F1-scores indicating moderate performance. The study highlights the challenges posed by imbalanced datasets and the limitations of the Decision Tree algorithm in this context. The results underscore the importance of proper pre-processing and feature extraction but also suggest the need for more advanced classification techniques and data balancing methods. This research contributes to the field by providing a detailed methodology and comprehensive evaluation metrics, offering insights into the application of machine learning for medical image classification. Future work should focus on improving classifier performance through data augmentation, advanced feature extraction, and exploring other machine learning models better suited for imbalanced datasets.

Keywords: Skin Diseases, Decision Tree, Imbalanced Dataset, Feature Extraction, Machine Learning.

Dataset link: https://www.kaggle.com/datasets/subirbiswas19/skin-disease-dataset

1. Introduction

Skin infections are a prevalent health concern caused by various pathogens, including bacteria and fungi. These pathogens are often present on the skin, and while typically harmless, they can cause infections when their numbers increase beyond the immune system's capacity to manage them [1]. Proper diagnosis and treatment of skin infections are essential to prevent complications and ensure effective management. This study focuses on two specific skin diseases: Chickenpox and Shingles. Chickenpox, caused by the varicella-zoster virus, is primarily a childhood illness, while Shingles, a reactivation of the same virus, predominantly affects adults. Accurate differentiation between these diseases is critical for appropriate treatment and prevention strategies.

The primary problem addressed in this research is the challenge of accurately classifying skin diseases using machine learning techniques on an imbalanced dataset [2]. Imbalanced datasets are common in medical imaging, where some conditions are more prevalent than others, leading to a skewed distribution of classes. This imbalance can significantly affect the performance of machine learning algorithms, causing them to favor the majority class and

resulting in poor recognition of the minority class. In this study, the dataset consists of images of Chickenpox and Shingles, with an inherent imbalance in the number of images for each class. To address this problem, this research aims to develop a robust classification model for skin diseases using the Decision Tree algorithm. The specific objectives are to preprocess the dataset using segmentation techniques, extract relevant features, and scale the data to standardize the input for the classifier. The Decision Tree algorithm [3]–[5] is chosen for its simplicity and effectiveness in handling categorical data. The performance of the classifier will be evaluated using accuracy, precision, recall, and F1-measure [6]–[8], providing a comprehensive assessment of its capability to distinguish between Chickenpox and Shingles.

This study seeks to answer several key research questions: Can the Decision Tree algorithm effectively classify an imbalanced dataset of skin disease images? How does the performance of the classifier, measured by accuracy, precision, recall, and F1-measure, vary under these conditions? Additionally, the research hypothesizes that proper preprocessing and feature extraction techniques can significantly enhance the classifier's performance, even in the presence of data imbalance. These questions are crucial for understanding the limitations and potential of using machine learning in medical image classification. The scope of this research is confined to the classification of Chickenpox and Shingles images obtained from the Kaggle dataset. The study's limitations include the inherent imbalance of the dataset and the potential for variability in image quality and resolution. While the Decision Tree algorithm provides a baseline for performance, future research could explore more advanced classifiers and data balancing techniques to further improve classification accuracy. The findings from this study are intended to provide insights into the challenges and solutions for classifying imbalanced medical image datasets [7], [9].

In terms of research contributions, this study offers a detailed methodology for pre-processing and classifying skin disease images, which can be a valuable reference for future studies in the field. The evaluation metrics provide a comprehensive understanding of the classifier's performance, highlighting areas for improvement and potential applications in medical diagnostics. By addressing the issue of data imbalance and exploring effective feature extraction techniques, this research contributes to the broader goal of enhancing the accuracy and reliability of machine learning models in healthcare.

2. Method

This study employs an experimental research design focusing on the classification of skin disease images, specifically Chickenpox and Shingles, using machine learning techniques. The research design includes image pre-processing, feature extraction, data scaling, and classification using the Decision Tree algorithm [8]. The performance of the classification model is evaluated using metrics such as accuracy, precision, recall, and F1-measure [10]. The overall goal is to develop a robust methodology that can handle the challenges posed by imbalanced datasets in medical imaging. A visual depiction of the complete research process is shown in **Figure 1**.

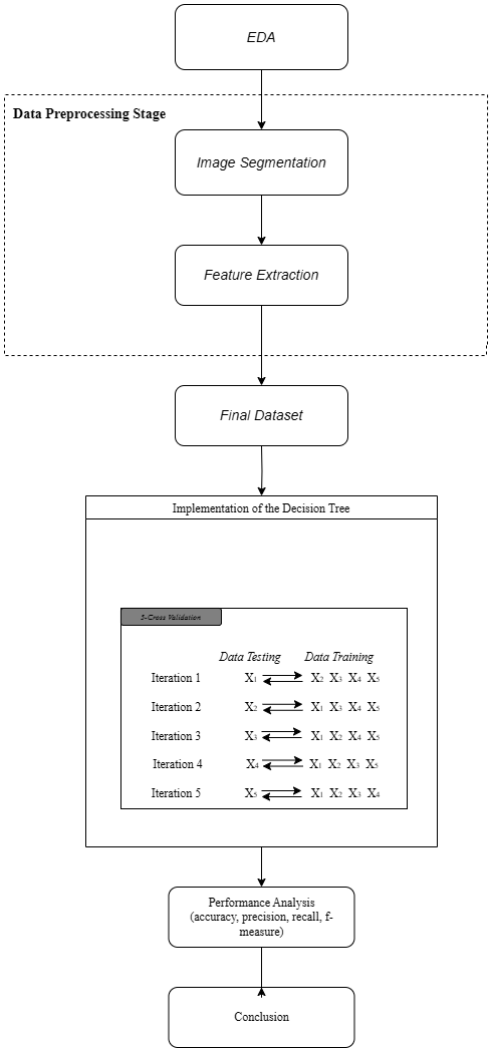

Figure 1: Decision Tree Assessment Workflow

Sample or Data Selection:

The dataset used in this study is sourced from Kaggle, a well-known platform for datasets and machine learning competitions. The dataset contains images of two skin diseases: Chickenpox and Shingles. The images were collected from various sources on the internet. The dataset is inherently imbalanced, with a higher number of images for one class compared to the other. This imbalance poses a significant challenge for the classification algorithm and necessitates careful handling during the pre-processing and training stages.

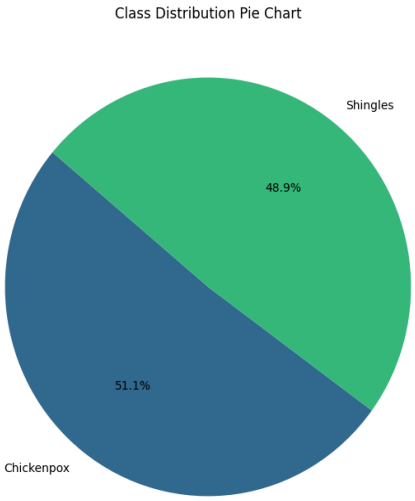

Figure 2: Class Distribution

Data Collection Process

The dataset is split into training and testing subsets, with 80% of the data allocated for training and 20% for testing [11]. This split ensures that the model has enough data to learn from while still being able to generalize to unseen examples. The images undergo a series of pre-processing steps, starting with segmentation using the Thresholding technique [12], [13]. This method isolates the regions of interest in the images, making it easier to extract relevant features [14], [15].

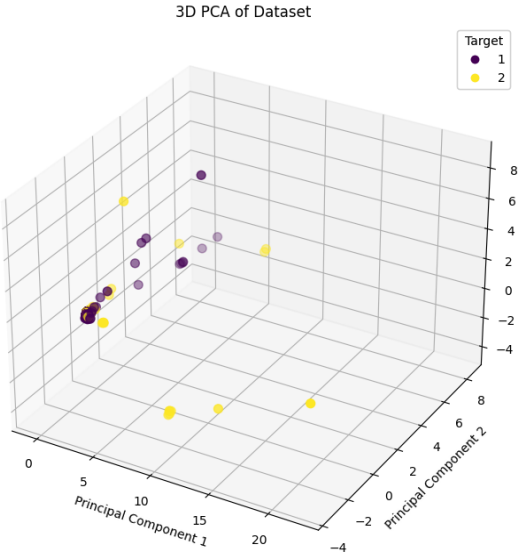

Figure 3: 3D PCA plot of the dataset

Data Analysis Methods

The data analysis methods in this study involve several key steps, each crucial for the successful classification of the skin disease images.

a. Segmentation

The segmentation process uses the Thresholding technique to isolate the regions of interest in the images [16]. The threshold value is chosen to maximize the contrast between the foreground (skin lesions) and the background. Mathematically, this can be expressed as:

$$T = \frac{1}{N}\sum_{i=1}^{N} I(x_i, y_i) \tag{1}$$

Where T is the threshold value, $I(x_i, y_i)$ is the intensity of pixel (x_i, y_i) , and N is the total number of pixels.

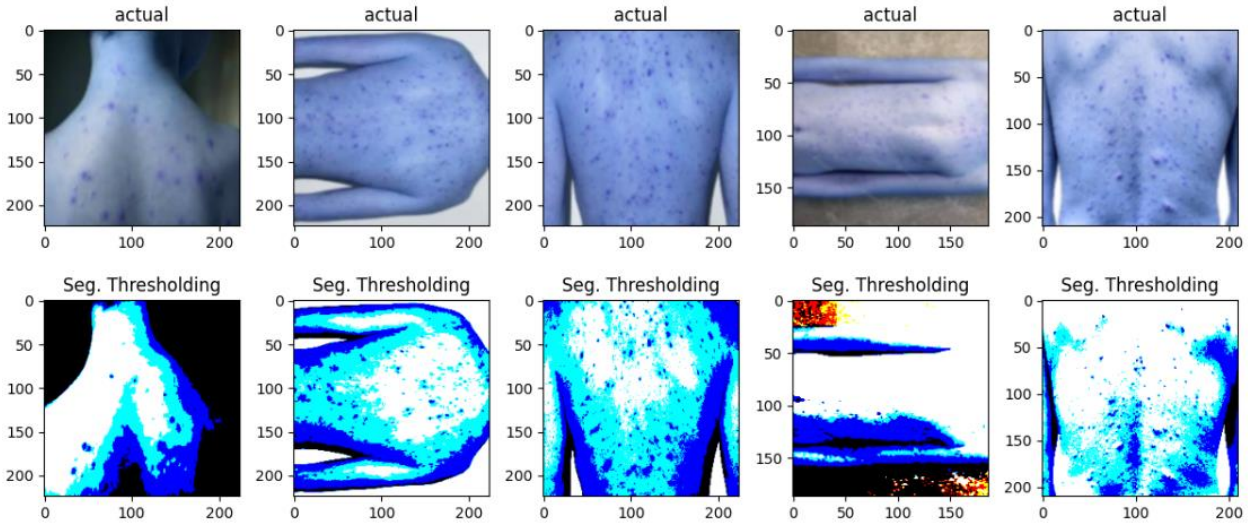

Figure 3: Class Chickenpox

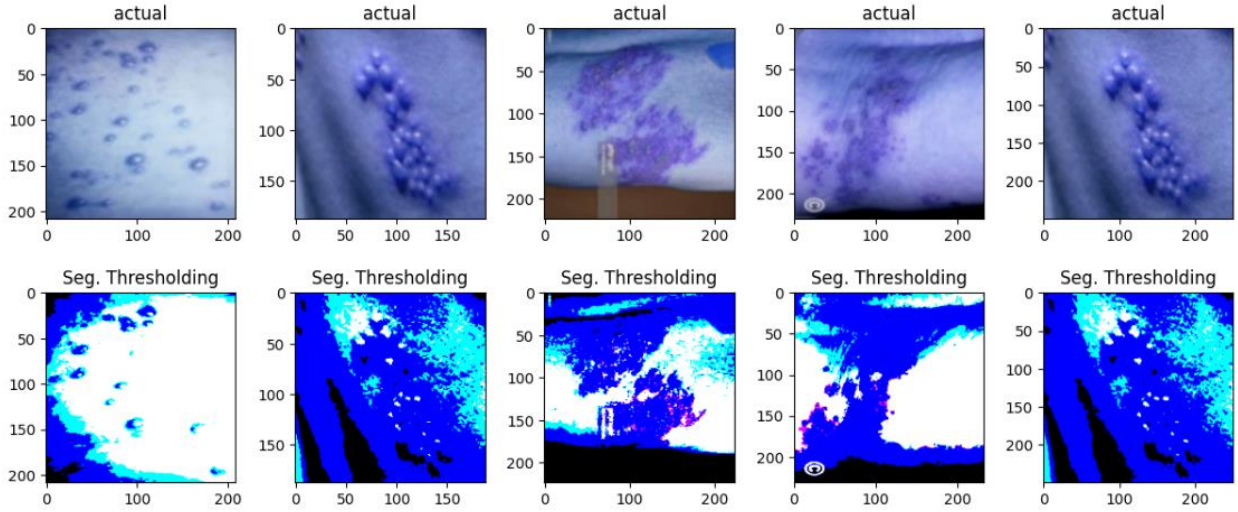

Figure 4: Class Shingles

b. Feature Extraction

Feature extraction is performed using Hu Moments, which are shape descriptors invariant to image transformations [17], [18]. Hu Moments are derived from the central moments of the image. The seven Hu Moments are calculated as follows [19]:

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

Where n_{ij} are the normalized central moments of the image. **Figure 5** visualizes the statistical distribution of Hu Moments for each class. **Figure 6** displays the frequency distribution of Hu Moments for each class.

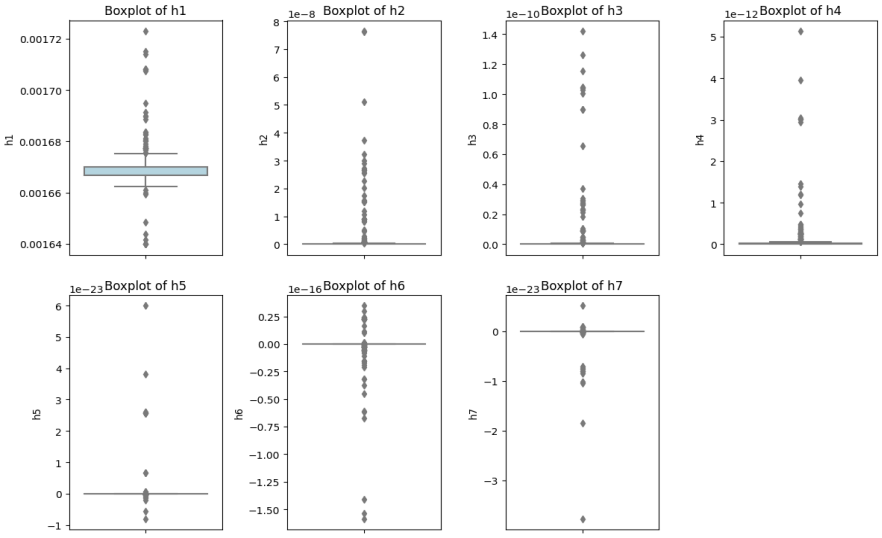

Figure 5: Boxplots of Hu Moments

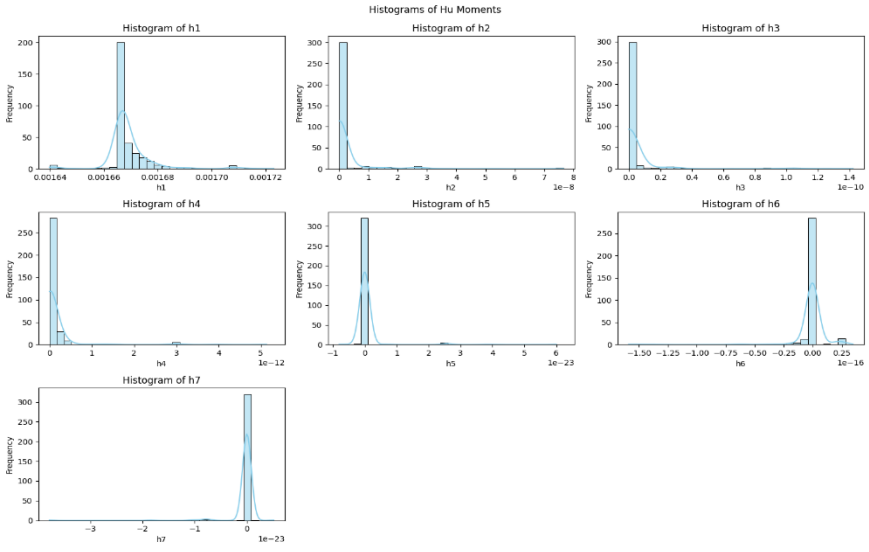

Figure 6: Histogram of Hu Moments

c. Data Scaling

To ensure that the features have a mean of 0 and variance of 1, the dataset is scaled using standardization. This process can be represented as:

$$X' = \frac{X - \mu}{\sigma} \tag{3}$$

Where X is the original feature value, μ is the mean of the feature values, and σ is the standard deviation. **Figure 7** highlights the correlation between different Hu Moments.

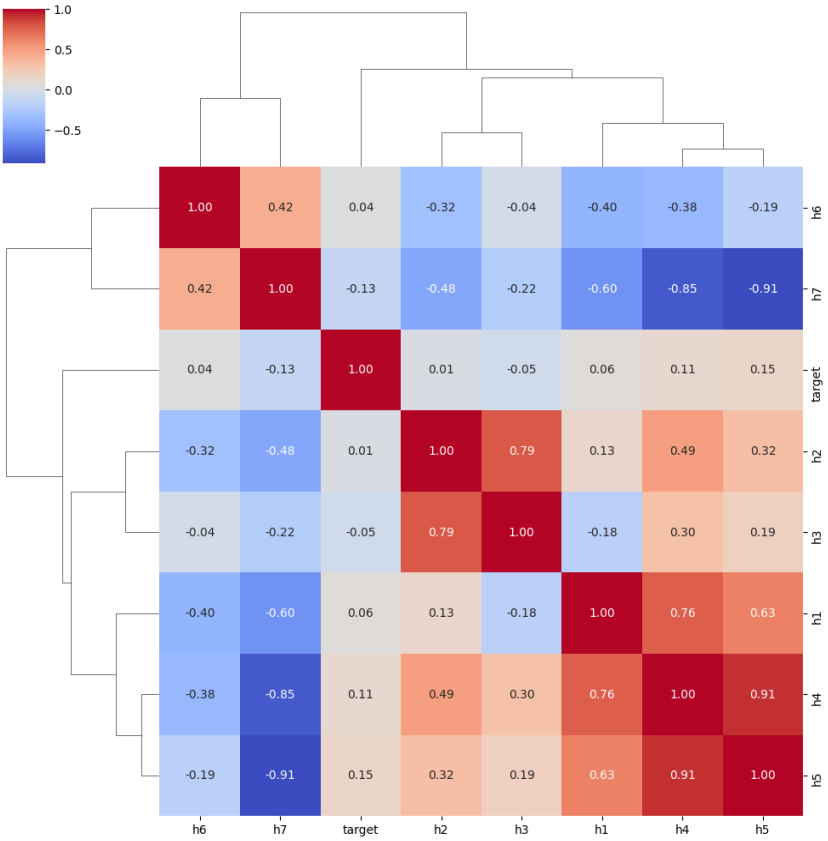

Figure 7: Correlation Heatmap of Hu Moments

d. Classification

The Decision Tree algorithm is used for classification. Decision Trees split the data based on feature values to create a tree structure, where each node represents a decision rule and each leaf node represents an outcome [4], [5]. The Gini impurity is used to determine the best splits [3], [20], [21]:

$$Gini(D) = 1 - \sum_{i=1}^{n} P_i^2 \tag{4}$$

Where D is the dataset, n is the number of classes, p_i and is the probability of class i.

e. Performance Evaluation

The classifier's performance is evaluated using the following metrics [22]:

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

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

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \quad (5)$$

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

3. Result and Discussion

. The data processing steps involved several key stages: segmentation using Thresholding, feature extraction using Hu Moments, and scaling of the dataset. Each image was segmented to isolate the regions of interest, which are the skin lesions in this case. Hu Moments, which are invariant to image transformations, were then extracted from the segmented images to serve as features for classification. Finally, the dataset was scaled to ensure that all features had a mean of 0 and a variance of 1, improving the performance of the Decision Tree classifier.

To evaluate the classifier, we used 5-fold cross-validation, which involves splitting the dataset into five subsets, training the model on four subsets, and testing it on the remaining subset. This process was repeated five times, with each subset used exactly once as the test set. The performance metrics—accuracy, precision, recall, and F1-score—were computed for each fold. The results of these metrics are summarized in the **Table 1**.

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

K-n	**Metrics**			
	Accuracy	***Precision***	***Recall***	***F-Measure***
K-1	65.67%	68.64%	65.67%	64.44%
K-2	65.67%	67.73%	65.67%	64.81%
K-3	68.66%	69.60%	68.66%	68.36%
K-4	62.12%	63.41%	62.12%	61.57%
K-5	68.18%	71.87%	68.18%	67.11%
$\sum \boldsymbol{Avg}$	66.06%	68.25%	66.06%	65.26%

The visualization of the performance metrics and the confusion matrix for the Decision Tree classifier are presented below. These visualizations provide a detailed view of how well the classifier distinguishes between Chickenpox and Shingles.

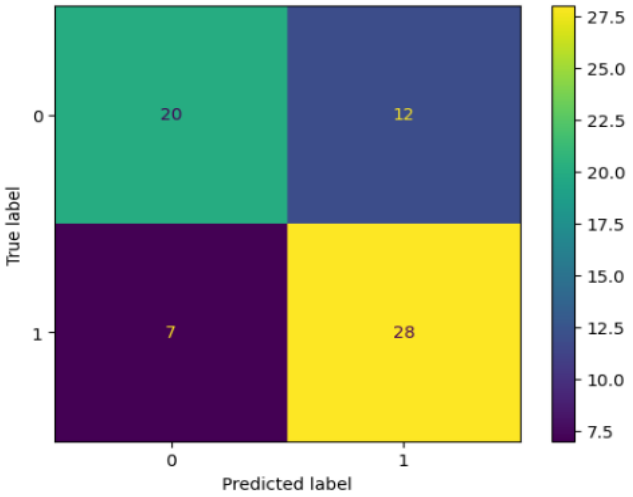

Figure 8: Confusion Matrix of the Decision Tree

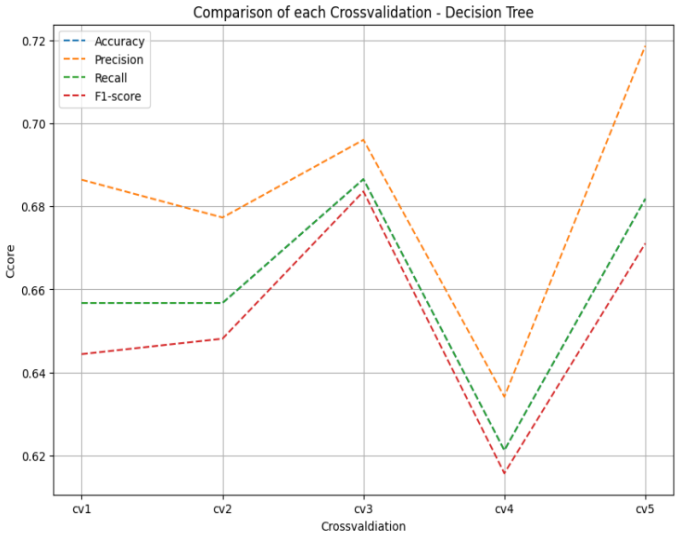

Figure 9: Performance Comparison of Each Cross-validation Fold

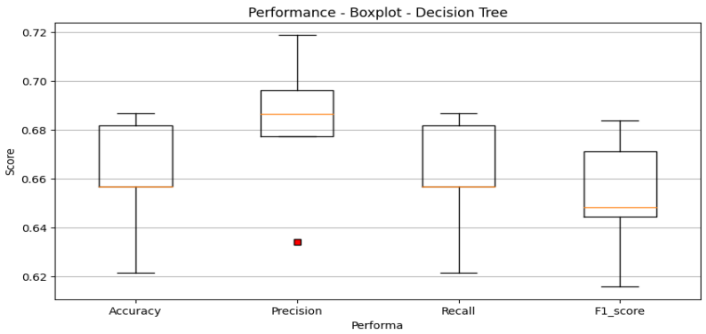

Figure 10: Performance Boxplot of the Decision Tree

Discussion

The results from the 5-fold cross-validation reveal several important insights. The mean accuracy of the Decision Tree classifier was approximately 66.06%, indicating a moderate level of performance. Precision, which measures the proportion of true positive predictions among all positive predictions, averaged 68.25%. Recall, which measures the proportion of true positive predictions among all actual positives, also averaged 66.06%. The F1-score, which is the harmonic mean of precision and recall, averaged 65.26%. These results suggest that the classifier performs reasonably well but has room for improvement, particularly in handling the imbalanced nature of the dataset.

Interpreting these results, it is evident that the Decision Tree classifier is somewhat effective in distinguishing between Chickenpox and Shingles. However, the moderate values of precision and recall indicate that the classifier occasionally misclassifies images, particularly those belonging to the minority class. This is a common issue in imbalanced datasets, where the classifier tends to be biased towards the majority class. Significant findings from this study include the relatively consistent performance across different folds, indicating that the model is stable but not highly accurate. The segmentation and feature extraction techniques used were effective in preparing the data for classification, but the Decision Tree algorithm's limitations are apparent. These findings are consistent with previous research that highlights the challenges of using decision trees on imbalanced datasets.

In terms of practical implications, the results suggest that while the Decision Tree algorithm can be used for initial classification tasks, further refinement is necessary. Techniques such as data balancing, advanced feature extraction, or using more sophisticated classifiers like ensemble methods or deep learning models could potentially improve performance. The main limitation of this research is the imbalanced nature of the dataset, which affects the classifier's performance. Future research should focus on addressing this limitation by exploring data augmentation techniques, synthetic data generation, or using different machine learning algorithms better suited for imbalanced datasets.

Based on these results, several recommendations for future research can be made. First, implementing data balancing techniques such as oversampling the minority class or under sampling the majority class could help improve classifier performance. Second, exploring more advanced classifiers, including ensemble methods like Random Forests or Gradient Boosting, could provide better results. Finally, incorporating additional features or using deep learning approaches might enhance the model's ability to accurately classify skin diseases. Visualizations of the performance metrics and confusion matrix provide further insights into the classifier's performance, highlighting areas where misclassifications occur and offering a visual representation of the classifier's effectiveness. These visualizations are crucial for understanding and interpreting the results, and they offer a foundation for further improvements in classification accuracy.

4. Conclusion

In summary, this study aimed to classify skin diseases, specifically Chickenpox and Shingles, using a Decision Tree algorithm on an imbalanced dataset. The results from the 5-fold cross-validation showed that the classifier achieved a mean accuracy of 66.06%, with precision, recall, and F1-scores indicating moderate performance. The findings highlight the challenges of using decision trees on imbalanced datasets, as the classifier showed a tendency

to misclassify images, particularly those from the minority class. The study successfully demonstrated the effectiveness of segmentation using Thresholding and feature extraction using Hu Moments, but also underscored the limitations of the Decision Tree algorithm in this context.

The research questions addressed whether the Decision Tree algorithm can effectively classify an imbalanced dataset and how its performance metrics vary under these conditions. The results suggest that while the Decision Tree algorithm has potential, it requires further refinement to improve its accuracy and reliability. This research contributes to the field by providing a detailed methodology for pre-processing and classifying skin disease images and offering insights into the performance of machine learning algorithms on imbalanced medical datasets. For future research, it is recommended to explore data balancing techniques, more sophisticated classifiers, and additional feature extraction methods to enhance classification accuracy. Implementing these improvements could significantly advance the application of machine learning in medical diagnostics, leading to better disease management and treatment outcomes.

References:

[1] C. R. Dhivyaa, "Skin lesion classification using decision trees and random forest algorithms," *J. Ambient Intell. Humaniz. Comput.*, 2020, doi: 10.1007/s12652-020-02675-8.

[2] M. A. Febriantono, "Classification of multiclass imbalanced data using cost-sensitive decision tree c5.0," *IAES Int. J. Artif. Intell.*, vol. 9, no. 1, pp. 65–72, 2020, doi: 10.11591/ijai.v9.i1.pp65-72.

[3] M. M. Ghiasi, "Decision tree-based diagnosis of coronary artery disease: CART model," *Comput. Methods Programs Biomed.*, vol. 192, 2020, doi: 10.1016/j.cmpb.2020.105400.

[4] D. Jalal, "Decision Tree and Support Vector Machine for Anomaly Detection in Water Distribution Networks," *2020 International Wireless Communications and Mobile Computing, IWCMC 2020*. pp. 1320–1323, 2020, doi: 10.1109/IWCMC48107.2020.9148431.

[5] O. J. Alajas, "Prediction of Grape Leaf Black Rot Damaged Surface Percentage Using Hybrid Linear Discriminant Analysis and Decision Tree," *2021 International Conference on Intelligent Technologies, CONIT 2021*. 2021, doi: 10.1109/CONIT51480.2021.9498518.

[6] T. E. Tarigan, E. Susanti, M. I. Siami, I. Arfiani, and ..., "Performance Metrics of AdaBoost and Random Forest in Multi-Class Eye Disease Identification: An Imbalanced Dataset Approach," *... Artif. Intell. ...*, 2023.

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

[8] A. Naswin and A. P. Wibowo, "Performance Analysis of the Decision Tree Classification Algorithm on the Pneumonia Dataset," *... Artif. Intell. Med. ...*, 2023.

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

[10] N. Litha and T. Hasanuddin, "Analisis Performa Metode Moving Average Model untuk Prediksi Jumlah Penderita Covid-19," *Indones. J. Data Sci.*, vol. 1, no. 3, pp. 87–95, 2020, doi: https://doi.org/10.33096/ijodas.v1i3.19.

[11] H. Azis, D. Widyawati, and ..., "Prediksi potensi donatur menggunakan model Logistic Regression," *Indones.*

J. ..., 2023.

[12] D. Lee, "Threshold-based quantification of fatty degeneration in the supraspinatus muscle on MRI as an alternative method to Goutallier classification and single-voxel MR spectroscopy," *BMC Musculoskelet. Disord.*, vol. 21, no. 1, 2020, doi: 10.1186/s12891-020-03400-4.

[13] J. Amin, "Diagnosis of COVID-19 infection using three-dimensional semantic segmentation and classification of computed tomography images," *Comput. Mater. Contin.*, vol. 68, no. 2, pp. 2451–2467, 2021, doi: 10.32604/cmc.2021.014199.

[14] F. Ramlie, "Classification performance of thresholding methods in the Mahalanobis–Taguchi system," *Appl. Sci.*, vol. 11, no. 9, 2021, doi: 10.3390/app11093906.

[15] Y. Liu, "Automatic multi-label ecg classification with category imbalance and cost-sensitive thresholding," *Biosensors*, vol. 11, no. 11, 2021, doi: 10.3390/bios11110453.

[16] 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.

[17] 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.

[18] 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.

[19] Y. Jusman, "Classification System for Leukemia Cell Images based on Hu Moment Invariants and Support Vector Machines," *Proc. - 2021 11th IEEE Int. Conf. Control Syst. Comput. Eng. ICCSCE 2021*, pp. 137–141, 2021, doi: 10.1109/ICCSCE52189.2021.9530974.

[20] R. Hazra, "Machine Learning for Breast Cancer Classification with ANN and Decision Tree," *11th Annual IEEE Information Technology, Electronics and Mobile Communication Conference, IEMCON 2020*. pp. 522–527, 2020, doi: 10.1109/IEMCON51383.2020.9284936.

[21] S. H. Asman, "Decision tree method for fault causes classification based on rms-dwt analysis in 275 kv transmission lines network," *Appl. Sci.*, vol. 11, no. 9, 2021, doi: 10.3390/app11094031.

[22] H. Azis and S. R. Jabir, "Chemical Composition and Aroma Profiling: Decision Tree Modeling of Formalin Tofu," *J. Embed. Syst. Secur. ...*, 2023.