Automatic Area Estimation of Mice Wound Images

Bruno Uhlmann Marcato[1], Camila Rodrigues Ferraz[2], Waldiceu Aparecido Verri Jr[3], Rubia Casagrande[3], Daniel Prado Campos[1], José Luis Seixas Junior[4], Rafael Gomes Mantovani[1]

Federal University of Technology – Paraná (UTFPR), Campus of Apucarana, Paraná, Brazil
University of Maryland Baltimore, School of medicine, Baltimore, Maryland, USA
State University of Londrina (UEL), Londrina, Paraná, Brazil
ELTE – Eötvös Loránd University, Faculty of Informatics, Budapest, Hungary

Abstract. Image segmentation is a classic computer vision set of techniques that partitions a digital image into discrete groups of pixel-image segments to inform object detection and related tasks. It has been successfully explored in biological studies, such as in the identification of wounds. However, recent approaches towards using black-box deep learning algorithms for image and semantic segmentation of objects have higher computational costs than classic techniques. In this study, we evaluated the effectiveness of thresholding and deep learning techniques for semantic segmentation of wound images of mice. Experiments were performed with a real dataset developed by the Pain, Neuropathy, and Inflammation Laboratory at the State University of Londrina with the approval of the University Ethics Committee on Animal Research and Welfare. The results were promising, showing that deep learning and thresholding were able to recognize wound areas, with an average IoU of 0.75 and 0.72, respectively. However, when estimating the wound areas, deep learning results were the most close to the ground truth.

CCS Concepts: • **Computing methodologies** → **Machine learning algorithms**.

Keywords: area estimation, image segmentation, thresholding, wound identification

1. INTRODUCTION

Wound assessment is crucial in medical research, particularly in understanding wound healing processes and evaluating treatment outcomes. Recent advances in Computer Vision (CV) and Machine Learning (ML) have opened up new possibilities for automated wound analysis [Zhang et al. 2022]. However, choosing the proper technique remains a challenge. In CV, traditional techniques for image segmentation have long been employed to delineate objects and regions of interest within images. These techniques encompass a spectrum of methods, including edge detection, region growing, and clustering algorithms such as K-means and Gaussian mixture models. Among these, thresholding is a fundamental approach, where pixel values are partitioned based on a predefined threshold to differentiate between foreground and background elements.

Thresholding methods involve setting intensity thresholds to segment wound regions based on color model values for each pixel. These techniques are the most common segmentation approach due to ease of use, simplicity, and fast computation. Still, they may struggle with complex textures or when the lighting conditions vary across the image [Alsahafi et al. 2023]. On the other hand, Deep Learning (DL) models, such as Convolutional Neural Networks (CNNs), have shown remarkable success in various image segmentation tasks [Long et al. 2015]. They can learn complex features from data but require substantial computational resources, time and labeled training data [Manakitsa et al. 2024].

Thus, in this paper we investigate the hypothesis that traditional thresholding methods can be as accurate as DL models in segmenting wounds. Experiments were performed with two main approaches:

thresholding techniques and state-of-the-art DL models for semantic segmentation. This paper is organized as follows: Sections 2 and 3 present the theoretical concepts necessary for understanding the work and related works found in the literature; Section 4 describes the proposed methodology for comparing and evaluating the segmentation techniques; Section 5 presents the experimental results; and finally, Section 6 presents conclusions and suggestions for future work.

2. BACKGROUND

Semantic segmentation is an active area of research in CV that involves identifying and classifying individual pixels within an image. Its applications span diverse fields, from autonomous driving to medical image analysis [Csurka et al. 2023]. The conventional methods for this task demand substantial amounts of meticulously annotated data, which is time-consuming and expensive or prone to noise generation. Most semantic segmentation networks commonly utilize cross-entropy as their loss function and assess network performance using the intersection-over-union (IoU) metric [Huang et al. 2020].

On the other hand, image segmentation [Minaee et al. 2020] seeks to separate an image into distinct regions based on characteristics such as color, texture, or intensity. Unlike semantic segmentation, which categorizes pixels into specific classes, non-semantic segmentation groups similar pixels or regions but does not assign semantic meaning. Image processing is used in traditional methods, while ML algorithms and CNNs are used in modern approaches.

2.1 Thresholding

Thresholding is a simple and crucial tool for dividing an image into regions of interest. Based on their intensity values, these techniques define a threshold separating image pixels into two distinct categories.Values below the found threshold are assigned to one class, while values above the threshold are assigned to the other class. Among the methods applied to determine the ideal threshold, we can list Otsu, Isodata, Mean, Li, and Yen [Van der Walt et al. 2014], each using different approaches to calculate threshold values.

2.2 Machine Learning

Machine Learning (ML) is a core field of Artificial Intelligence (AI) focused on automating tasks through data-driven training. It involves algorithms learning from past experiences to improve future performance. ML can be supervised, unsupervised, or reinforcement learning, with supervised learning being the most common [Marsland 2015]. In supervised learning, algorithms are trained with correct input-output pairs provided by specialists, enabling them to generate correct outputs for new inputs after training.

Different learning algorithms can be explored when dealing with classification tasks. While many algorithms exist, some demonstrate consistent performance across a wide range of problems and tasks. One such algorithm is Random Forest (RF) [Breiman 2001], which often excel in predictive tasks because they can mitigate overfitting and handle high-dimensional data effectively. Therefore, despite the availability of numerous classifiers, the robustness and versatility of algorithms like Random Forest make them a compelling choice for various machine learning applications.

Unlike traditional ML algorithms, DL [Aggarwal 2018] has been widely used to solve image classification and segmentation problems. It extracts high-level abstract features, but its use depends on the data amount used to train models, i.e., it is ideal to have a massive dataset with hundreds of thousands or millions of samples. In this study, we mention the U-Net algorithm as a CNN architecture developed for image segmentation, which is efficient in several applications in the medical context [Punn and Agarwal 2022; Azad et al. 2022]. The network design consists of a symmetric

network with descending *pooling* layers to encode low-level information into a high-dimensional representation, ascending convolution (or transposed convolution), and up-sampling layers to reconstruct the segmented image.

3. RELATED WORKS

3.1 Thresholding for Image Segmentation

Many studies review thresholding algorithms, providing concise evaluations of their efficacy. In [Pare et al. 2020], the authors presented a comprehensive review of optimization algorithms applied to multilevel thresholding in image segmentation. A survey of 157 relevant research publications is conducted, encompassing parametric and non-parametric approaches. The study categorizes image thresholding methods to address diverse segmentation challenges. Optimization algorithms are increasingly utilized across numerous domains, including image-processing tasks like enhancement, compression, classification, and pattern recognition. Swarm intelligence algorithms are notably gaining traction for multilevel thresholding in gray-scale and colored natural and satellite images. However, the effectiveness of specific algorithms varies depending on image types, requiring tailored approaches for different image classes.

In a different study [Hosny et al. 2023], the authors introduced a modified optimization algorithm for image segmentation, employing a hybrid approach of Otsu's and Kapur's entropy methods to determine optimal thresholds. The proposed algorithm outperformed other techniques in image segmentation performance. A specific enhancement for satellite image segmentation is proposed, utilizing chaotic initialization and a hybrid fitness function, demonstrating superior results. Furthermore, [N and S 2016] introduced a locally adaptive thresholding method for image binarization, employing local mean and standard deviation to distinguish foreground and background pixels. It compares the performance of Niblack and Sauvola local thresholding algorithms, mainly focusing on medical image applications. Evaluation metrics include the Jaccard Similarity Coefficient and Peak Signal Noise Ratio (PSNR). Results demonstrated that the Niblack algorithm outperformed Sauvola in reducing background noise, as indicated by segmentation quality metrics.

3.2 Deep Learning for Semantic Segmentation

Recently, most of the research in the field has turned to DL algorithms for solving the semantic segmentation of wounds. For instance, in [Wang et al. 2020], semantic segmentation of ulcers' images was conducted using cutting-edge models like U-Net, FCN-VGG16, Mask-RCNN, and MobileNetV2. U-Net demonstrated the highest accuracy, achieving an FScore of 0.915 and the highest recall of 0.912. In parallel, [Kaymak et al. 2020] explored Fully Connected Networks (FCN) for segmenting skin lesions, comparing their performance with other networks like U-Net and SegNet.

Another study [Niri et al. 2020] focused on segmenting diabetic foot ulcers using DL models such as U-Net, V-Net, and SegNet. Based on accuracy, IoU, and FScore metrics, the evaluation favored U-Net across all three criteria, achieving scores of 0.949, 0.948, and 0.972, respectively. Furthermore, [Kang and Nguyen 2019] proposed a hybrid framework that combines RF with DL for image segmentation. Their approach, tested on hand segmentation and other semantic segmentation datasets, achieved real-time segmentation with limited computational resources, showcasing its efficacy in various applications.

4. METHODOLOGY

An overview of the flow of experiments, including sub-steps, is shown in Figure 1. The following subsections give additional details regarding them: the image dataset, data preparation, classification algorithms used, models' training and evaluations, and the area estimation.

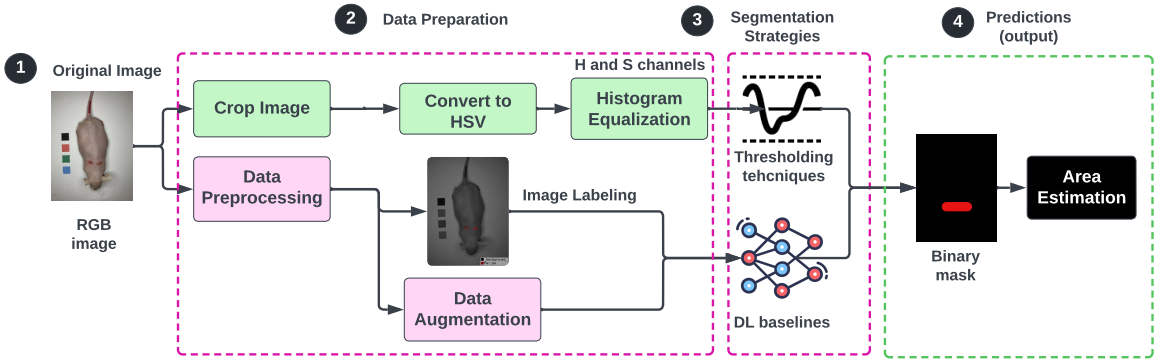


Fig. 1: Experimental methodology for automated area estimation of mice wounds. Adapted from [Marcato et al. 2024].

4.1 Image Dataset

The dataset used in this study comprises 71 images of mice with wounds on their backs. It was developed by the Pain, Neuropathy, and Inflammation Laboratory at the State University of Londrina (UEL) with the approval of the UEL Ethics Committee on Animal Research and Welfare (process number 15654.2019.33). Images were acquired with no protocol, so there are files with different resolutions (36 images have 1024×768 pixels, while the remaining have 4032×3024 pixels). The classes (pixel labels) were defined manually using the `Labelme` tool[1].

4.2 Data Preprocessing

When applying thresholding methods, each image segmented the region of interest to the wound to avoid unnecessary information in the image histogram. More specifically, the images, with a resolution of 256×256 pixels, are cropped in the regions [80, 210] on the vertical axis and [80, 200] on the horizontal axis, where all wounds are found for all images.

Initially, the images are converted to the HSV color space, and equalization is applied to the H and S channels. Histogram equalization is helpful in fulfilling pixel gradation level and adding color contrast between background and foreground. The reason for using HSV instead of RGB is that it separates color information (chroma) from intensity or lighting (luma). Because the values are separated, it is possible to construct a histogram or thresholding rule using only saturation and hue.

Considering the baselines, images were resized to 256×256 pixels and normalized from $[0, 255]$ to interval $[0, 1]$. Data Augmentation (DA) was applied to increase the dataset size (71 images). Five transformations were applied: horizontal and vertical flips, rotations (up to 35 degrees), salt and pepper noise, and image translation [Marcato et al. 2024]. DA expanded the dataset to 426 images for training, ensuring model precision and preventing overfitting.

4.3 Segmentation Techniques

Considering thresholding methods, six strategies were explored:

—**Otsu**: is a thresholding technique that calculates the ideal threshold to segment an image, finding values that minimize the variance within classes while maximizing the variance between classes of pixels;

[1] `https://github.com/wkentaro/labelme`

—**Isodata**: is an iterative algorithm that calculates the threshold based on the average of the pixel values. It divides the image pixels into two classes, updating the threshold until the difference between successive thresholds is less than a predefined value;

—**Mean**: is a simple method that calculates the threshold as the average of the intensity values of the pixels in the image;

—**Li**: is a method based on image histogram analysis and minimization of entropy between classes. It calculates the optimal threshold that maximizes the entropy between the two classes resulting from the segmentation;

—**Yen**: is a technique based on histogram analysis that determines the optimal threshold considering the variability between classes and the entropy of the image and

—**Voting**: *Voting* application of all the methods mentioned above.

We evaluated thresholding techniques with and without Binary Closing, a morphological operation that smooths object boundaries and improves image connectivity. Binary Closing involves dilation followed by erosion, reducing noise, and closing small gaps in object outlines. By comparing model performance with and without Binary Closing, we aimed to assess the efficacy of this operation in improving segmentation accuracy.

We also explored two different DL baselines for wound recognition, previously evaluated in [Marcato et al. 2024]: i) U-Net: a state-of-the-art DL architecture for Semantic Segmentation; and ii) RF + VGG16: a cheap alternative combining VGG16 latent features trained in a Random Forest model. U-Net was trained with Adam optimizer and a $\alpha = 0.0001$ learning rate. The optimized loss function was binary cross entropy. The activation functions between convolutional layers are ReLu functions and a sigmoid in the last fully connected layer. U-Net was trained for 100 epochs using batch size = 2. The RF setup was defined with 100 trees and the '*Gini*' index as the attribute evaluation criterion. It was fed with 64 feature maps extracted from the first two convolutional layers of the VGG16[2] neural network [Simonyan and Zisserman 2015], pre-trained in the `ImageNet` dataset.

4.4 Area Estimation

We defined a green square marker of a fixed size of $1cm \times 1cm = 1cm^2$ to estimate the wound areas. These squares are annotated in all the original images. Thus, by counting the number of pixels in the marker, we establish a pixel-to-area ratio that translates pixel measurements into real-world dimensions. The wound areas from binary mask images are considered the ground truth. This ratio is then applied to the thresholding/semantic segmentation outputs using their obtained IoU scores. The IoU scores are multiplied by the ground truth area, resulting in the estimated wound area in the output. This approach leverages the IoU score as a scaling factor, adjusting the ground truth area based on the segmentation accuracy.

4.5 Experimental Setup

We evaluated DL models using a cross-validation resampling with five folds. Moreover, to prevent any data leakage, we performed DA only on the training set after separating training and testing sets in each iteration. All the strategies (Thresholding, UNet, and RF) were evaluated using the Intersection Over Union (IoU) metric. This measure divides the intersection of two masks by their union, obtaining the perfect prediction when both are equal. IoU is one of the most used metrics in image and semantic segmentation [Goyzueta et al. 2021].

To ascertain the statistical significance of our findings, we evaluated the results using non-parametric Wilcoxon with a significance level $\alpha = 0.05$. This method enabled us to rigorously compare the

[2]Visual Geometry Group, 16 layers.

performance of our threshold techniques against the baselines. By employing statistical tests, we can assess the significance of performance differences by considering the different evaluated techniques. All the code was developed in Python: Threshold techniques used `scikit-image` library; DL algorithm used `PyTorch`; RF and the ML methodological functions used `scikit-learn` and `Keras`. Finally, DA methods used `albumentations`. The code repository of this study is also publicly available[3].

5. RESULTS

5.1 Predicting Wound Pixels

Table I shows the general results of the Thresholding techniques applied to all the available images. In the table, the mean IoU and standard deviation values are reported. The best results are reported in bold. We preferred the method with the lowest sd when two techniques obtained the same mean value. The last two rows of the table present DL baselines. Considering results without Binary Closing, the Mean technique presented the best result among thresholding methods, with values close to the RF+VGG baseline, and showed a difference of 10% about the UNet baseline. On the other hand, when using Binary Closing, the Isodata and Otsu techniques were the best strategies, outperforming the RF+VGG baseline, with a slight difference between UNet.

Table I: Mean IoU and standard deviation results obtained by experiments

Technique		**Mean IoU (sd)**	
		Without Closing	**With Closing**
Thresholding	Isodata	0.633 (0.212)	**0.717 (0.194)**
	Li	0.593 (0.198)	0.682 (0.190)
	Mean	**0.645 (0.217)**	0.682 (0.201)
	Otsu	0.631 (0.213)	0.717 (0.195)
	Yen	0.497 (0.234)	0.617 (0.242)
Voting	Thr. Voting	0.633 (0.210)	0.714 (0.196)
DL Baselines	U-Net	**0.752 (0.197)**	**0.752 (0.197)**
	RF + VGG	0.644 (0.241)	0.644 (0.241)

The non-parametric Wilcoxon test with $\alpha = 0.05$ (95% of significance) was applied to assess the statistical significance of these results. For Mean and UNet methods, a p-value < 0.001 was obtained, meaning a statistical difference in favor of UNet. Otherwise, for tests with Mean and RF, Isodata and UNet, and Isodata with UNet (p-values of 0.313, 0.124, and 0.057, respectively), the null hypothesis is considered, meaning no statistical difference between the distributions.

Figure 2 depicts individual image results for the best techniques, considering both scenarios: with and without Binary Closing. The x-axis shows all the individual images in the dataset, while the y-axis segmentation techniques: the darker the cell, the better the IoU values. There are some images whose predictions for all the methods are inaccurate. It happens with images 28, 29, and 60, presenting lower values of IoU. These images depict minor wounds in an advanced stage of healing, almost closed. At this point, the wounds may have minimally affected pixels, making segmentation challenging. Moreover, as healing progresses, the color of the wounds may blend with that of the surrounding tissue, further complicating segmentation. On the other hand, images 70 and 71 achieved accurate results with all methods. This is likely because these images depict wounds at an early stage and bigger size, with significant regions and clear differentiation from the mice's skin.

[3] `https://github.com/BrunoMarcato/MiceWoundSegmentation`

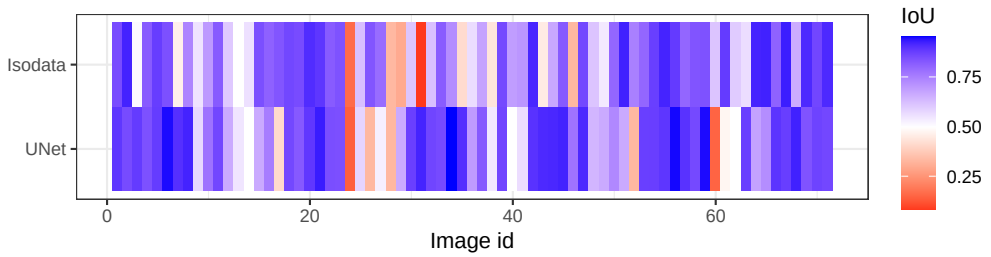
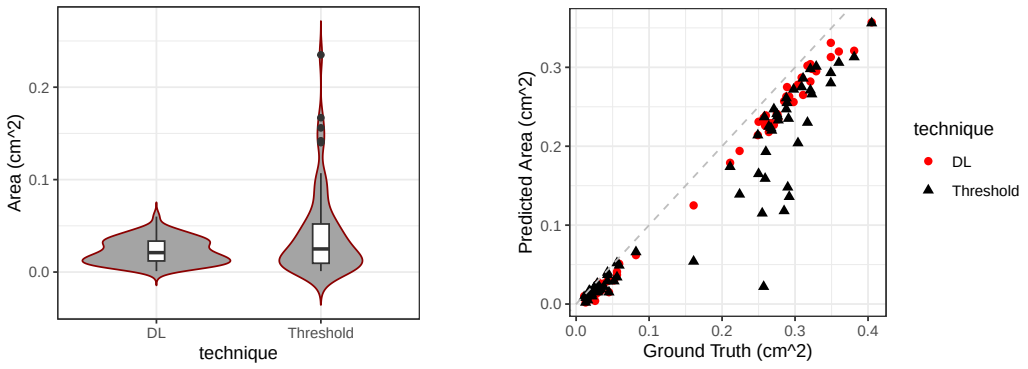


Fig. 2: IoU heatmap of best Thresholding technique and UNet

5.2 Estimating Wound Areas

The results of estimating wound areas using DL and thresholding techniques show notable differences, as shown in 3. Figure 3a depicts the distribution of area errors for each technique. DL obtained an average absolute error of 0.021, while thresholding obtained 0.025 when considering all the wound areas. Although error estimations of both techniques may seem close, DL predictions show a more significant concentration of values close to zero, indicating individual lower errors. The thresholding technique shows a broader distribution with some outliers.

Additionally, Figure 3b demonstrates the individual differences by plotting ground truth areas against predicted areas. The diagonal line represents perfect predictions. The closer to the diagonal line, the better the technique and the estimated area. Figure highlights that DL red points are better than thresholding points (black triangles). An inaccurate group of images is concentrated in the middle of the chart. These images contain animals with tiny wounds that are hard to predict. Nevertheless, DL provided better estimations than thresholding. Thus, it is safe to state that DL generally outperforms the thresholding technique.



(a) Models' errors

(b) Images' errors

Fig. 3: Area errors obtained by the induced models.

6. CONCLUSION

In this study, we investigated Thresholding techniques for mice wound image segmentation. Experiments were carried out with an image dataset generated by the Pain, Neuropathy, and Inflammation Laboratory at the State University of Londrina, composed of 71 images showing wounds in mice. The Isodata, Otsu, Mean, Li, and Yen methods were applied considering the application of Binary Closing method to fill regions not wholly identified.

Threshold methods obtained promising results in terms of IoU. The best thresholding technique was the Isodata with Binary Closing, with an average IoU of 0.72 compared to 0.75 for UNet. When applying the non-parametric Wilcoxon test, there was no statistical difference between the technique's performances, which favors thresholding techniques that require a few seconds to be applied in contrast to the hours required to induce DL models and does not require training set.

Both strategies provided accurate results when estimating the wound areas, with DL being more consistent even when estimating tiny wounds in the most challenging images. For future works, we can explore different DL architectures and traditional ML while automating the entire pipeline, providing a more robust comparison.

REFERENCES

Aggarwal, C. C. *Neural Networks and Deep Learning: A Textbook*. Springer International Publishing, Cham, 2018.

Alsahafi, Y. S., Elshora, D. S., Mohamed, E. R., and Hosny, K. M. Multilevel threshold segmentation of skin lesions in color images using coronavirus optimization algorithm. *Diagnostics* 13 (18), 2023.

Azad, R., Aghdam, E. K., Rauland, A., Jia, Y., Avval, A. H., Bozorgpour, A., Karimijafarbigloo, S., Cohen, J. P., Adeli, E., and Merhof, D. Medical image segmentation review: The success of u-net, 2022.

Breiman, L. Random forests. *Machine learning* 45 (1): 5–32, 2001.

Csurka, G., Volpi, R., and Chidlovskii, B. Semantic image segmentation: Two decades of research, 2023.

Goyzueta, C. A. R., De la Cruz, J. E. C., and Machaca, W. A. M. Integration of u-net, resu-net and deeplab architectures with intersection over union metric for cells nuclei image segmentation. In *2021 IEEE Engineering International Research Conference (EIRCON)*. pp. 1–4, 2021.

Hosny, K. M., Khalid, A. M., Hamza, H. M., and Mirjalili, S. Multilevel thresholding satellite image segmentation using chaotic coronavirus optimization algorithm with hybrid fitness function. *Neural Computing and Applications* 35 (1): 855–886, Jan, 2023.

Huang, Y., Tang, Z., Chen, D., Su, K., and Chen, C. Batching soft iou for training semantic segmentation networks. *IEEE Signal Processing Letters* vol. 27, pp. 66–70, 2020.

Kang, B. and Nguyen, T. Q. Random forest with learned representations for semantic segmentation. *IEEE Transactions on Image Processing* 28 (7), 2019.

Kaymak, R., Kaymak, C., and Ucar, A. Skin lesion segmentation using fully convolutional networks: A comparative experimental study. *Expert Systems with Applications* vol. 161, pp. 113742, 2020.

Long, J., Shelhamer, E., and Darrell, T. Fully convolutional networks for semantic segmentation, 2015.

Manakitsa, N., Maraslidis, G. S., Moysis, L., and Fragulis, G. F. A review of machine learning and deep learning for object detection, semantic segmentation, and human action recognition in machine and robotic vision. *Technologies* 12 (2), 2024.

Marcato, B. U., Pierotti, S. M., Ritter, P. D., Ferraz, C. R., Verri Jr, W. A., Casagrande, R., Seixas Junior, J. L., and Mantovani, R. G. Semantic segmentation of mice wounds. In *Anais do XV Computer on the Beach, 10 a 13 de abril de 2024*. pp. 23–29, 2024.

Marsland, S. *Machine learning: an algorithmic perspective*. CRC press, 2015.

Minaee, S., Boykov, Y., Porikli, F., Plaza, A., Kehtarnavaz, N., and Terzopoulos, D. Image segmentation using deep learning: A survey, 2020.

N, S. and S, V. Image segmentation by using thresholding techniques for medical images. *Computer Science & Engineering: An International Journal* vol. 6, pp. 1–13, 02, 2016.

Niri, R., Hassan, D., Yves, L., and Treuillet, S. Semantic segmentation of diabetic foot ulcer images: Dealing with small dataset in dl approaches, 2020.

Pare, S., Kumar, A., Singh, G. K., and Bajaj, V. Image segmentation using multilevel thresholding: A research review. *Iranian Journal of Science and Technology, Transactions of Electrical Engineering* 44 (1): 1–29, Mar, 2020.

Punn, N. S. and Agarwal, S. Modality specific u-net variants for biomedical image segmentation: a survey. *Artificial Intelligence Review* vol. 55, pp. 5845–5889, 2022.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*. CoRR, San Diego, CA, USA, 2015.

Van der Walt, S., Schönberger, J. L., Nunez-Iglesias, J., Boulogne, F., Warner, J. D., Yager, N., Gouillart, E., and Yu, T. scikit-image: image processing in python. *PeerJ* vol. 2, pp. e453, 2014.

Wang, C., Anisuzzaman, D., Williamson, V., Dhar, M. K., Rostami, B., Niezgoda, J., Gopalakrishnan, S., and Yu, Z. Fully automatic wound segmentation with deep convolutional neural networks, 2020.

Zhang, R., Tian, D., Xu, D., Qian, W., and Yao, Y. A survey of wound image analysis using deep learning: Classification, detection, and segmentation. *IEEE Access* vol. 10, pp. 79502–79515, 2022.