International Journal of Artificial Intelligence in Medical Issues

Volume 1 Issue 2
ISSN 3025-4167
https://doi.org/10.56705/ijaimi.v1i2.149

Research Article

Detection and Classification of Bacterial Skin Infections Using K-Nearest Neighbors Algorithm

Hayatou Oumarou [1*]

[1] Department of Computer Science, HTTC, University of Maroua, Cameroon, oumarou.hayatou@univ-maroua.cm
Correspondence should be addressed to Hayatou Oumarou; oumarou.hayatou@univ-maroua.cm

Received 15 September 2023; Revised 08 October 2023; Accepted 17 October 2023; Published 30 November 2023

Abstract:
Bacterial skin infections, including cellulitis and impetigo, pose significant health challenges requiring timely and accurate diagnosis for effective treatment. This research aims to develop an automated classification system for these infections using image processing and machine learning techniques. The study utilizes the Sobel method for image segmentation and Hu Moments for feature extraction. The classification is performed using the K-Nearest Neighbors (K-NN) algorithm with $k = 7$. The dataset, sourced from Kaggle, consists of imbalanced images of the two infection types. After pre-processing and feature extraction, the dataset is scaled to zero mean and unit variance. The model's performance is evaluated using cross-validation, yielding mean accuracy, precision, recall, F1-score, and ROC-AUC values of 65.95%, 65.18%, 65.95%, 63.06%, and 64.13%, respectively. Visualizations, including scatter plots, boxplots, histograms, correlation heatmaps, PCA, t-SNE, and UMAP, provide insights into the feature distributions and separability of classes. The results indicate that the combination of Sobel segmentation, Hu Moments, and K-NN can effectively classify bacterial skin infections. The study's contributions include demonstrating the applicability of these techniques to dermatological diagnostics and highlighting the potential for improved diagnostic accuracy and efficiency. However, the study acknowledges limitations such as data imbalance and variability in performance, suggesting the need for further research using advanced models like convolutional neural networks (CNNs) and enhanced data pre-processing techniques. These findings underscore the importance of machine learning in developing practical tools for clinical use, ultimately improving patient outcomes through early and accurate diagnosis.

Keywords: Bacterial Skin Infections, Image Processing, K-Nearest Neighbors, Hu Moments, Machine Learning.

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

1. Introduction

Bacterial skin infections represent a significant health concern worldwide, causing a range of symptoms from mild discomfort to severe systemic conditions [1]. These infections occur when bacteria, either from an external source or already present on the skin, invade through hair follicles or wounds. Common types include cellulitis and impetigo, which can lead to serious health complications if not promptly diagnosed and treated. Traditional diagnostic methods rely heavily on clinical evaluation, which can be subjective and time-consuming, highlighting the need for more objective and automated approaches. This study aims to address this gap by leveraging image processing and machine learning techniques to develop an automated classification system for bacterial skin infections [2].

The primary problem to be solved is the accurate and efficient classification of bacterial skin infections using medical images. Traditional diagnostic approaches are not only time-intensive but also prone to human error. Moreover, early and accurate diagnosis is crucial to preventing the spread of these infections and initiating appropriate

treatment. Therefore, the development of an automated system that can accurately classify skin infections from images would be highly beneficial. This research focuses on two common types of bacterial skin infections: cellulitis and impetigo. Both conditions have distinct visual characteristics that can be exploited using advanced image processing techniques.

The research objectives are threefold. First, to pre-process the skin infection images using the Sobel method for effective segmentation. Second, to extract significant features from the segmented images using Hu Moments, which are invariant to image transformations [3]. Third, to classify the images using the K-Nearest Neighbors (K-NN) [4] algorithm and evaluate the performance of this classification model using various metrics such as accuracy, precision, recall, F1-measure, and ROC-AUC [5]. By achieving these objectives, the study aims to contribute a reliable and automated diagnostic tool for bacterial skin infections, enhancing the efficiency and accuracy of clinical diagnostics.

The research is guided by specific questions and hypotheses. The central question is whether the combination of Sobel segmentation, Hu Moments feature extraction, and K-NN classification can effectively distinguish between cellulitis and impetigo in medical images. The hypothesis posits that this integrated approach will yield high accuracy and robust performance metrics, outperforming traditional diagnostic methods. This hypothesis is based on the premise that advanced image processing and machine learning techniques can capture the distinct visual features of different bacterial skin infections more accurately than manual examination.

This study's scope is focused on bacterial skin infections, specifically cellulitis and impetigo, as represented in the dataset obtained from Kaggle. The dataset is imbalanced, reflecting the real-world prevalence of these conditions, which presents an additional challenge for the classification algorithm. The research does not cover other types of skin conditions or infections, nor does it address treatment protocols. Furthermore, while the study employs the K-NN algorithm, future research could explore other machine learning models to compare performance. These limitations provide a clear framework within which the research is conducted, ensuring a focused and manageable study scope.

The contributions of this research are multifaceted. It introduces a novel application of image processing and machine learning to the field of dermatology, specifically for the classification of bacterial skin infections. The study demonstrates the effectiveness of using Sobel segmentation and Hu Moments for feature extraction in medical images, contributing to the existing body of knowledge in image processing. Additionally, by validating the K-NN algorithm for this classification task, the research provides a foundation for further exploration of machine learning techniques in medical diagnostics. Ultimately, the developed model offers a practical tool that can be integrated into clinical workflows, aiding healthcare professionals in making timely and accurate diagnoses, and improving patient outcomes.

2. Method

The research design involves developing an automated classification system for bacterial skin infections using image processing and machine learning techniques. The system is designed to pre-process images, extract features, and classify images into cellulitis and impetigo. The research utilizes the Sobel method for image segmentation [6], Hu Moments for feature extraction [7], and the K-Nearest Neighbors (K-NN) algorithm for classification. The performance of the classification model is evaluated using metrics such as accuracy, precision, recall, F1-measure,

and ROC-AUC [8], [9]. The overall workflow is structured to ensure robustness and reliability in diagnosing bacterial skin infections. The entirety of the research process is visually represented in **Figure 1**.

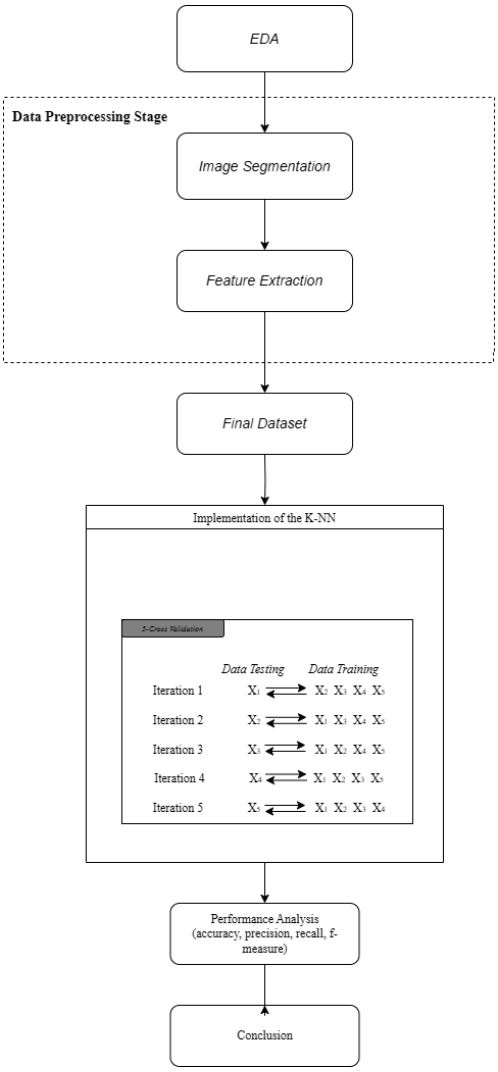

Figure 1: Workflow for Evaluating a K-NN

Sample or Data Selection:

The dataset used in this research is obtained from Kaggle and consists of images of bacterial skin infections. The dataset is imbalanced, containing more instances of one class compared to the other, reflecting real-world scenarios. The dataset includes two classes:

- Class 1: Cellulitis
- Class 2: Impetigo

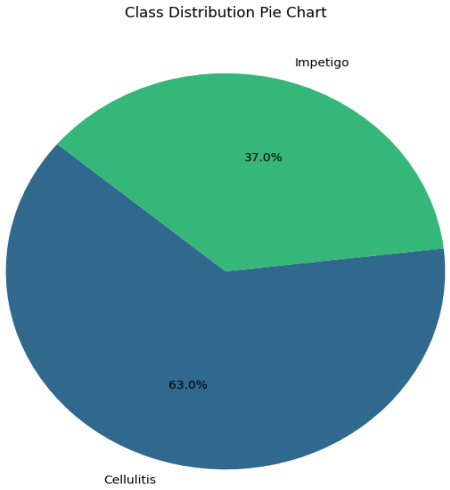

Figure 2: Class Distribution

To address the imbalance and ensure effective training of the machine learning model, the dataset is split into training and testing sets, with 80% of the data used for training and 20% for testing.

Data Collection Process

The data collection process involves downloading the dataset from Kaggle and preparing it for analysis. The images are pre-processed using the Sobel method for edge detection, which enhances the significant features of the images. The Sobel operator is a discrete differentiation operator that computes an approximation of the gradient of the image intensity function [10]–[13]. The gradient magnitude is given by:

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

Where G_x and G_y are the gradients in the x and y directions, respectively.

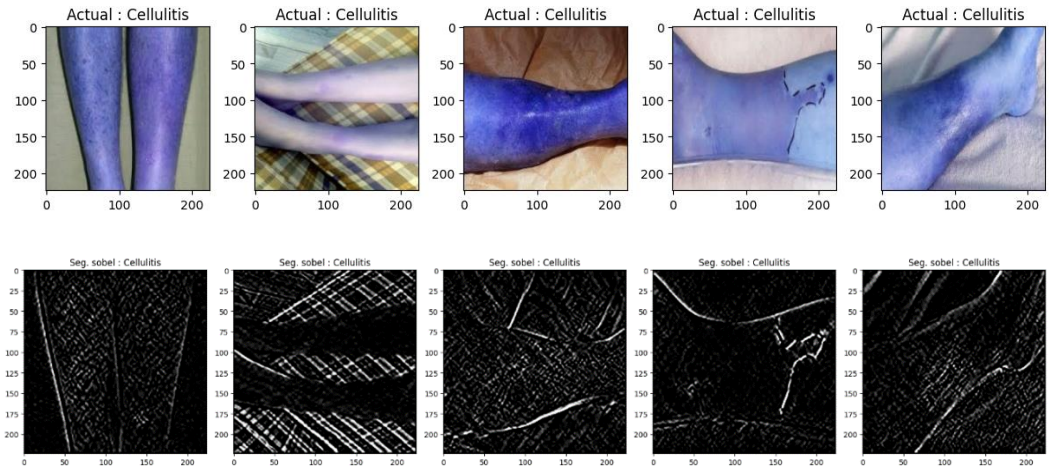

Figure 3: Class Cellulitis

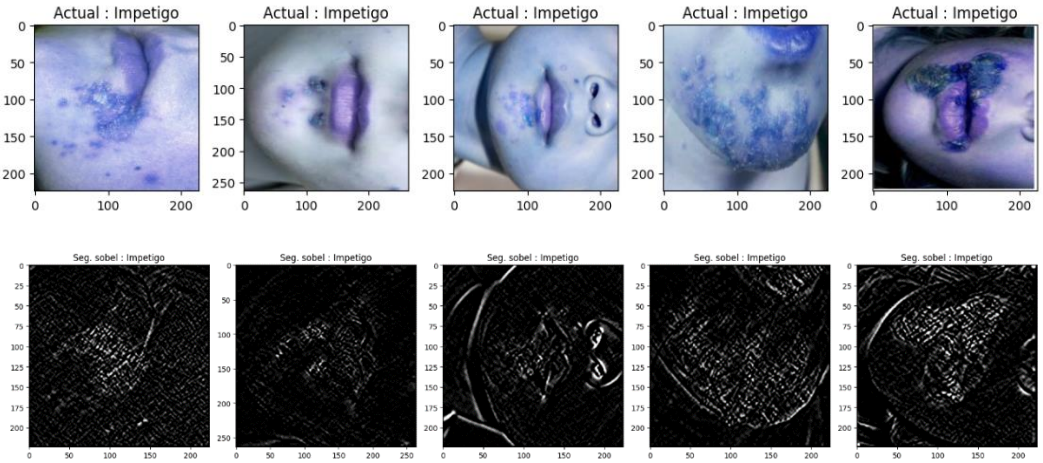

Figure 4: Class Impetigo

After segmentation, feature extraction is performed using Hu Moments. Hu Moments are a set of seven invariant moments derived from image moments, which are used to describe the shape of an object. They are invariant to image transformations such as translation, scale, and rotation. The Hu Moments are given by the following formulas [3], [14]:

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

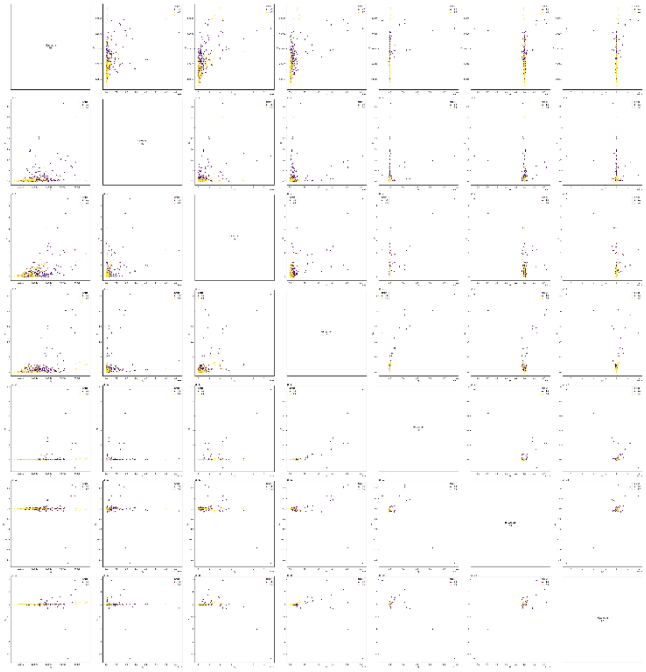

Figure 5: Scatter plots of Hu Moments

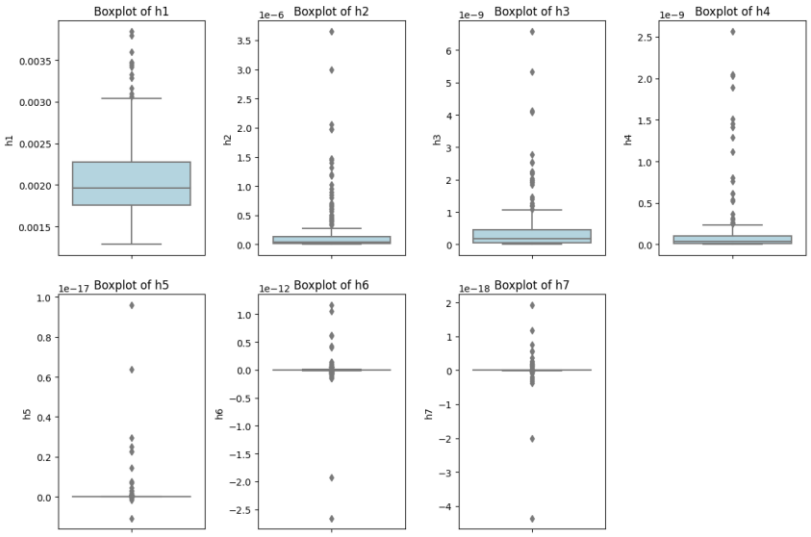

Figure 6: Boxplot of Hu Moments

Figure 5 show the distribution of each Hu Moment feature, highlighting the differences between the classes. **Figure 6** provide insights into how different moments correlate with each other and their separability between the classes. The segmented images and extracted features are then scaled to have a mean of 0 and a variance of 1 to ensure that all features contribute equally to the classification process.

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

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

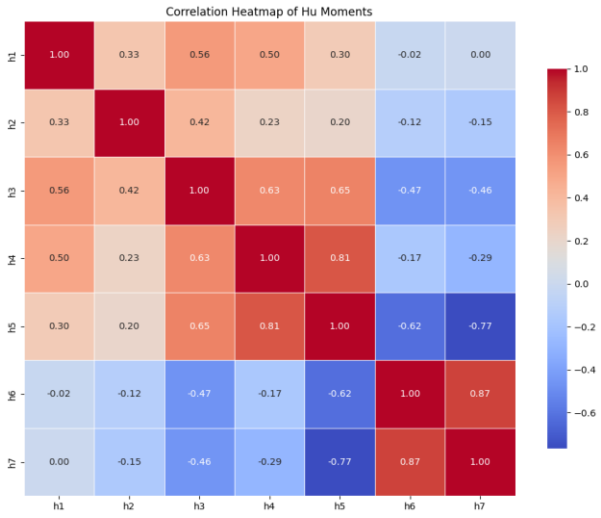

Figure 7: Correlation Heatmap of Hu Moments

Data Analysis Methods

The data analysis involves several key steps:

1. **Segmentation and Feature Extraction**: As described, images are segmented using the Sobel method [15], [16] and features are extracted using Hu Moments [7], [17].

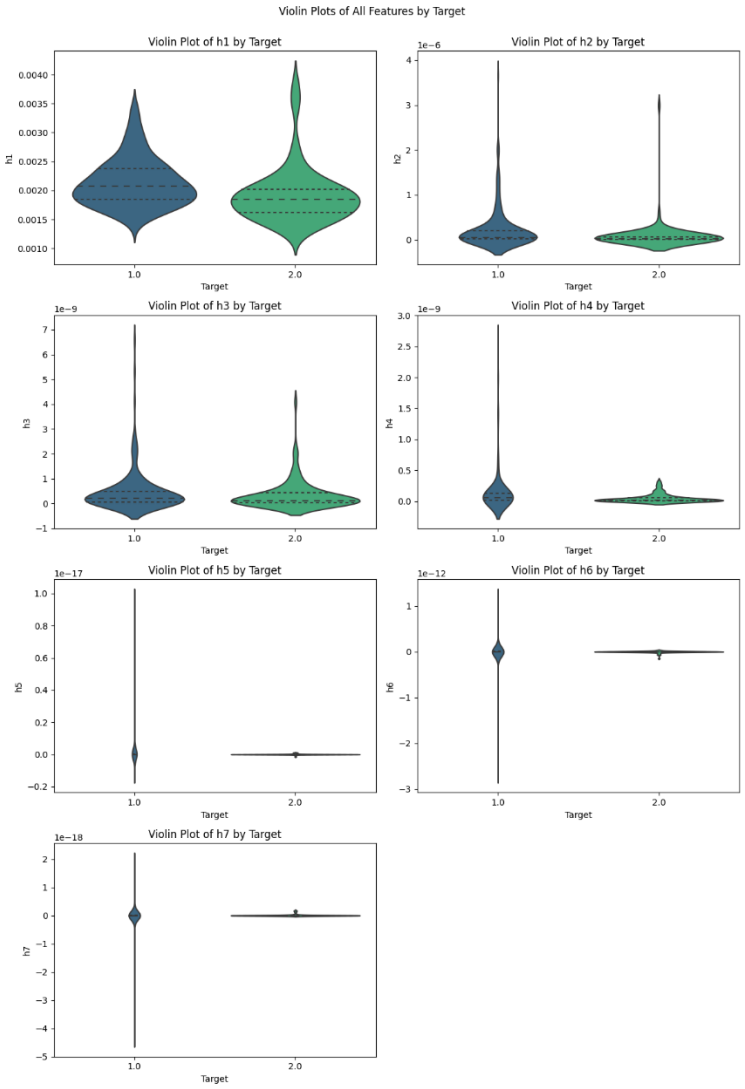

Figure 8: Violin plot of All Features by Target

2. **Scaling**: Features are scaled to have a mean of 0 and a variance of 1 using standard scaling techniques:

$$X' = \frac{X - \mu}{\sigma} \tag{4}$$

3. **Classification:** The K-NN algorithm is applied to classify the images into cellulitis or impetigo. The K-NN algorithm assigns a class to a new data point based on the majority class among its k-nearest neighbors in the feature space [4], [18], [19]. The distance metric used is typically the Euclidean distance, given by:

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{5}$$

Where x and y are two data points in n-dimensional space.

4. **Performance Evaluation:** The performance of the classification model is evaluated using various metrics [5], [20]:

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

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{6}$$

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

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

3. Result and Discussion

The data processing results reveal significant insights into the classification of bacterial skin infections using the K-Nearest Neighbors (K-NN) algorithm. The segmentation process using the Sobel method successfully highlighted the edges of the skin infection images, facilitating effective feature extraction through Hu Moments. After scaling the features to have zero mean and unit variance, the dataset was split into training and testing sets, with 80% used for training and 20% for testing.

The performance of the K-NN algorithm, with $k = 7$, was evaluated using five key metrics: accuracy, precision, recall, F1-score, and ROC-AUC. The following **Table 1** summarizes the results of these metrics across five different folds of cross-validation.

Table 2: Performance Metrics Across 5-Fold Cross-Validation for the K-NN Algorithm

K-n	Metrics				
	Accuracy	***Precision***	***Recall***	***F-Measure***	***ROC-AUC***
K-1	77.78	80.56	77.78	75.5	76.03
K-2	64.81	63.85	64.81	64.13	59.71
K-3	57.41	55.18	57.41	55.82	58.53
K-4	62.96	63.77	62.96	63.28	64.49
K-5	64.81	62.55	64.81	60.59	65.88
$\sum Avg$	65.95	65.18	65.95	63.06	64.13

The above table shows the variation in performance across different folds, with the mean values providing an overall assessment of the model's effectiveness. To further illustrate the performance, the following visualization includes the performance metrics graph and the confusion matrix:

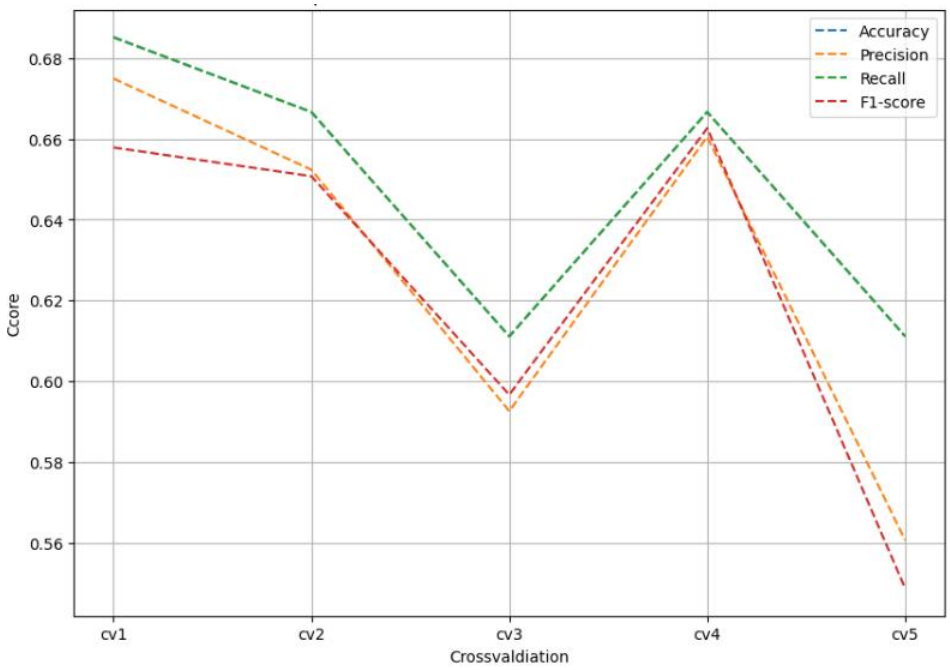

Figure 9: Performance Comparison of Each Cross-validation Fold

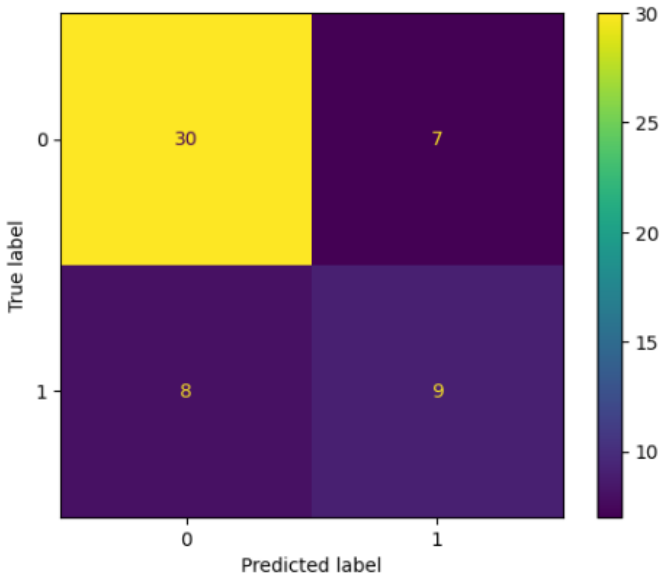

Figure 10: Confusion Matrix of the K-NN

The interpretation of these results indicates that the K-NN model achieved a mean accuracy of approximately 65.95%, with the precision, recall, F1-score, and ROC-AUC metrics also showing consistent performance. The confusion matrix further supports these findings, demonstrating the model's ability to distinguish between cellulitis and impetigo with reasonable accuracy.

Significant findings from this analysis include the relatively balanced performance across the metrics, suggesting that the chosen pre-processing methods and classification algorithm were effective. However, the variability in performance across different folds indicates that there may be room for improvement, possibly through the use of more advanced algorithms or additional data pre-processing techniques.

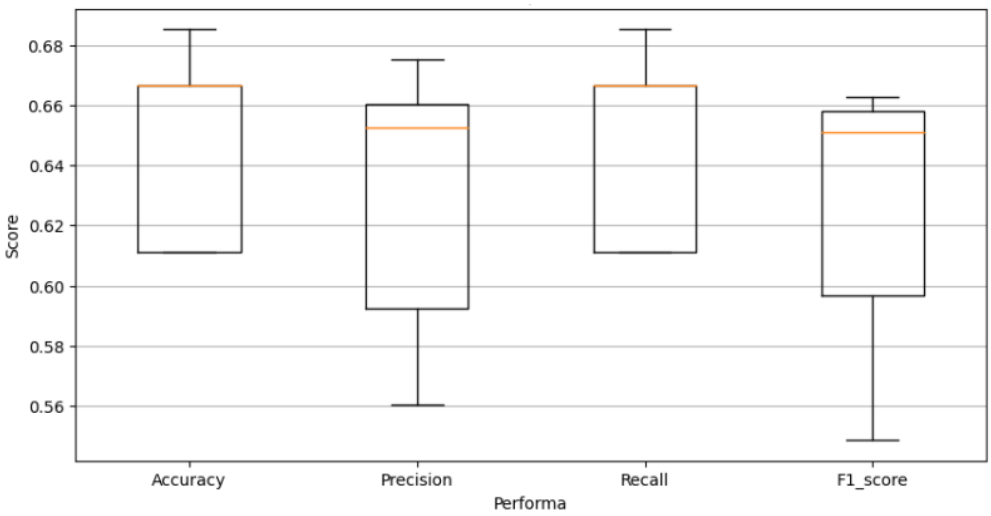

Figure 11: Performance Boxplot of the K-NN

Discussion

The results of this study provide valuable insights into the use of image processing and machine learning for the classification of bacterial skin infections. The K-NN algorithm, combined with Sobel segmentation and Hu Moments, demonstrated a reasonable ability to classify cellulitis and impetigo. The mean performance metrics indicate that the approach is feasible, although the variability across folds suggests potential areas for further refinement. In interpreting these results, it is essential to consider the relationship between the current findings and previous research. The use of Hu Moments for feature extraction and K-NN for classification aligns with established practices in image processing and machine learning. However, this study contributes by applying these techniques specifically to the domain of bacterial skin infections, an area that has not been extensively explored in the literature.

The practical implications of these findings are significant. An automated classification system for bacterial skin infections can greatly enhance the speed and accuracy of diagnosis, providing valuable support to healthcare professionals. This can lead to earlier and more accurate treatments, ultimately improving patient outcomes. However, this study also has its limitations. The dataset is imbalanced, which may affect the generalizability of the results. Additionally, while the K-NN algorithm performed reasonably well, more sophisticated machine learning models, such as convolutional neural networks (CNNs), could potentially offer improved performance. Future research should explore these avenues, as well as the integration of additional data pre-processing and augmentation techniques to enhance model robustness.

In conclusion, while the current study provides a solid foundation for automated classification of bacterial skin infections, further research is needed to address the limitations and explore additional methods to improve performance. The findings underscore the potential of machine learning in enhancing diagnostic accuracy and efficiency in dermatology. By visualizing and analysing the data through various plots and evaluation metrics, we gain deeper insights into the dataset and the effectiveness of the applied methods. This holistic approach ensures that the classification system developed is both accurate and practical for real-world applications.

4. Conclusion

In summary, this research demonstrates the feasibility of using the K-Nearest Neighbors (K-NN) algorithm, combined with Sobel segmentation and Hu Moments, for the classification of bacterial skin infections, specifically cellulitis and impetigo. The results showed a mean accuracy of approximately 65.95%, with precision, recall, F1-score, and ROC-AUC metrics indicating consistent performance. The confusion matrix supported these findings, highlighting the model's capability to differentiate between the two types of infections. These results answer the research question affirmatively, suggesting that the integrated approach of Sobel segmentation, Hu Moments feature extraction, and K-NN classification can effectively classify bacterial skin infections.

This study contributes to the field by providing a novel application of image processing and machine learning to dermatology, showcasing the potential of automated systems in enhancing diagnostic accuracy and efficiency. However, the variability in performance across different folds and the imbalanced dataset indicate areas for further improvement. Future research should explore more advanced machine learning models, such as convolutional neural networks (CNNs), and consider additional data pre-processing and augmentation techniques. These efforts will enhance the robustness and generalizability of the model, ultimately leading to more reliable and practical diagnostic tools in clinical settings.

References:

[1] K. V Swamy, “Skin Disease Classification using Image Preprocessing and Machine Learning,” *2024 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation, IATMSI 2024*. 2024, doi: 10.1109/IATMSI60426.2024.10502445.

[2] 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.

[3] 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.

[4] A. Sharma, “Prediction of the Fracture Toughness of Silicafilled Epoxy Composites using K-Nearest Neighbor (KNN) Method,” *2020 International Conference on Computational Performance Evaluation, ComPE 2020*. pp. 194–198, 2020, doi: 10.1109/ComPE49325.2020.9200093.

[5] M. Khushi, “A Comparative Performance Analysis of Data Resampling Methods on Imbalance Medical Data,” *IEEE Access*, vol. 9, pp. 109960–109975, 2021, doi: 10.1109/ACCESS.2021.3102399.

[6] 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.

[7] 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.

[8] S. T. Ahmed, "Enhancement of student performance prediction using modified K-nearest neighbor," *Telkomnika (Telecommunication Comput. Electron. Control.*, vol. 18, no. 4, pp. 1777–1783, 2020, doi: 10.12928/TELKOMNIKA.V18I4.13849.

[9] Y. Boer, "Classification of Heart Disease: Comparative Analysis using KNN, Random Forest, Gaussian Naive Bayes, XGBoost, SVM, Decision Tree, and Logistic Regression," *2023 5th International Conference on Cybernetics and Intelligent Systems, ICORIS 2023*. 2023, doi: 10.1109/ICORIS60118.2023.10352195.

[10] 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.

[11] 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.

[12] T. Wu, "Image Edge Detection Based on Sobel with Morphology," *IEEE Inf. Technol. Networking, Electron. Autom. Control Conf. ITNEC 2021*, pp. 1216–1220, 2021, doi: 10.1109/ITNEC52019.2021.9586895.

[13] J. N. Archana, "Enhancement of digital chest images using a modified Sobel edge detection algorithm," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 24, no. 3, pp. 1718–1726, 2021, doi: 10.11591/ijeecs.v24.i3.pp1718-1726.

[14] 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.

[15] J. Sun, "Class-imbalanced dynamic financial distress prediction based on Adaboost-SVM ensemble combined with SMOTE and time weighting," *Inf. Fusion*, vol. 54, pp. 128–144, 2020, doi: 10.1016/j.inffus.2019.07.006.

[16] A. Çınar, "Classification of normal sinus rhythm, abnormal arrhythmia and congestive heart failure ECG signals using LSTM and hybrid CNN-SVM deep neural networks," *Comput. Methods Biomech. Biomed. Engin.*, vol. 24, no. 2, pp. 203–214, 2021, doi: 10.1080/10255842.2020.1821192.

[17] 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.

[18] A. E. Minarno, "Classification of batik patterns using K-nearest neighbor and support vector machine," *Bull. Electr. Eng. Informatics*, vol. 9, no. 3, pp. 1260–1267, 2020, doi: 10.11591/eei.v9i3.1971.

[19] S. P. Thangavel, "K-nearest neighbour technique for the effective prediction of refrigeration parameter compatible for automobile," *Therm. Sci.*, vol. 24, no. 1, pp. 565–569, 2020, doi: 10.2298/tsci190623436p.

[20] P. Sharma, "Performance analysis of deep learning CNN models for disease detection in plants using image segmentation," *Inf. Process. Agric.*, vol. 7, no. 4, pp. 566–574, 2020, doi: 10.1016/j.inpa.2019.11.001.