The Influence of Color on Semantic Ulcer Segmentation Using Deep Learning Models

Muhammad Usman(✉) and Jacek Rumiński

FETI, Department of Biomedical Engineering, Gdansk University of Technology, 80-233 Gdansk, Poland
muhammad.usman1@pg.edu.pl
http://www.pg.edu.pl

Abstract. Accurate and early image analysis is crucial for proper diagnosis and treatment. Convolutional neural networks (CNNs) have the ability to precisely classify and segment the affected wound areas, which has revolutionized biomedical imaging for automatic detection and diagnosis. However, CNN models often face challenges like generalization and overfitting issues due to limited wound image data. In addition, RGB images are computationally intensive. They may prompt the model to focus on irrelevant features, such as color variations, instead of the most important ones, like silhouettes, ultimately leading the model to the overfitting problem. This study investigates the influence of color on ulcer semantic segmentation. We explore different color-to-grayscale conversion techniques to study the learning behavior of the CNN model. Traditionally, image conversion from an RGB color space to grayscale uses fixed transformation parameters, e.g., YUV color model weights, which are standardized for human observers. The YUV color model separates luminance (Y) from chrominance (U and V), allowing efficient brightness and color information representation. However, machine-based processing does not require 'seeing' the content but focuses on extracting the essential features for model training, data augmentation, or future design of task-specific sensors. We designed a universal model architecture to analyze the learning procedure's ability to learn task-specific weights for color-to-grayscale conversion. The experiment results show how grayscale images based on learned weights could be used in task-specific semantic segmentation compared to color images and grayscale images obtained using traditional conversion techniques. The grayscale images based on learned weights are computationally efficient, and simplified feature representation helps the model emphasize the most relevant attributes while suppressing irrelevant ones. The source code is available on [1].

Keywords: Convolution neural network · Foot ulcer segmentation · Image Modalities · learned grayscale · RGB and gray-scale images

1 Introduction

The amount of money spent on wound care is increasing in the healthcare sector. Approximately 8 million people were injured in 2018, and Medicare costs (USA)

P. Ladyzynski et al. (Eds.): NBC/PCBBE 2025, IFMBE Proceedings 131, pp. 147–164, 2025.
https://doi.org/10.1007/978-3-031-96538-8_13

were estimated to be between $28.1 billion and $96.8 billion [2]. Wounds that do not heal properly or where the healing process does not restore anatomical and functional integrity after three months are considered chronic wounds. Time is of the essence; the longer a wound is allowed to deteriorate, the more difficult it is to heal, and early diagnosis is the most effective means of reducing wound care costs [3,4].

Chronic wounds are more likely to form in persons who are obese or have diabetes. Diabetes affects 34.2 million Americans and 463 million people globally, and by 2030, this number is expected to increase by 25% [5]. Diabetic foot ulcers (DFUs) are a relatively common form of chronic ulcers on the lower extremities. Regardless of the presence of peripheral vascular disease, neuropathy can diminish or eliminate the sensation of pain in the foot, thus often resulting in diabetic foot ulcers (DFUs), ranging in depth from shallow to deep. An increased mortality rate may result from these wounds or ulcers if they are not adequately managed [6,7].

For appropriate wound care and management, early and precise assessment of the severity of foot wounds can be crucial. Accurate tissue analysis and wound area assessment are essential for proper treatment and diagnosis. Wound attributes such as area, volume, and stage can be identified with the help of precise wound image segmentation [8]. In turn, these characteristics can be used to assess and treat chronic wounds, track the healing trajectory of the wound, plan future interventions, estimate the risk of hospitalization for the patient, or estimate the healing time [9], which can significantly lower hospitalization and amputation rates. It takes a great deal of effort and knowledge for professionals to manually mark the wound area and determine its severity without any inaccuracies. These factors are detrimental to the management of DFU because appropriate wound healing requires prompt and precise wound care decisions and treatments. Therefore, an automated approach is crucial for quick wound assessment, investigation, and treatment. Automatic segmentation of wound images can benefit significantly from computer-aided diagnosis and ailment location [10].

In binary wound segmentation, each pixel in the image is categorized as wound or non-wound. In the beginning, researchers proposed traditional wound segmentation methods involving learning algorithms. For example, methods [11–14] utilize watershed segmentation, threshold-based techniques, feature extraction methods, and similar approaches for semantic wound segmentation. Challenges faced by these methods include sensitivity to lighting conditions, skin color variations, resolution, effective feature engineering, and manual parameter optimization.

Over the years, with the unprecedented success of convolution neural networks (CNNs) in the medical imaging domain, researchers have also introduced deep learning-based segmentation frameworks that outperformed the traditional methods for wound segmentation. A CNN model for diabetic wound segmentation based on MobileNet-v2 was presented in [15]. The prediction masks from MobileNet-2 were converted into binary masks. Subsequently, to create the final segmentation mask, post-processing techniques for hole-filling and noise removal

were applied to the predicted mask. Liu et al. [16] introduced a dual-view segmentation technique that uses a learnable lighting correction module as a prepossessing step. This method employed characteristics from original and illumination-corrected images, enhancing the final result of wound segmentation. Cao et al. [17] proposed a hierarchical segmentation and multilevel classification paradigm to classify diabetic foot ulcers (DFUs) into five grades based on Wagner's wound assessment criteria.

The above methods face challenges in manual feature engineering or encounter overfitting and generalization problems due to various factors-for example, lack of available training data, skin color variations, and different lighting conditions. In semantic segmentation, especially in the ulcer wound image task, structural and textural information like wound silhouettes are very important. Moreover, the training data for wound images is also limited, and the model focuses on learning the color variations while neglecting the task-specific salient features. As a result, the model struggles with generalization and overfitting problems. Gray-scale images are a simplified representation that has the potential to help the model (e.g., using augmentation) focus on task-specific relevant information while suppressing irrelevant ones.

In [18], authors review 20 studies on measuring wound sizes using image processing techniques. Many of the studies used color-to-grayscale conversion as the first preprocessing step. All these approaches used traditional conversion methods based on fixed weights of transformations between RGB input data and luminance/intensity related color models such as HSV (hue, saturation value) or YUV/YCbCr (Y-luminance, UV or CbCr two chrominance components). These studies, and others presented in the next section, do not analyze to what extent color improves semantic segmentation results compared to single-component images, especially when the grayscale image is obtained not for fixed color-to-grayscale conversion weights but for learned domain-specific weights. It is also challenging to represent colors in digital images relatively constant over varying illuminations and different imaging sensors. Single-component images can potentially reduce the influence of color changes on the subsequent processing steps, and therefore, many image processing tasks use color-to-grayscale conversion. However, reducing the color components is a potential limitation of the information contained in an image. Hence, a need exists to analyze the influence of color and color reduction on wound semantic segmentation.

This study aims to investigate the effect of color on ulcer wounds in semantic segmentation. First, it is investigated to what extent domain-learned color conversion weights can improve the results of semantic wound segmentation compared to traditionally used color-conversion methods based on fixed, YUV color model weights. Second, we aim to analyze how color improves wound semantic segmentation results compared to grayscale images, primarily obtained using domain-learned color conversion weights. To reach these aims, first, we adopt different traditional RGB-grayscale conversion methods to convert the input RGB image to a single component representation. Then, we propose a learned grayscale technique that adaptively maps RBG color images to the learned

grayscale feature maps. We train the semantic segmentation models using various color compositions and compare the results.

The key contributions of our work are as follows:

a) We designed a simple, universal model architecture to learn the task-specific weights for color-to-grayscale conversion.
b) We demonstrated that using learned weight for color-to-grayscale conversion leads to much better results than traditional RGB-grayscale conversion methods.
c) We show that for the best model architecture, no statistical difference was found comparing segmentation metrics obtained for models trained on RGB images and grayscale images converted with learned weights.
d) We demonstrated that color preprocessing using the adjusted retinex method [19] can improve semantic segmentation results for the applied (not tuned) model for the similar but new domain.
e) We compared the influence of several color preprocessing steps on the semantic segmentation of ulcer images.

2 Related Work

This section reviews the previous works on wound image segmentation, including deep-learning and traditional machine-learning-based segmentation approaches.

The first step in diagnosing and treating of chronic wounds is to measure the wound area. Several traditional image-processing methods have been applied to classify wound tissue. For example, color descriptors and texture detectors have been used to automate the monitoring of the wound healing process, to extract information from wound images, and to classify skin patches as normal or abnormal [20,21] However, these experiments did not provide reliable tools for process automation Likewise, Song et al. [22] delineated 49 attributes employing a feature engineering approach based on traditional machine learning methods, including K-mean clustering, thresholding, region growth in grayscale and RGB, as well as edge detection. Subsequently, the resulting features were fed into a multilayer perceptron (MLP) and radial basis function (RBF) to learn and evaluate the segmentation results. Some other techniques include creating a red-yellow-black-white probability map [23], which is then optimized for thresholding or region growth [24]. However, feature engineering techniques require human involvement in feature selection and cannot adapt to image irregularities, leading to degraded predictive performance. These methods are subject to several constraints, including sensitivity to skin color, the involvement of a certain degree of feature engineering, the lack of full automation of the end-to-end process, and the manual parameter tuning involved. The authors in [25] review different methods to measure wound area, including elliptical estimation, square counting, photogrammetry using imaging devices, and 3D methods such as laser 3D scanners as well as stereophotography and 2-camera systems. Similarly, a study in [26] evaluates the accuracy of digital planimetry (DP) with adaptive

calibration for measuring wound area on curved surfaces. The DP with adaptive calibration exhibits significantly lower error (0.60% vs. 2.65% and 2.23%) and higher precision, eliminating systematic errors, compared to the SilhouetteMobile device and a standard DP. The method is 4.4 times more accurate and 7.4 times more precise, making it one of the best methods for measuring wound area on curved surfaces.

Recent advancements in deep learning have enabled automated feature extraction and data-driven learning. Researchers now leverage convolutional neural networks (CNNs) to identify wound-affected tissues and segment chronic wound areas effectively [15]. To mitigate the above limitation in chronic wound segmentation, Goyal et al. [27] exploit fully convolution networks (FCN) using different model configurations, including FCN-32 s, FCN-8 s, FCN-16 s, and AlexNet. These models were first pretrained in ImageNet [28] and Pascal VOC [29] dataset and then evaluated on the DFU dataset, consisting of 600 wound images. FCN-16 architecture performed adequately, achieving a 79.4% dice score. However, segmentation images with smaller wound areas and irregular borders remained challenging. U-Net being a famous architecture in semantic segmentation tasks, Niri et al. [30] employed it on the ESCALE dataset for diabetic food ulcer segmentation. In another approach, authors [31] introduced a preprocessing stage before the training of the CNN network to filter noise effects. Later, transfer learning techniques were introduced to elevate the generalization ability of the deep learning model and trained on the Mask-RCNN model to segment chronic wound images [32].

An ensemble network including LinkNet and U-Net employing pretrained EfficientNetB1 and EfficientNetB2 encoders, respectively, with additional pretraining using the Medetec dataset was proposed by Mahbod et al. [33] for the FUSeg challenge 2021 [33]. The segmentation performance was improved through fusion techniques, test time augmentation, and fivefold cross-validation. To constrain boundary information for the automated segmentation of foot wounds, Edge-OCRNet was first presented in [34] and used the ConvNeXt backbone architecture and edge loss function. The Edge-OCRNet segmented mask was then post-processed to enhance the predicted mask. Global and local attributes are equally crucial for accurate segmentation tasks. A WSNet with a global-local architecture was then introduced by Subba et al. [35], which uses the entire image and its patches to extract high-level semantic information and local context in wound images. The model is first trained to classify wounds, and then the segmentation data is used to refine the model further. WSNet achieves a dice score of 84.7% when tested on its own donated dataset, WOUNDSEG, which includes eight different types of wounds.

Color-to-grayscale conversion was used in many approaches focused on wound image analysis. All proposed methods use traditional color-to-grayscale conversion based on HSV, YCbCr, and other color models designed with a human-in-the-loop. In [36], the authors use color-to-grayscale conversion for the reference ruler tick detection and wound area measurement before applying ISODATA classification to find a threshold for the image. The combination of RGB

and grayscale images was proposed in [37]. The authors converted the RGB Region-Of-Interest using ITU-R BT.709 luma coefficients. Next, they modeled the marginal distributions of the three main classes of the ROI gray image using a linear combination of discrete Gaussians (LCDG). In [38], the authors used color-to-grayscale conversion before Otsu's thresholding to obtain the wound segment. The color-to-grayscale conversion was also used [39] before image segmentation. The Authors used grayscale versions of images to apply normalized cuts for segmentation-based detection of wheals on the skin. In [40], the authors selected the saturation plane of the HSV color model to create a single grayscale image used for wound segmentation with the active contour algorithm.

Analyzing research on the processing of wound color images reveals that many approaches use color information in semantic segmentation. However, the existing approaches have not demonstrated the role of color in semantic segmentation. Moreover, all studies that use color-to-grayscale conversion in an initial wound color image processing phase use fixed conversion weights. These weights are established based on human vision properties using color models such as YUV or HSI. So, there is a need to design and perform a study to answer research questions, including the following: 1) Does the machine-learned conversion from color to grayscale enhance semantic segmentation outcomes compared to using fixed conversion weights? 2) To what extent does color information in wound images improve the semantic segmentation results compared to the use of single component images obtained either by the traditional color-to-grayscale conversion using fixed weights or by the use of learned conversion weights?

3 Method

This section describes the proposed approach to investigate the influence of color in semantic segmentation tasks. In particular, we investigate how learned weights for color-to-grayscale conversion can impact the model's learning ability to focus on specific features with less information than RGB input.

We employed two different U-Net architectures in various configurations to investigate/ evaluate the research question. One model uses a U-Net architecture and a VGG-16 as an encoder, pretrained on the ImageNet dataset. The second model comprises the LiteSeg model using pretrained MobileNet as a backbone. Using pretrained encoders for the initial extraction of features is vital for available datasets with a limited number of cases. Other parts or encoder/decoder architecture are fully trained. Figure 1 presents a block diagram of our architecture. Overall, this architecture consists of four main blocks. First, the input RGB image is passed through a specially designed layer that could be attached to any model to learn the color-to-grayscale conversion weights. This layer is skipped for original or preprocessed RGB inputs. Additionally, we use different color conversion techniques to convert an input RGB image into: 1)Gray World white balance [41], 2) Adjusted retinex white balance [19], 3) grayscale using Luma (L), and 4) grayscale using weights. We followed traditional ITU-R BT weights to convert the RGB image into a grayscale image. After color conversion,

images are fed to the base model. The model's head performs pixel classification for semantic segmentation.

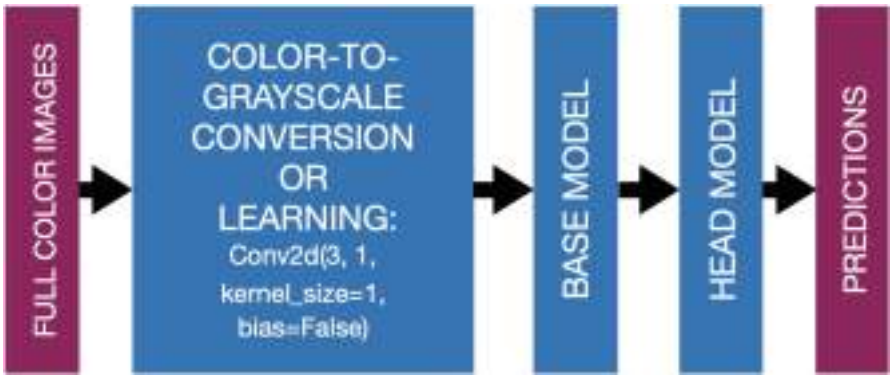

Fig. 1. Description of the proposed method.

3.1 Color Conversion

Luma (Y') or luminance (Y) is often used in many practical computer vision applications as the lightness dimension for color images. Typically, it is calculated as the weighted average of the gamma-corrected color components, i.e., R, G, and B. The weights, as presented in Eq. (1), were formulated to correspond with the cone sensitivity functions in the retina, guaranteeing that grayscale conversion is coherent with human vision.

$$Y' = W_1 \cdot R + W_2 \cdot G + W_3 \cdot B \tag{1}$$

where W_1, W_2, and W_3 are the weights corresponding to each RGB color.

As shown in Eqs. (2), (3), and (4), different sets of W_1-W_3 weights values have been proposed to represent the RGB contribution to perceived lightness, e.g., [42–44]:

$$Y'(Rec.601, SDTV) = 0.2989 \cdot R + 0.5870 \cdot G + 0.1140 \cdot B \tag{2}$$

$$Y'(Rec.709, HDTV) = 0.2126 \cdot R + 0.7152 \cdot G + 0.0722 \cdot B \tag{3}$$

$$Y'(Rec.2020, UHDTV) = 0.2627 \cdot R + 0.6780 \cdot G + 0.0593 \cdot B \tag{4}$$

The ITU-R BT.601 weights, as in Eq. (2), are often used as a popular standard for color-to-grayscale conversion (e.g., in the popular Python framework PIL). These traditional color-to-grayscale conversion methods, as in Eq. (2)–(4), were defined for human observers considering the sensitivity functions of cones in the retina. However, a machine has no such limitation, and the input grayscale

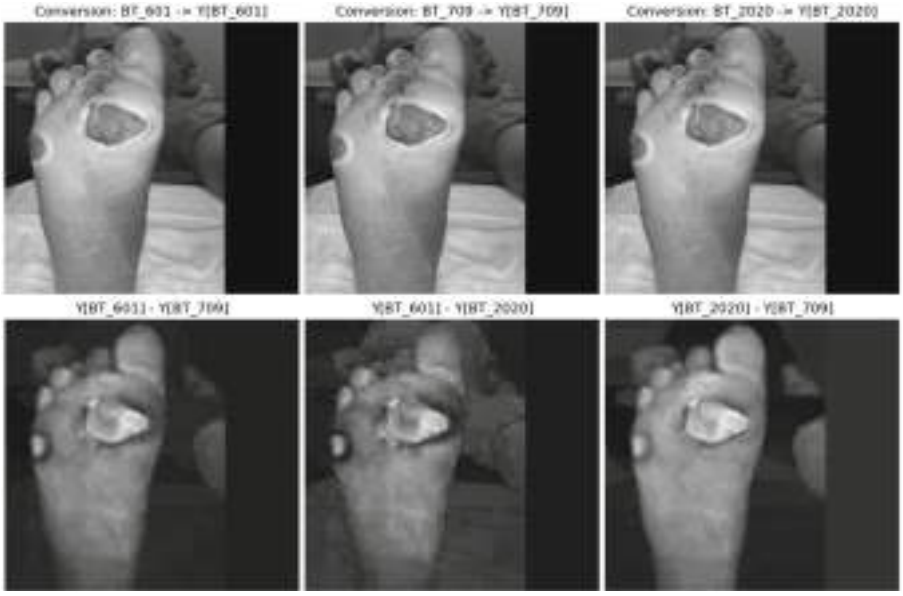

Fig. 2. Top (from left): RGB to grayscale conversion outputs using different ITU-R standards (the PIL.Image.convert('L') method uses ITU-R BT.601). Bottom (from left): differences between some conversion results.

image can be a result of data acquisition using a dedicated, single-channel camera or can be a result of a dedicated preprocessing step aimed at improving the generalization capabilities of a deep learning (DL) model. Figure 4 depicts different grayscale images using different conversion methods (Fig. 2).

In this work, we leverage the ability of CNN to learn data-driven features for learned grayscale transformation from an RGB image to a single color space. For this purpose, a CNN layer is introduced in the color conversion stage, presented in Fig. 1. In deep learning frameworks, the cross-correlation operation is used for convolution (since weights are learned). This operation is defined as in Eq. (5).

$$\mathrm{Y}(C_{\mathrm{out}}) = \mathrm{bias}(C_{\mathrm{out}}) + \sum_{k=0}^{C_{\mathrm{in}}-1} \mathrm{W}(C_{\mathrm{out}}, k) \cdot \mathrm{input}(k) \tag{5}$$

where C is the number of input/output channels, and W is the learned weights.

Our first layer is defined as Conv2D (3, 1, kernel = 1, bias = False) without activation function. Therefore, this CNN layer receives 3 channels (RGB) at the input with a kernel size of 1 and outputs a single-channel feature map without bias, as presented in Eq. (6)–(7).

$$\mathrm{Y}(C_{\mathrm{out}}) = 0 + \sum_{k=0}^{2} \mathrm{W}(C_{\mathrm{out}}, k) \cdot \mathrm{input}(k) \tag{6}$$

$$\mathrm{Y}(C_{\mathrm{out}}) = W_0 \cdot \mathrm{input}(0) + W_1 \cdot \mathrm{input}(1) + W_2 \cdot \mathrm{input}(2) \tag{7}$$

The results is equivalent to (1) with learnable weights W_0, W_1, W_2. The resultant single-channel feature map could be equivalent to a grayscale image (when a grayscale color map is used), which is optimized for machine learning rather than human visual interpretation. This CNN-based color space transformation enables the model to emphasize vital information and learn low-level features, like texture, edges, contours, and patterns. As a result of this approach, the model can avoid focusing on irrelevant information such as hue variations and extract robust feature representations, resulting in better generalization and less tendency to overfitting. This is especially beneficial when the training data is small.

4 Experiments Settings

In our experimental analysis, we used two publicly available datasets, including Foot Ulcer Segmentation Challenge (FUSeg) [45] and Advancing the Zenith of Healthcare Wound and Vascular Center (AZH-WVC) [15] datasets. Two semantic segmentation models, LiteSeg_MobileNet and UNet_VGG, with pretrained weights, are used as base models in this experiment. First, we train each model using RGB image as input and evaluate their performance on the test dataset; then, we train the model using the color conversion module in the preprocessing stage and compare the accuracy results from both configurations. For color conversion, we train models using each color-to-grayscale conversion method, including learned weights, RGB adj. retinex, RGB adj. retinex learn gray, RGB gray world, RGB gray world learn gray, Luma, and Weights.

4.1 Datasets

Details of the datasets used for training and testing are given below.

The Chronic Wound Dataset [15]**:** AZH-WVC includes 1109 photos of 889 patients with ulcer wounds. The dimensions of each image are 512×512 pixels. There are 832 photos in the training dataset and 278 in the test dataset.

FUSeg Dataset [45]**:** The Foot Ulcer Segmentation (FUSeg) dataset comprises 1210 clinical images. 810 images and their corresponding labels are available for training, while 200 images are reserved for validation. The remaining 200 images and their masks were not made public by the challenge organizer. All photos in this collection possess dimensions of 512×512 pixels.

In this study, we used only the FUSeg dataset for training. The evaluation was performed on the FUSeg validation and AZH-WVC test subsets. It is essential

to underline that no models were trained using any image from AZH-WVC to show the generalization properties of the trained models using different color-preprocessing approaches.

Figure 3 presents examples of images from both datasets. Please note that examples from the AZH-WVC dataset are presented as original images (without preprocessing). These images contain a dominant area of zero-padding values, which is challenging for segmentation using a model trained on another dataset.

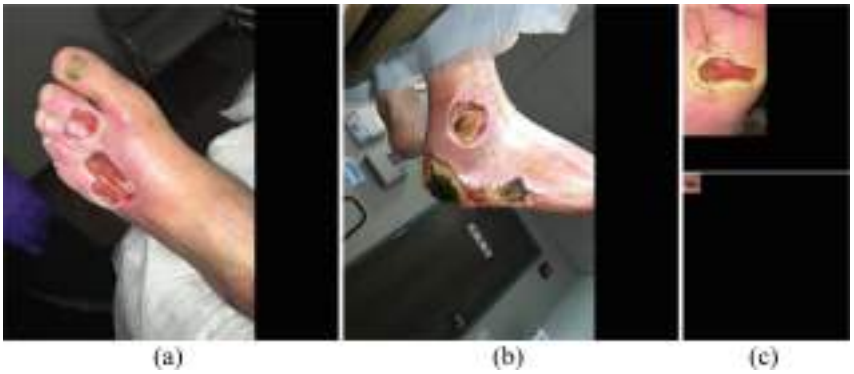

Fig. 3. Examples of wound images from FUSeg and datasets. Images (a) and (b) are from FUSeg, while images in (c) are from the AZH dataset.

4.2 Model Training

Each model with each configuration was independently (random start) trained 10 times. As a result, 80 models (8×10) were obtained for the UNet_VGG architecture, and 80 models (8×10) were obtained for the LiteSeg architecture. For statistical analysis, we used 9 best models for each configuration (one worst was skipped to keep the variance comparable between experiments). Usually, the best models are selected, but to analyze the uncertainty, we used 9 best models to analyze mean values and standard deviations of IoU, Dice, FPR, and FNR metrics.

All experiments were performed using identical training codes (except for differences in model definition) prepared in PyTorch. No augmentation was used. Other training parameters are the Adam optimizer (initial learning rate = 0.001), number of epoches = 60, cross-entropy loss function, and the StepLR learning rate scheduler with gamma = 0.1 and step size = 10.

5 Results

This section illustrates the experiment results and evaluates the segmentation performance. Different performance metrics are considered in this evaluation,

including IoU, Dice, False positive rate (FPR), and False negative rate (FNR). As an additional metric, we calculate evaluation accuracy. It was greater than 94%. However, the accuracy is calculated for both classes (with dominant background), so it is not very informative. Therefore, we report statistics on the above-mentioned segmentation metrics.

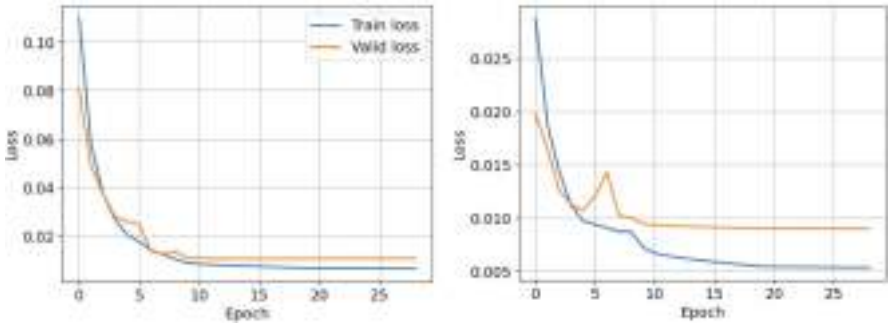

Fig. 4. Example training curves (in blue) and validation curves (in orange) for the UNet_VGG model (left) and LiteSeg model (right) on RGB images preprocessed using the adjusted Retinex method.

Figure 5 presents an example of the qualitative results of wound segmentation from the UNet-VGG model when using learned weights color-to-grayscale conversion. Visual inspection reveals that the model effectively segments the wound image precisely and accurately. While comparing the predicted mask with the actual wound image, we observe that our model with the learned weights module can accurately encompass the contour of the wounded area, which is identical to the RGB image.

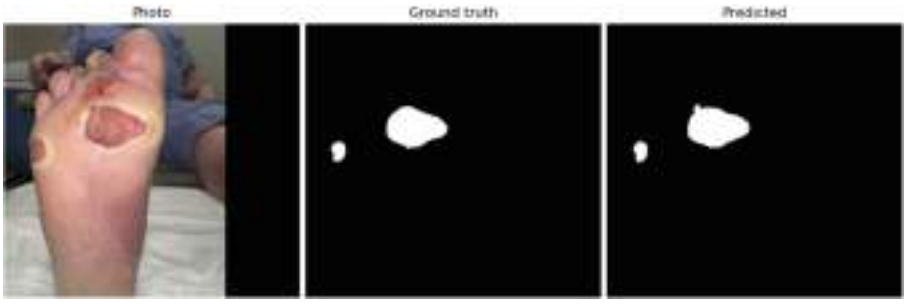

Fig. 5. From left: an example of the original photography of a foot ulcer, ground truth mask, and predicted mask using UNet-VGG model with learned weights for color-to-grayscale conversion.

Fig. 6. From left: the result of color image preprocessing (conversion to grayscale using the BT-601 standard), the output of the first layer of the UNet_VGG model (Conv2D with kernel size = 1 and no bias), the output of the color image conversion using weights of the first layer with the overlaid, predicted ulcer segment.

Figure 6 plots three different images after the color conversion process. The image on the left is after color-to-grayscale conversion using the BT-601 standard, while the image in the middle is extracted from the output of the learned weights module of the UNet-VGG model. Similarly, the image on the right is produced using the learned weights (extracted from the first layer) and manually applied to the RGB image. This shows how three learned weights, i.e., W_1, W_2, and W_3 (Eq. 1), can be applied after the knowledge obtained from the machine learning process. The last image demonstrates the segmented ulcer overlaid on the new grayscale image. It is noticeable that these new grayscale images explicitly highlight the shape of the affected area and delineate a distinct boundary line of the wound.

We further analyze the significance of each color conversion method for the given RGB input by extending our evaluation procedure with quantitative data from our experiment findings. For this evaluation, we include UNet_VGG and LiteSeg_MobileNet deep learning architectures. Both models are trained on the FUSeg dataset and tested on the FUSeg validation subset and on the AZH test dataset.

Table 1 and 2 show the testing results on the FUSeg and AZH datasets, respectively. It is noticeable that the model achieves significantly lower results using traditional, fixed conversion weights to the RGB image. However, the learned grayscale image still achieves comparable IoU and Dice scores, while the training parameters for the learned grayscale image are three times less than those for the RGB image.

We performed tests to analyze the statistical difference between results for DICE metrics obtained for selected groups. For example, the results obtained for 1) RGB input images and 2) grayscale images converted with the learned weights were not statistically different for the FUSeg validation set (p-value = 0.052 for the Mann-Whitney test, Table 1) and statistically different for the AZH test set (but p-value = 0.042 for Mann-Whitney test).

Table 1. Performance comparison of UNet_VGG on FUSeg dataset (G - grayscale)

Config	mIoU	Std IoU	mDice	Std Dice	mFPR	Std FPR	mFNR	Std FNR
RGB	0.769	0.071	0.868	0.048	0.041	0.014	0.091	0.059
G: learned weights for RGB	0.732	0.030	0.845	0.020	0.052	0.008	0.141	0.098
RGB: adj. retinex	0.785	0.015	0.880	0.009	0.051	0.008	0.088	0.047
G: l. weights adj. retinex	0.695	0.056	0.819	0.041	0.043	0.012	0.163	0.065
RGB: gray world	0.777	0.009	0.875	0.005	0.045	0.008	0.087	0.012
G: l. weights gray world	0.723	0.022	0.839	0.015	0.055	0.019	0.171	0.175
G: Luma	0.373	0.101	0.536	0.100	0.067	0.096	0.452	0.084
G: Weights	0.445	0.055	0.614	0.053	0.068	0.037	0.346	0.077

Table 2. Performance comparison of UNet_VGG on AZH dataset

Config	mIoU	Std IoU	mDice	Std Dice	mFPR	Std FPR	mFNR	Std FNR
RGB	0.454	0.080	0.620	0.078	0.005	0.003	0.375	0.081
G: learned weights for RGB	0.377	0.083	0.542	0.094	0.018	0.024	0.479	0.102
RGB: adj. retinex	0.503	0.034	0.668	0.030	0.006	0.002	0.347	0.048
G: l. weights adj. retinex	0.304	0.103	0.456	0.126	0.006	0.004	0.581	0.137
RGB: gray world	0.409	0.066	0.577	0.067	0.004	0.002	0.444	0.070
G: l. weights gray world	0.415	0.061	0.584	0.061	0.011	0.005	0.458	0.120
G: Luma	0.116	0.050	0.203	0.082	0.014	0.019	0.815	0.095
G: Weights	0.179	0.077	0.296	0.106	0.017	0.014	0.737	0.099

Table 3 and Table 4 summarize the testing results of the LiteSeg_MobileNet model on the FUSeg dataset and AZH dataset, respectively. Similar to the UNet_VGG, the LiteSeg_MobileNet model also performs adequately with learned grayscale images compared to the RGB input.

6 Discussion

In this section, we discuss the effects of color on semantic segmentation in wound images. As described earlier, deep learning models often face challenges in the accurate segmentation of ulcer wound images due to limited data availability. Semantic segmentation focuses on pixel-level classification by assigning a label to each pixel in a given image. Therefore, local and global features are instrumental for accurate performance. Local features capture fine-grained details, helping the model to find wound contours or boundary edges, while global features focus on contextual information, understanding wound shapes and similar skin patterns.

We are investigating the impact of learned weights on a single-channel image, which could aid the model in identifying the most pertinent features early on. Figure 6 presents three images produced from an RGB image (see Fig. 5 (left)), using three different color conversion methods. Compared to the image on the left

Table 3. Performance comparison of LiteSeg_MobileNet on FUSeg dataset

Config	mIoU	Std IoU	mDice	Std Dice	mFPR	Std FPR	mFNR	Std FNR
RGB	0.744	0.012	0.853	0.008	0.053	0.011	0.102	0.019
G: learned weights for RGB	0.675	0.025	0.806	0.018	0.069	0.016	0.144	0.041
RGB: adj. retinex	0.732	0.009	0.845	0.006	0.057	0.011	0.109	0.026
G: l. weights adj. retinex	0.676	0.027	0.806	0.019	0.085	0.020	0.125	0.041
RGB: gray world	0.733	0.013	0.846	0.009	0.059	0.013	0.104	0.020
G: l. weights gray world	0.674	0.031	0.805	0.022	0.077	0.022	0.136	0.030
G: Luma	0.229	0.106	0.361	0.135	0.023	0.006	0.673	0.115
G: Weights	0.259	0.083	0.405	0.108	0.052	0.057	0.597	0.107

Table 4. Performance comparison of LiteSeg_MobileNet on AZH dataset

Config	mIoU	Std IoU	mDice	Std Dice	mFPR	Std FPR	mFNR	Std FNR
RGB	0.369	0.071	0.535	0.075	0.005	0.003	0.493	0.077
G: learned weights for RGB	0.237	0.046	0.381	0.061	0.005	0.002	0.660	0.106
RGB: adj. retinex	0.440	0.037	0.610	0.035	0.007	0.004	0.415	0.059
G: l. weights adj. retinex	0.344	0.073	0.507	0.082	0.011	0.006	0.540	0.143
RGB: gray world	0.374	0.053	0.542	0.057	0.006	0.003	0.475	0.064
G: l. weights gray world	0.272	0.100	0.418	0.133	0.008	0.005	0.628	0.155
G: Luma	0.021	0.012	0.040	0.023	0.001	0.001	0.969	0.021
G: Weights	0.027	0.019	0.052	0.034	0.002	0.003	0.961	0.020

(BT-601 standard), the image in the middle (learned weighted image) amplifies the wound boundaries, which helps the model refine local features. Through this analysis, we conclude that machines can learn to interpret colors in many ways and differently than humans. This interpretation can be advantageous in improving the generalization properties of the trained ulcer segmentation model using color-to-grayscale conversion.

The quantitative results clearly show the difference in segmentation metrics between grayscale images obtained using traditional conversion methods and weights learned during the model training. For example, for the UNet_VGG model and FUSeg validation dataset (Table 1), the mean Dice value for grayscale images converted using learned weights is 0.845, while the same metric for traditional conversion methods is 0.536 and 0.614. Interestingly, the mean Dice value obtained for input RGB images was comparable to the result for grayscale images converted using learned weights (no statistical difference). The results obtained for the AZH test subset show lower values of all metrics. This could be understood since these images are much smaller and different, and examples from this domain were not used during training. However, for both models, much better results were obtained when adjusted retinex preprocessing was used (simple white balance) compared to the original RGB images. This demonstrates

the need for color calibration for RGB images (or huge training datasets with sufficient representation of all domains - often impossible) or the potential use of augmentation with grayscale images obtained for learned weights.

Furthermore, results from Tables 1, 2, 3, and 4 also show that fixed-weight methods are in line with human observation, while machines interpret the same color differently. The fixed-weight techniques do not help the deep learning models, leading to poor prediction performance. On the other hand, learned weights images present (in some configurations) comparable results with RGB images. Usually, the goal of research is to find the best model. In performed experiments, the UNet_VGG performed better for both datasets. For this model, the values of segmentation metrics for grayscale images obtained with learned weights were comparable with those obtained for RGB images. Therefore, the general conclusion is that such preprocessing is valuable and worth further investigation. The conversion weights can be learned for larger datasets and used in other preprocessing and augmentation experiments.

This study has several limitations, including a few datasets used in experiments, a limited number of models investigated, etc. We do not have many fixed-based conversion methods from three channels to a single channel, e.g., Cr, Cb, a*, b*, etc. However, preliminary results suggest the potential role of machine-oriented data preprocessing instead of traditional human-related color-to-grayscale conversion. Other methods can be compared in future research.

Other researchers used traditional color conversion methods to grayscale (e.g., [36–40]). In contrast, learned-based color-to-grayscale conversion methods can be potentially better for many tasks, mainly when fully explainable methods are applied for wound image analysis. In such cases, preprocessing using learned weights can be introduced instead of traditional conversion. Further, explainable image analysis methods can be used for wound shape analysis (e.g., monitoring of the healing process), wound classification, etc.

Thus, from these experiments, we validate potential applications of learned grayscale features in semantic segmentation, especially in ulcer wound segmentation, where wound shape, contour, edges, and texture attributes are very relevant. Additionally, learned grayscale images can partially reduce the complexity of the model, help the model learn the most critical features, and potentially avoid overfitting issues (they are less sensitive to color variations).

7 Conclusion

Early and precise assessment of ulcer wounds is crucial for proper treatment and management. Visual inspection by clinical professionals for diagnosis is labor-intensive and susceptible to human error. In contrast, computer-aided methods are highly efficient and can help early diagnosis. However, due to scarce data, deep learning methods face challenges of generalization and overfitting issues. This study aimed to address these challenges by exploring the impact of color on the segmentation of wound images. We designed a simple, universal model architecture to learn the task-specific weights for color-to-grayscale conversion.

We demonstrated that using learned weights for color-to-grayscale conversion leads to significantly better results in the semantic segmentation of wounds, achieving a mean Dice value of 0.845, compared to 0.536 and 0.614 for traditional RGB-to-grayscale conversion methods. For the best model, no statistical difference was found when comparing DICE metrics obtained for models trained on RGB images ($DICE = 0.868 \pm 0.048$) and grayscale images converted with learned weights (0.845 ± 0.020). The p-value calculated using the Mann-Whitney test for the FUSeg validation set was > 0.05. We also demonstrated that color preprocessing using adjusted retinex can improve semantic segmentation results for similar but new domains for the applied (not tuned) model. Using grayscale images as input data presents competitive performance and lower computation complexity than RGB images. Additionally, using grayscale images as input data can potentially reduce the number of model parameters. The color-to-grayscale conversion weights can be learned for much more extensive (mono-domain or multi-domain) datasets and further used in other experiments for preprocessing, augmentation, and explainable image analysis methods.

References

1. Usman M (2025) Learned Weights Foot Ulcer Segmentation Repository. https://github.com/usmanraza121/Learned_weights_FUSeg
2. Sen CK (2019) Human wounds and its burden: an updated compendium of estimates. Adv Wound Care 8(2):39–48
3. Sun B et al (2022) An optimally designed engineering exosome-reductive COF integrated nanoagent for synergistically enhanced diabetic fester wound healing. Small 18(26):2200895
4. Boulton AJM (2008) The diabetic foot. In: Controversies in treating diabetes: clinical and research aspects, pp 251–267
5. Chang M, Nguyen TT (2021) Strategy for treatment of infected diabetic foot ulcers. Acc Chem Res 54(5):1080–1093
6. Armstrong DG et al (2023) Diabetic foot ulcers: a review. Jama 330(1):62–75
7. Patry J et al (2021) Outcomes and prognosis of diabetic foot ulcers treated by an interdisciplinary team in Canada. Int Wound J 18(2):134–146
8. Yan J et al (2022) Treatment of diabetic foot during the COVID-19 pandemic: a systematic review. Medicine 101(35):e30139
9. Oota SR et al (2021) Healtech-a system for predicting patient hospitalization risk and wound progression in old patients. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision, pp 2463–2472
10. Jørgensen LB et al (2016) Methods to assess area and volume of wounds-a systematic review. Int Wound J 13(4):540–553
11. Song B, Sacan A (2012) Automated wound identification system based on image segmentation and artificial neural networks. In: 2012 IEEE international conference on bioinformatics and biomedicine, pp 1–4. https://api.semanticscholar.org/CorpusID:47463
12. Computerized segmentation and measurement of chronic wound images. Comput Biol Med 60:74–85. ISSN: 0010-4825. https://doi.org/10.1016/j.compbiomed.2015.02.015

13. Hu X et al (2019) Topology-preserving deep image segmentation. In: Advances in neural information processing systems, vol 32
14. Babu KS, Ravi KYB, Sabut S (2017) An improved watershed segmentation by flooding and pruning algorithm for assessment of diabetic wound healing. In: 2017 2nd IEEE international conference on recent trends in electronics, information & communication technology (RTEICT). IEEE, pp 679–683
15. Wang C et al (2020) Fully automatic wound segmentation with deep convolutional neural networks. Sci Rep 10(1):1–9 Article number: 21897
16. Liu H et al (2022) Wound Segmentation with Dynamic Illumination Correction and Dual-view Semantic Fusion. arXiv preprint arXiv:2207.05388
17. Cao C et al (2023) Nested segmentation and multi-level classification of diabetic foot ulcer based on mask R-CNN. Multimedia Tools Appl 82(12):18887–18906
18. Ferreira Filipe et al (2021) A systematic investigation of models for color image processing in wound size estimation. Computers 10(4):43
19. Lam EY (2005) Combining gray world and retinex theory for automatic white balance in digital photography. In: Proceedings of the ninth international symposium on consumer electronics (ISCE 2005). IEEE, pp 134–139
20. Veredas FJ, Mesa H, Morente L (2015) Efficient detection of wound-bed and peripheral skin with statistical colour models. Med Biol Eng Comput 53:345–359
21. Loizou CP et al (2012) Evaluation of wound healing process based on texture analysis. In: 2012 IEEE 12th international conference on bioinformatics & bioengineering (BIBE). IEEE, pp 709–714
22. Song B, Sacan A (2012) Automated wound identification system based on image segmentation and artificial neural networks. In: 2012 IEEE international conference on bioinformatics and biomedicine. IEEE, pp 1–4
23. Fauzi MFA et al (2021) Segmentation and management of chronic wound images: a computer-based approach. In: Chronic wounds, wound dressings and wound healing, pp 115–134
24. Fauzi M et al (2015) Computerized segmentation and measurement of chronic wound images. Comput Biol Med 60:74–85
25. Foltynski P, Ciechanowska A, Ladyzynski P (2021) Wound surface area measurement methods. Biocybern Biomed Eng 41(4):1454–1465
26. Foltynski P, Ladyzynski P (2022) Digital planimetry with a new adaptive calibration procedure results in accurate and precise wound area measurement at curved surfaces. J Diab Sci Technol 16(1):128–136
27. Goyal M et al (2017) Fully convolutional networks for diabetic foot ulcer segmentation. In: 2017 IEEE international conference on systems, man, and cybernetics (SMC). IEEE, pp 618–623
28. Deng J et al (2009) Imagenet: a large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. IEEE, pp 248–255
29. Hoiem D, Divvala SK, Hays JH (2009) Pascal VOC 2008 challenge. World Lit Today 24(1):1–4
30. Rania N, Douzi H, Yves L, Sylvie T (2020) Semantic segmentation of diabetic foot ulcer images: dealing with small dataset in DL approaches. In: El Moataz A, Mammass D, Mansouri A, Nouboud F (eds) ICISP 2020, vol 12119. LNCS. Springer, Cham, pp 162–169. https://doi.org/10.1007/978-3-030-51935-3_17
31. Cui C et al (2019) Diabetic wound segmentation using convolutional neural networks. In: 2019 41st annual international conference of the IEEE engineering in medicine and biology society. IEEE, pp 1002–1005

32. Muñoz PL, Rodríguez R, Montalvo N (2020) Automatic segmentation of diabetic foot ulcer from mask region-based convolutional neural networks. J Biomed Res Clin Invest 1(1.1006) (2020)
33. Mahbod A et al (2022) Automatic foot ulcer segmentation using an ensemble of convolutional neural networks. In: 2022 26th international conference on pattern recognition (ICPR). IEEE, pp 4358–4364
34. Yi H et al (2022) OCRNet for diabetic foot ulcer segmentation combined with edge loss. In: Diabetic foot ulcers grand challenge. Springer, pp 31–39
35. Oota SR et al (2023) WSNet: towards an effective method for wound image segmentation. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision, pp 3234–3243
36. Cazzolato MT et al (2020) Semi-automatic ulcer segmentation and wound area measurement supporting telemedicine. In: 2020 IEEE 33rd international symposium on computer-based medical systems (CBMS). IEEE, pp 356–361
37. Garcia-Zapirain B et al (2017) Automated framework for accurate segmentation of pressure ulcer images. Comput Biol Med 90:137–145
38. Satheesha TY, Satyanarayana D, Prasad G (2015) Early detection of melanoma using color and shape geometry feature. J Biomed Eng Med Imaging 2(4):33
39. Bulan O, Artan Y (2016) Wheal detection from skin prick test images using normalized-cuts and region selection. In: 2016 IEEE International Conference on Image Processing (ICIP). IEEE, pp 1250–1253
40. Hettiarachchi NDJ et al (2013) Mobile based wound measurement. In: 2013 IEEE point-of-care healthcare technologies. IEEE, pp 298–301
41. Lukac R (2018) Single-sensor imaging: methods and applications for digital cameras. CRC Press
42. International Telecommunication Union (2011) Recommendation ITU-R BT.601-7: Studio encoding parameters of digital television for standard 4:3 and widescreen 16:9 aspect ratios. Technical report BT.601-7. ITU. https://www.itu.int/dms_pubrec/itu-r/rec/bt/r-rec-bt.601-7-201103-i!!pdf-e.pdf
43. International Telecommunication Unit (2015) Recommendation ITU-R BT.709-6
44. Sugawara M, Choi S-Y, Wood D (2014) Ultra-high-definition television (Rec. ITU-R BT.2020). IEEE Signal Process Mag 31(3):170–174
45. Wang C et al (2022) FUSeg: The Foot Ulcer Segmentation Challenge. arXiv preprint arXiv:2201.00414 (2022)