

Please cite the Published Version

Cassidy, Bill , McBride, Christian, Kendrick, Connah, Reeves, Neil D, Pappachan, Joseph M, Fernandez, Cornelius J, Chacko, Elias, Brüngel, Raphael, Friedrich, Christoph M, Alotaibi, Metib, AlWabel, Abdullah Abdulaziz, Alderwish, Mohammad, Lai, Kuan-Ying and Yap, Moi Hoon  (2025) An enhanced harmonic densely connected hybrid transformer network architecture for chronic wound segmentation utilising multi-colour space tensor merging. *Computers in Biology and Medicine*, 192 (Pt A). 110172 ISSN 0010-4825

DOI: <https://doi.org/10.1016/j.compbiomed.2025.110172>

Publisher: Elsevier

Version: Published Version

Downloaded from: <https://e-space.mmu.ac.uk/640043/>

Usage rights:  [Creative Commons: Attribution-Noncommercial-No Derivative Works 4.0](https://creativecommons.org/licenses/by-nc-nd/4.0/)

Additional Information: This is an open access article published in *Computers in Biology and Medicine*, by Elsevier.

Enquiries:

If you have questions about this document, contact openresearch@mmu.ac.uk. Please include the URL of the record in e-space. If you believe that your, or a third party's rights have been compromised through this document please see our Take Down policy (available from <https://www.mmu.ac.uk/library/using-the-library/policies-and-guidelines>)



An enhanced harmonic densely connected hybrid transformer network architecture for chronic wound segmentation utilising multi-colour space tensor merging

Bill Cassidy^{a,*,} Christian McBride^a, Connah Kendrick^a, Neil D. Reeves^b, Joseph M. Pappachan^c, Cornelius J. Fernandez^d, Elias Chacko^e, Raphael Brüngel^{f,g,h}, Christoph M. Friedrich^{f,g}, Metib Alotaibiⁱ, Abdullah Abdulaziz AlWabelⁱ, Mohammad Alderwishⁱ, Kuan-Ying Lai^j, Moi Hoon Yap^{a,c}

^a Department of Computing and Mathematics, Manchester Metropolitan University, Dalton Building, Chester Street, Manchester, M1 5GD, UK

^b Medical School, Faculty of Health and Medicine, Health Innovation Campus, Lancaster University, LA1 4YW, UK

^c Lancashire Teaching Hospitals NHS Foundation Trust, Preston, PR2 9HT, UK

^d United Lincolnshire Hospitals NHS Trust, Greetwell Road, Lincoln, LN2 5QY, UK

^e Jersey General Hospital, St Helier, JE1 3QS, Jersey

^f Department of Computer Science, University of Applied Sciences and Arts Dortmund (FH Dortmund), Emil-Figge-Str. 42, 44227 Dortmund, Germany

^g Institute for Medical Informatics, Biometry and Epidemiology (IMIBE), University Hospital Essen, Zweigertstr. 37, 45130 Essen, Germany

^h Institute for Artificial Intelligence in Medicine (IKIM), University Hospital Essen, Girardetstr. 2, 45131 Essen, Germany

ⁱ University Diabetes Center, King Saud University Medical City, Riyadh, Saudi Arabia

^j Independent researcher, Taiwan

ARTICLE INFO

MSC:

41A05

41A10

65D05

65D17

Keywords:

Chronic wounds

Pressure ulcers

Venous ulcers

Arterial ulcers

Diabetic foot ulcers

Deep learning

Hybrid transformer

Segmentation

Synthetic wounds

Wounds analysis

ABSTRACT

Chronic wounds and associated complications present ever growing burdens for clinics and hospitals world wide. Venous, arterial, diabetic, and pressure wounds are becoming increasingly common globally. These conditions can result in highly debilitating repercussions for those affected, with limb amputations and increased mortality risk resulting from infection becoming more common. New methods to assist clinicians in chronic wound care are therefore vital to maintain high quality care standards. This paper presents an improved HarDNet segmentation architecture which integrates a contrast-eliminating component in the initial layers of the network to enhance feature learning. We also utilise a multi-colour space tensor merging process and adjust the harmonic shape of the convolution blocks to facilitate these additional features. We train our proposed model using wound images from light skinned patients and test the model on two test sets (one set with ground truth, and one without) comprising only darker skinned cases. Subjective ratings are obtained from clinical wound experts with intraclass correlation coefficient used to determine inter-rater reliability. For the dark skin tone test set with ground truth, when comparing the baseline results ($DSC = 0.6389$, $IoU = 0.5350$) with the results for the proposed model ($DSC = 0.7610$, $IoU = 0.6620$) we demonstrate improvements in terms of Dice similarity coefficient (+0.1221) and intersection over union (+0.1270). Measures from the qualitative analysis also indicate improvements in terms of high expert ratings, with improvements of > 3% demonstrated when comparing the baseline model with the proposed model. This paper presents the first study to focus on darker skin tones for chronic wound segmentation using models trained only on wound images exhibiting lighter skin. Diabetes is highly prevalent in countries where patients have darker skin tones, highlighting the need for a greater focus on such cases. Additionally, we conduct the largest qualitative study to date for chronic wound segmentation. All source code for this study is available at: <https://github.com/mmu-dermatology-research/hardnet-cws>.

* Corresponding author.

E-mail address: bill.cassidy@stu.mmu.ac.uk (B. Cassidy).

<https://doi.org/10.1016/j.combiomed.2025.110172>

Received 21 October 2024; Received in revised form 4 April 2025; Accepted 7 April 2025

Available online 2 May 2025

0010-4825/© 2025 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

1. Introduction

Diabetes is now regarded as a global epidemic, resulting in most part from a systematic increase in populations becoming overweight and obese [1]. Programmes that target the condition have historically shown only short-term benefits, with longer-term effects yet to be established [2,3]. The situation is similar for obesity [4], a common factor in diabetes occurrence [5]. Arterial leg ulcers (ALUs) and diabetic foot ulcers (DFUs) are a debilitating and costly complication of diabetes [1], with recent findings suggesting an association between DFU episodes and all-cause resource utilisation and increased mortality risk [6]. Venous leg ulcers (VLUs) and pressure ulcers (PRUs) are the most common types of complex skin ulcers [7], with ulcer prevalence in the diabetic population estimated to be 13% in North America [8]. The incidence of chronic wounds is high and is estimated to continue on an upward trajectory [9].

Patients diagnosed with DFU are two to three times more likely to die than patients without and are predisposed to numerous comorbidities, including peripheral artery disease, cardiovascular disease, neuropathy, retinopathy, and nephropathy. VLUs and DFUs often result in significantly impaired quality of life [10–12]. Occurrence of ulcers is linked to an increased incidence of both amputation and mortality, especially in the presence of advanced age, peripheral artery disease and anemia [10,13,14]. Chronic wounds exert a significant physical and emotional burden on patients [15,16], with depression being associated with an increased risk at initial and subsequent occurrence [17,18].

Chronic wounds are typically correlated with comorbidities such as diabetes, vascular deficits, hypertension, and chronic kidney disease [19]. Diabetic neuropathy is highly prevalent in DFU cases and is the primary cause of DFU formation [20], meaning that patients have lost sensation in their foot due to nerve damage [21]. This means that patients often go through long periods not realising they have a DFU until the wound becomes much worse and leads to other serious complications. Infection affects more than 50% of all DFU cases [22] and represents one of the most common causes of diabetes related hospitalisation [20]. Diabetic leg and foot ulcers are amongst the most expensive wound types to treat in the United States [19]. For VLUs, the recurrence rate within 3 months after wound closure is as high as 70% [10].

Management of chronic wounds can be a long and difficult task, for both patient and clinician. This is especially true for wounds that are not caught early, and require more intensive treatment programmes. This can mean frequent visits to clinics or hospitals for assessment by experts [23,24]. Even after accomplished wound healing, recurrences are frequent and often lead to minor or major amputation of lower extremities [25,26]. The post COVID-19 climate poses further risks and challenges to the treatment of chronic wounds, given that diabetic patients are placed in the high-risk category. To this end, recent years have seen an increased research interest in the remote detection and monitoring of wounds using non-contact methods [27–29].

Evolving current telemedicine systems to include remote wound monitoring represents an opportunity to reduce risks to vulnerable patients and to ease significantly overburdened healthcare systems [30]. Furthermore, the advent of cheap consumer mobile devices and easily accessible cloud platforms promotes the idea of making these technologies available to poorer regions, where patients may experience reduced access to expert healthcare providers. Low cost, easy-to-use non-invasive devices that can detect and monitor wounds could act as a mechanism to promote patient engagement with the monitoring of their health.

A growing body of evidence has shown the ability of convolutional neural networks (CNNs) to equal or surpass experienced dermatologists for detection and classification in related domains [31–37]. In this regard, deep learning may be able to assist in providing more objective results in domains which are prone to high levels of subjectivity. Changes to wound area have been shown to be a robust predictor in

healing status [38]. Segmentation of chronic wounds allows for more accurate assessment of changes to wound shape and size over time when compared to more generalised localisation techniques. In the next section, we discuss the recent notable developments in this domain.

2. Related work

Studies on deep learning tasks related to chronic wounds have become a growing interest in the research community in recent years due to the possible benefits that such technologies might offer in real-world clinical settings [39,40]. In this section, we examine the more prominent studies conducted in chronic wound segmentation research that have helped to guide the experiments presented in this paper.

Goyal et al. [41] were one of the first research groups to investigate chronic wound segmentation using convolutional neural networks (CNNs). They trained a number of fully convolutional networks (FCN) to segment DFU wounds and associated periwounds using a dataset comprising 600 DFU images together with ground truth masks which were provided by wound experts at Lancashire Teaching Hospitals (LTH), UK. A two-tier transfer learning approach using two publicly available general image datasets was used - Pascal VOC and ImageNet segmentation datasets. The DFU segmentation dataset was divided into 420 training images, 60 validation images, 120 test images, and 105 images of healthy feet. In the joint segmentation of wound and periwound regions the highest performing model was FCN32-s with a Dice similarity coefficient (DSC) of 0.899. For segmentation of ulcer regions only, the highest performing model was FCN-16s, reporting a DSC of 0.794. For segmentation of only periwounds, the highest performing model was FCN-16s, reporting a DSC of 0.851. This work noted that the FCN-AlexNet and FCN-32s models were less accurate in the segmentation of irregular boundaries, and that the smaller pixel strides used in FCN-16s and FCN-8s resulted in improved detection of such examples. This study also observed an overlap of periwound and wound regions in prediction results due to ambiguities in feature boundaries. A limitation of this work is the small number of samples used in the experiments, which may make the results difficult to generalise across more diverse datasets.

Wang et al. [42] conducted wound segmentation experiments using MobileNetV2, which was pretrained using the Pascal VOC segmentation dataset. For training and testing, they used a newly introduced dataset of 1109 DFU images ($train = 831; test = 278$). A localisation method was used as a preprocessing stage to exclude non-DFU wound regions from images before the segmentation stage. As a post-processing step, morphological algorithms were used (small region removal and hole-filling). Their test results reported a mean DSC of 0.9047. However, this work presents several limitations. First, all wound images were very small patches that are heavily padded to a resolution of 224×224 pixels. Wound pixels therefore comprised only very small regions of the images. Excluding padding, the average size of the wound regions in the training set is 71×104 pixels, and the average wound region size in the test set is 70×101 pixels. At such low resolutions, as small as 17×18 pixels, a large number of wound features may be lost. They also tested their model on the Medetec dataset, and obtained a DSC of 0.9405.

In later works, Wang et al. [43] conducted the Foot Ulcer Segmentation Challenge (FUSC) 2021 whereby a new DFU dataset was released ($train = 810, val = 200, test = 200$). This new dataset comprised of examples with less significant padding compared to their prior dataset, with images exhibiting more foot and background features. The winner of the FUSC 2021, Mahbod et al. [44], achieved an image-based DSC of 0.8880, which was 1.67% lower than the prior DSC reported by Wang et al. [42]. This may indicate that the task was more difficult when larger wound images were introduced. In the FUSC 2021, models were required to learn features that are more complex that were absent from the prior experiments conducted by Wang et al. [43] which used a smaller dataset comprising notably smaller wound regions and thus fewer features.

Scebba et al. [45] noted the numerous challenges associated with wound segmentation, including wound type heterogeneity, variance in tissue colouration, wound shapes, background features, anatomical location, variety of image capturing scenarios, and non-standard specifications of capture devices. They observed that standardisation initiatives in medical wound photography may lead to additional workload burdens on clinical routine, and that the proposal of standards would likely not result in a desired consistent approach in real-world scenarios. Their proposed method utilised a MobileNet localisation model to assist a U-Net segmentation model to reduce non-wound features. This study used a total of five chronic wound datasets: (1) SwissWOU - a private dataset of DFU ($n = 1096$) and systemic sclerosis digital ulcers ($n = 63$), (2) SIH (second healing intention dataset) ($n = 58$) [46], (3) DFUC2020 ($n = 2000$) [47], (4) FUSC ($n = 60$) [43], (5) Medetec ($n = 53$) [48]. We observe that for some of the datasets used in this study, complete sets were not utilised in the experiments. For the FUSC, Medetec, and SIH datasets, only a selection of images were used. The authors experimented using a range of well-known segmentation networks, both with and without localisation preprocessing (manual and automated). When tested using only the SwissWOU DFU images (10% of all patients), their results showed that U-Net was the highest performing network ($MCC = 0.85$, $IoU = 0.75$). Their test results for the SwissWOU systemic digital ulcers, Medetec, SIH, and FUSC images also showed U-Net to be the best performing network ($MCC = 0.8725$, $IoU = 0.7875$).

HardNet-DFUS (Harmonic Densely Connected Network), proposed by Liao et al. [49], was the winning entry for the DFUC 2022, achieving a DSC of 0.7287. The design is based on a prior work, HardNet-MSEG [50], and is the basis of our proposed methods in the present paper. HardNet-DFUS uses inter-layer connections which were configured according to the required block depth n . Therefore, when $n = 9$, the resulting factors are 1, 3, and 9, allowing for shortcuts to the 1st, 3rd, and 9th convolutions. This results in the removal of the power of 2 constraint found in the original block design. A block depth of 3, 9, and 15 was selected for the final design, replacing the original depth of 4, 9, and 16. This results in reduced data movement using the same number of convolutional layers. Additionally, they replaced the receptive field blocks (RFB) in the decoder with a large window attention (Lawin) transformer. The original HardNet network mainly utilised 3×3 convolutions to increase computational density, which changes the model from being memory-bound to compute-bound [51]. To increase accuracy further, they used an ensembling strategy using 5-fold cross validation and test time augmentation (TTA). Augmented images were added to the test set when testing the sub-models, with the output averages used as the final prediction results. However, they found that this method was not consistent, and would sometimes degrade performance in terms of DSC and IoU.

Ramachandram et al. [52] proposed a chronic wound segmentation network for tissue type segmentation (AutoTissue) and wound segmentation (AutoTrace) designed for use in a commercial mobile app. The AutoTrace model implemented a typical auto-encoder design using depth-wise separable convolution layers, attention gates, and strided depth-wise convolutions resulting in downsample activations which act as an alternative to fixed max-pooling. Additive attention gates were added to skip connections to regulate activations from previous network layers. Bilinear upsampling was used in the decoder blocks followed by depth-wise separable convolution layers, helping to reduce memory requirements. The AutoTissue segmentation model implemented EfficientNetB0 as the encoder path, with a decoder comprising 4 layers with each layer utilising two-dimensional bilinear upsampling followed by 2 depth-wise convolution layers. AutoTrace was trained with a private dataset comprising 467,000 wound images, while AutoTissue was trained with a second private dataset comprising 17,000 wound images. For both datasets, both images and ground truth labels were obtained from hospitals in North America, allowing for a diverse range of wound images. However, details were not disclosed

regarding the exact composition of the datasets. The study reported an mIoU of 0.8644 for wound segmentation and an mIoU of 0.7192 for tissue and wound segmentation. Clinicians rated 91% (53/58) of the results as between fair and good for segmentation and tissue segmentation quality. Qualitative assessment of this type is rare in chronic wound related deep learning studies. However, the sample size used is limited, whereby only 58 examples were rated.

Swerdlow et al. [53] used a private dataset exhibiting stages 1–4 PRUs, acquired from eKare Inc. Mask R-CNN with a ResNet101 backbone was trained for segmentation and classification of each PRU stage of development. The dataset comprised 969 PRU images ($t_{train} = 848$, $t_{test} = 121$). The study reported a DSC of 0.92 for stage 1 PRU, 0.85 for stage 2 PRU, 0.93 for stage 3 PRU, and 0.91 for stage 4 PRU. The wound image acquisition protocol indicated that images be taken from approximately 40–65 cm distance from the wound. Additionally, the study excluded PRU wounds that were smaller than 2×2 cm, which may have limited testing of the model's true ability to segment a range of wound sizes.

The use of different colour spaces in CNNs was explored by Gowda et al. [54]. Their classification experiments on the CIFAR-10, CIFAR-100, SVHN, and ImageNet datasets showed that different classes were sensitive to models trained on different colour spaces. They trained a series of DenseNet models using multiple image datasets that had been converted to different colour spaces, with each DenseNet using a different colour space as input. The outputs from each DenseNet were then used as input into a final dense layer to generate weighted predictions from each sub-DenseNet. Increased computational overhead, a result of using multiple DenseNets, was addressed by using smaller and wider DenseNets. This work showed that training with images from multiple colour spaces provided comparable results to significantly larger models, such as DenseNet-BC-190-40, with a reduction of more than 10M parameters.

In later CNN-based colour space studies, Simon et al. [55] trained classification models using RGB and luminance images. Their experiments utilised a ResNet101 pretrained model for feature learning and an SVM for the classifier. They trained and tested their model with the Describable Texture Dataset (DTD) and the Flickr Material Dataset (FMD). Compared to prior works, for the DTD, they reported an accuracy improvement of 0.73%, and for the FMD they reported an accuracy improvement of 6.95%.

In more recent work, McBride et al. [56] conducted preliminary experiments which merged individual colour channels from different colour spaces into single tensors when training a chronic wound U-Net segmentation model. They found that different colour channel merging operations using RGB, CIELAB, and YCrCb colour spaces improved segmentation performance by 0.0264 for IoU and 0.0348 for DSC when testing on the FUSC dataset. However, this study was limited by the use of only a simple U-Net model.

One of the most prominent aspects of chronic wound research in deep learning, as highlighted by our literature review, has been a lack of substantial publicly available fully annotated datasets. Another notable factor in the field is a lack of focus on patients exhibiting darker skin tones. The biases towards lighter skin tones present in deep learning models in dermatology research is well established [57]. Benčević et al. [58] observed significant bias in skin lesion segmentation against darker skin cases when performing in and out-of-sample evaluation. Furthermore, they also found that methods used to mitigate bias do not result in significant bias reduction. Most of the publicly available chronic wound datasets comprise cases that were collected from lighter skin patients. While some datasets do contain examples with darker skin tones, these are not quantified. In the next section, we discuss the chronic wound datasets that we used in our experiments.

Table 1

A summary of public and private chronic wound image datasets used in our experiments. Note that the Train, Val, and Test columns show how the datasets were originally divided. YWHD - Yang Wound Healing dataset; AZH - Advancing the Zenith of Healthcare Wound Care dataset; FUSC - Foot Ulcer Segmentation Challenge dataset; GIS-W - Google Image Search wound images; CWDB - Complex Wound DB; Wseg - Wound Segmentation dataset; KSUMC - King Saud University Medical City dataset.

Name	Resolution	Task	Train	Val	Test	Total	Status
Medetec [48]	560 × (347–444)	Segmentation	–	–	–	608	Public
YWHD [46]	5184 × 3456	–	–	–	–	201	Public
Alzubaidi [72]	Various	Classification	–	–	–	493	Public
AZH [42]	224 × 224	Segmentation	831 ^b	–	278 ^b	1109	Public
DFUC 2022 ^a [68]	640 × 480	Segmentation	2000 ^b	–	2000 ^b	4000	Public
FUSC [73]	512 × 512	Segmentation	810 ^b	200 ^b	200	1210	Public
Fitzpatrick17k [74]	Various	–	–	–	–	16,529	Public
WoundsDB [75]	4896 × 3264	Multimodal	–	–	–	188 ^b	Public
GIS-W	Various	–	–	–	–	186	Public
CWDB [76]	Various	Segmentation	–	–	–	27 ^b	Public
Wseg [77]	331 × 331	Segmentation	–	–	–	2686	Public
KSUMC	Various	Multimodal	–	–	–	115	Private

^a Includes pathology class and anatomical location labels.

^b Includes ground truth masks available to the present study.

3. Chronic wound datasets

Large medical imaging datasets present notable challenges when used to train deep learning networks [57]. Issues such as image duplication, image and feature similarity [59], varying image quality, label noise and the presence of visual artefacts can significantly impact model performance [60–65].

Our research group has been responsible for the release of the first substantial publicly available DFU wound datasets with ground truth labels [66–68]. With the release of each dataset, we have conducted yearly challenges in association with the International Conference on Medical Image Computing and Computer Assisted Intervention [47, 66,69–71]. Our datasets comprise of over 20,000 high quality DFU wound photographs together with internationally coordinated clinical labelling provided by experts in podiatry. Table 1 shows a summary of all the datasets used in our chronic wound segmentation experiments. We use 10 public datasets, 1 private dataset, and a dataset comprising Google Image Search images which we collected using the Creative Commons License search option to remove copyrighted images from search results. These images vary significantly, both in size and quality. To obtain these images, we used search terms such as “diabetic foot ulcer”, “neuropathic ulcer”, “venous ulcer”, “pressure ulcer”, “wound”, and “chronic wound”.

The private dataset used in our experiments is the The King Saud University Medical City (KSUMC) dataset. This dataset comprises 115 DFU wound images and was obtained from the King Saud University Medical City, Saudi Arabia. The images were acquired using a Fujifilm Finepix SL260 digital camera at various resolutions and orientations. The KSUMC dataset was obtained with ethical approval from King Saud University Medical City, Saudi Arabia (REF: 24/1159/IRB).

3.1. Expert wound delineation

All training, validation and test cases for the DFUC 2022 dataset were delineated with the location of DFUs in polygon coordinates. The VGG Image Annotator tool [78,79] was used to delineate images with polygons indicating the ulcer region. The ground truth was produced by five healthcare professionals who specialise in treating diabetic foot ulcers and associated pathology, comprising consultant physicians and podiatrists, all with more than 5 years professional experience. Each clinical expert delineated a portion of the dataset. The instruction for annotation was to delineate each DFU with a polygon region.

Prior to the delineation of the full dataset, 800 cases were provided to two clinical experts for the purposes of evaluating expert agreement. The 800 cases represent 20% of the full dataset and were chosen at random. Agreement was assessed using intersection over union (IoU) and DSC. The delineation between experts in terms of IoU and DSC is

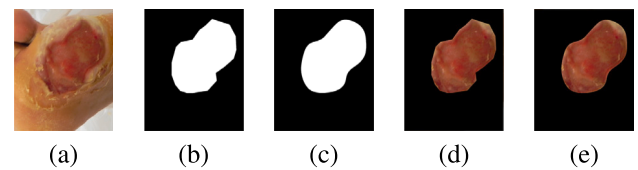


Fig. 1. Illustration of an image from the DFUC 2022 training set and corresponding masks: (a) original image; (b) original mask based on clinician delineation; (c) original mask processed using active contour model; (d) original image with clinician delineation mask overlaid; and (e) original image with original mask processed using active contour model overlaid. Note that images were cropped for illustration purposes.

0.5876 ± 0.2670 and 0.6981 ± 0.2544 respectively, with a recorded accuracy of 0.9869 ± 0.0291 . The delineations for the 800 cases used for reliability evaluation were not used in the subsequent release of the full dataset.

The use of active contour masks when used as ground truth has been shown to provide superior agreement with machine predicted results in chronic wound segmentation tasks [68]. Therefore, in our experiments, for the DFUC 2022 dataset we use ground truth masks that have been processed using the original polygon delineations with an active contour model applied to smooth delineated vertices. The active contour model masks were produced using the MATLAB (The MathWorks, Inc., Massachusetts) method created by Kroon et al. [80], using default parameters. Fig. 1 shows an example of the two different mask types applied to a training image from the DFUC 2022 dataset. To further validate that the smoothing effect did not alter the delineation of the experts, we measure the similarity of the masks produced by the clinicians and the masks post-processed by active contour on the training set. The DSC is 0.9620 ± 0.0259 and the accuracy is 0.9991 ± 0.0012 . These evaluations support our statement that the pre-processing stage has provided a smoothing effect, but did not significantly alter the experts' delineation.

4. Method

This section details the training, validation, and testing workflow, proposed model architecture, and corresponding metrics used for our segmentation experiments.

4.1. Metrics

We utilised a series of widely used evaluation metrics to determine the accuracy of the models trained, validated, and tested in our wound segmentation experiments. IoU and DSC were selected as the main metrics for determining segmentation model accuracy. DSC was chosen

for its representation as the harmonic mean of precision and recall, giving a balanced evaluation between false positive and false negative predictions. A true positive prediction is defined as a segmentation result where the wound pixels are correctly segmented. A false positive prediction is defined as a segmentation result which includes non-wound pixels. A true negative prediction is defined as a segmentation result where non-wound pixels have not been segmented. A false negative prediction is defined as a segmentation result where wound pixels have not been segmented. The mathematical expressions for IoU and DSC are shown in Eqs. (1) and (2) respectively:

$$IoU = \frac{|X \cap Y|}{|X \cup Y|} \quad (1)$$

$$DSC = 2 \times \frac{|X \cap Y|}{|X| + |Y|} \quad (2)$$

where X represents the ground truth mask and Y represents the prediction mask.

We also utilise two additional statistical hypothesis testing metrics to better understand the Type I and Type II errors associated with deep learning segmentation algorithm performance. The two additional metrics we use are False Positive Error (FPE) and False Negative Error (FNE) which are defined as follows:

$$FPE = \frac{FP}{FP + TN} \quad (3)$$

$$FNE = \frac{FN}{FN + TP} \quad (4)$$

where FP is the total number of false positive predictions, TN is the total number of true negative predictions, and FN is the total number of false negative predictions.

4.2. Baseline experiments

The first stage in our experiments was to determine the effectiveness of a range of deep learning segmentation networks using the largest publicly available chronic wound dataset (DFUC 2022). We obtained a series of baselines for training, validation, and test results for a selection of newer segmentation architectures using the DFUC 2022 dataset. A single dataset (DFUC 2022) was used for the purposes of brevity, given the number of different model architectures requiring training, validation, testing, and subsequent analysis. This information also serves as a useful reference to researchers who do not have access to the private KSUMC dataset used later in the present study. Due to the limited availability of chronic wound images exhibiting darker skin tones with ground truth segmentation masks, later testing experiments presented in this paper require the use of images from both the DFUC 2022 training and testing sets. Therefore, the initial baseline test set has a different composition compared to the test sets used in the later experiments for the proposed model. A second baseline is established with the new test sets for the proposed model to enable fair comparison in subsequent experiments. For the initial baseline experiments, we focus on a selection of more advanced CNN architectures that were not included in the previous baseline experiments reported for DFUC 2022 [71]. For all baseline experiments, the DFUC 2022 dataset images and masks were unchanged from their original resolution (640 × 480 pixels). A total of 200 images were taken at random from the training set for use as the validation set during training. No augmentation or post-processing methods were used in any of the baseline experiments. All baseline models were trained for 300 epochs with a batch size of 2 using the Adam optimiser with a learning rate of 0.001, and a weight decay of 0.0001. All models were trained without the use of pretrained weights. The best model for each experiment was selected from the 300 epochs training schedule determined by the highest validation IoU and DSC values. The hardware and software configuration for all experiments completed in the present paper was as follows: Debian GNU/Linux 10 (buster) operating system, AMD Ryzen 9 3900X 12-Core CPU, 128 GB RAM, NVIDIA GeForce RTX 3090 24 GB GPU. Models

were trained with Tensorflow 2.4.1 and Pytorch 1.13.1 using Python 3.7.13.

The results of the baseline experiments are summarised in Table 2. HardNet-DFUS is clearly the best overall performing network in terms of training ($IoU = 0.7889$, $DSC = 0.8743$), validation ($IoU = 0.6068$, $DSC = 0.7101$), and test metrics ($IoU = 0.5421$, $DSC = 0.6520$, $FPE = 0.0255$, $FNE = 0.3278$). We observe that the EfficientNet U-Nets record lower training and validation loss rates at 0.1558 (EffNetB0 U-Net) and 0.3485 (EffNetB1 U-Net) respectively. These loss rates are significant, a reduction of 0.1043 for B0 train loss and a reduction of 0.1125 for B1 validation loss. However, these performance gains are not reflected in the test loss results when comparing the EfficientNets with HardNet. The notable differences between validation and test results for the best overall performing network (HardNet-DFUS) may be indicative of the random nature of the validation set, which might not fully represent the range of features present in the test set. We observe that the deeper U-Net variants such as U-Net++ and ResUNet++ demonstrated particularly low metrics, which may be a consequence of the relatively small size of the DFUC 2022 dataset and the larger size of these network architectures.

In addition to the range of network architectures reported on in Table 2, we also trained, validated, and tested a number of vision transformer (ViT) segmentation models. However, the test results for the ViTs were well below those reported in Table 2. As reported by [90], ViTs require substantial amounts of training data and are not suitable for use with very small datasets such as those used in the present paper. [90] observed that representation similarity between ViTs trained on small and large datasets comprising of > 1M images differed substantially. They posit that this may be due to a reduction in inductive bias (the relationship between closely positioned input features). Their experiments show that lower layers of ViTs are not able to sufficiently learn local relationships when small amounts of complex data are used. Conversely, recent work by [91] suggests that ViTs might be trained on smaller datasets using self-supervised inductive biases. However, even in these scenarios, datasets of up to 100,000 images were used, which although might be considered small in deep learning terms, is still significantly greater than the current publicly available chronic wound datasets.

We compared a selection of ground truth masks with model predictions for the best performing network in the initial baseline results, which was HardNet-DFUS. Fig. 2 shows 3 cases with original image, ground truth labels, and corresponding baseline model predictions. The first row shows a case where the ground truth mask includes the wound and periwound as a single region, whereas the model predicted only the unhealed wound region. The second row shows a case where the two wound regions are separated by epithelial skin, indicating significant healing between the two non-healed regions. The corresponding prediction shows that only the main wound region was predicted by the model. The third row shows a case where two large wound regions are separated by an epithelial region. The ground truth includes both wound regions and the partially healed region. However, the prediction includes only the non-healed regions. These examples demonstrate the significant challenges inherent in human expert wound delineation and how delineation of wound regions can be highly subjective. We asked two clinical experts in wound care (a consultant surgeon and a consultant podiatrist) to indicate agreement with the ground truth labels and corresponding model predictions for the 3 cases shown in Fig. 2. Both experts agreed that the model predictions, although not perfect, were of higher quality than the ground truth labels. Both experts indicated that the automated segmentation of non-healed wound regions was more important than segmentation of healed tissue in terms of automated wound monitoring. We note that these qualitative observations are preliminary and are not to be considered conclusive. The intention is to demonstrate issues present in both expert delineation and limitations of the baseline model. Larger scale qualitative assessment is explored later in the paper.

Table 2

Baseline results for a selection of deep learning segmentation networks trained, validated and tested on the DFUC 2022 dataset (image size = 640×480 pixels). IoU - intersection over union; DSC - Dice similarity coefficient; FPE - false positive error; FNE - false negative error; DCSA - deeper more compact split-attention; MBS - multi-branch segmentation; EffNet - EfficientNet. ConvNeXt U-Net was trained using the *convnext* base backbone. Note that none of the networks evaluated used pretraining.

Model	Epoch	Train IoU	Train loss	Train DSC	Val IoU	Val loss	Val DSC	Test IoU	Test DSC	FPE	FNE
ResUNet++ [81]	152	0.6015	0.3238	0.7213	0.4495	0.6245	0.5767	0.3798	0.4969	0.4315	0.3967
U-Net++ [82]	279	0.6694	0.2662	0.7826	0.5147	0.4505	0.6451	0.3996	0.5179	0.4057	0.4152
Attention U-Net [83]	65	0.6710	0.2671	0.7835	0.5352	0.4157	0.6552	0.4135	0.5302	0.3760	0.4238
DCSAU-Net [84]	245	0.5657	0.3653	0.6887	0.4467	0.5395	0.5712	0.3627	0.4736	0.4498	0.4298
MBSNet [85]	81	0.6979	0.2332	0.7999	0.5195	0.4524	0.6446	0.3977	0.5102	0.4240	0.3979
ResNet50 U-Net [86]	196	0.6424	0.2892	0.7578	0.4924	0.4612	0.6211	0.3732	0.4915	0.3878	0.4712
MobileNetV2 U-Net [86]	34	0.6884	0.2485	0.7946	0.5624	0.3912	0.6844	0.4406	0.5597	0.3542	0.3975
ConvNeXt U-Net [87]	98	0.5529	0.3574	0.6869	0.4339	0.5016	0.5728	0.3087	0.4289	0.4157	0.5476
EffNetB0 U-Net [87]	258	0.7817	0.1558	0.8686	0.5846	0.3693	0.7044	0.4616	0.5784	0.3474	0.3813
EffNetB1 U-Net [87]	38	0.6856	0.2388	0.7935	0.5844	0.3485	0.7038	0.4584	0.5785	0.3396	0.3807
EffNetB2 U-Net [87]	184	0.7575	0.1748	0.8515	0.5843	0.3613	0.7026	0.4461	0.5641	0.3648	0.3828
UNeXt [88]	96	0.4580	0.4844	0.5895	0.4398	0.5128	0.5695	0.3383	0.4596	0.0464	0.4660
HarDNet-DFUS [89]	33	0.7889	0.2601	0.8743	0.6068	0.4610	0.7101	0.5421	0.6520	0.0255	0.3278

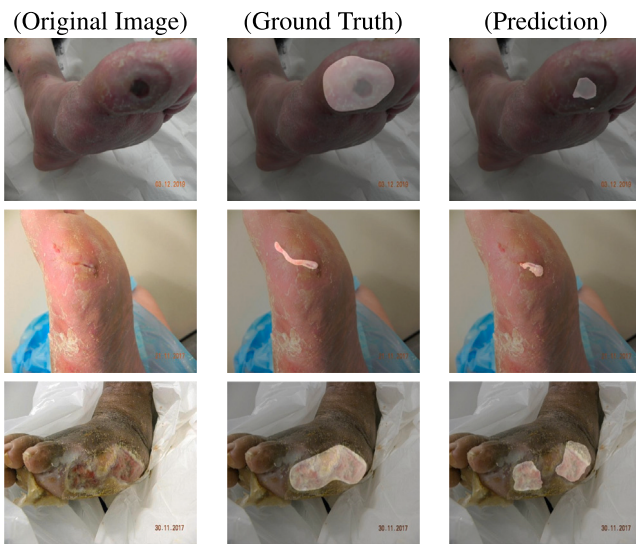


Fig. 2. Illustration of 3 cases from the DFUC 2022 dataset where clinical experts determined the baseline model predictions (HarDNet-DFUS) to be superior to the ground truth labels. The first column shows the original images, the second column shows the ground truth, and the third column shows the model predictions.

We observe that many of the segmentation models that performed highly in other medical imaging domains, such as DCSAU-Net which reported state-of-the-art performance on polyp, multiple myeloma plasma cells, ISIC 2018, and brain tumour segmentation, did not perform well when trained and tested on DFU wounds. We posit that this is due to the larger range of features found across chronic wounds at different stages of development, in addition to the significant visual complexity of such wounds.

4.3. Construction of training, validation, and test sets

The aim of our work is to determine the effectiveness of a segmentation model, trained and validated only on patients with lighter skin tones, to segment wounds on patients with darker skin tones. To this end, we construct a series of datasets for use in our experiments. Our approach for this task was to use all publicly available chronic wound datasets that have ground truth masks, together with all datasets that we have access to privately. Test set images were selected based on Fitzpatrick [92] skin types IV (moderate brown skin), V (dark brown skin), and VII (deeply pigmented dark brown or black skin). To create the first test set (test set A), we gathered all images with masks exhibiting darker skin tones from the DFUC 2022 dataset (68 images and corresponding masks from the training and test sets), the

Table 3

Summary of the composition of the new dataset used for training, validation, and testing purposes. Note that the training and validation sets comprise only of wound images from light skinned patients, whereas the test set (test set A) comprises only wound images from patients with darker skin tones.

Dataset	Train	Validation	Test set A
DFUC 2022	3893	0	68
AZH	824	173	81
CWDB	0	24	3
FUSC	0	795	190
WoundsDB	0	188	0
Total	4717	1180	342

AZH dataset (81 images and corresponding masks from the training and test sets), the CWDB dataset (3 images and corresponding masks), and the FUSC dataset (190 images and corresponding masks from the training and validation sets). Test set A comprises all publicly available wound images with segmentation masks from patients with dark skin tones. To create the new training set, we combined the remaining DFUC 2022 training and test sets (3893 images and corresponding masks) with 824 images and corresponding masks from the AZH training and test sets. For the validation set, we use the remaining 173 AZH images and corresponding masks together with all 24 CWDB images and masks, all 795 FUSC training and validation images and masks, and all 188 WoundsDB images and masks. Finally, we created a second test set (test set B) which comprises the same number of images as test set A ($n = 342$) and includes only dark skin tone wound images which do not have ground truth masks and will be assessed qualitatively. Test set B includes wound images from the Alzubaidi dataset ($n = 52$), the Fitzpatrick17k dataset ($n = 4$), the FUSC test set ($n = 35$), the GISW dataset ($n = 13$), the Medetec dataset ($n = 8$), the Wseg dataset ($n = 115$), and the KSUMC dataset ($n = 115$). A summary of the dataset composition for training, validation, and testing (test set A) is shown in Table 3. A summary of test sets A and B is shown in Table 4.

4.4. HarDNet-DFUS architecture

Following the analysis of the initial baseline results, we selected the HarDNet-DFUS network architecture used for the winning entry for DFUC 2022, proposed by [89]. This non-symmetrical hybrid transformer segmentation model demonstrated the highest performance in our baseline tests, as shown in Table 2, achieving a test DSC of 0.6520 and a test IoU of 0.5421. The harmonic element of the network design that is used for the naming of the network is derived from the harmonic pattern of the number of layers used in each HarDNet convolution block. In the encoder, HarDNet performs channel splitting on the convolutional outputs in accordance to the number of output connections per layer. This results in an input channel count equal to the number of output channels for each 3×3 convolutional layer. The decoder

Table 4

Summary of the composition of the two dark skin tone test sets used in our experiments. Test set A = 342 images and corresponding masks taken from the DFUC 2022, AZH, and FUSC datasets; test set B = 342 images (with no masks) taken from the Alzubaidi, Fitzpatrick17k, FUSC, GIS-W, Medetec, Wseg, and KSUMC datasets.

Dataset name	Images	Masks	Test set
DFUC 2022	68	68	A
AZH	81	81	A
CWDB	3	3	A
FUSC	190	190	A
Alzubaidi	52	0	B
Fitzpatrick17k	4	0	B
FUSC	35	0	B
GIS-W	13	0	B
Medetec	8	0	B
Wseg	115	0	B
KSUMC	115	0	B
Total	684	342	A & B

implements a series of Lawin (Large Window Attention) Transformers. Multi-scale features are captured using a Multi-Layer Perception (MLP) decoder and an MLP-Mixer together with Spatial Pyramid Pooling (SPP). The MLP-Mixer comprises two layer types: one with MLPs independently applied to image patches for the purpose of mixing per-location features, and a second using MLPs which are applied across patches to enable spatial information to be mixed to enhance spatial representations, as originally proposed by [93]. SPP is a pooling layer with no fixed-size constraints whereby spatial information is retained in local spatial bins where the outputs of each filter are pooled, allowing for multi-scale representations of features [94]. The decoder design essentially allows for capture of richer contextual data at different scales, utilising transformer elements (Lawin) to focus on improved learning of global relationships. Deep supervision is employed in the decoder to aid regularisation in feature learning and to improve convergence behaviour. This involves the use of companion losses which are calculated at different layers in the network, with the final loss calculated as the output loss plus the sum of the companion losses [95]. Edge loss is also used to enhance the fine-grained details at the edges of prediction masks. Finally, an Exponential Moving Average (EMA) function is used during training which maintains moving averages of trainable parameters via an exponential decay. [96] demonstrated that EMA models generalised better and had improved robustness to noisy labels.

4.5. HardNet-CWS architecture

We propose a modified HardNet-DFUS network architecture, henceforth “HardNet-CWS” (Chronic Wound Segmentation), which utilises the following novel enhancements:

1. Implementation of an improved multi-colour space tensor merging process that builds on concepts proposed in our previous recent works.
2. Modification of the network encoder stem layers using combined instance-batch normalisation in the first encoder block, and switch normalisation in the second encoder block.
3. Replacement of ReLU6 activation functions with Parameterised Rectified Linear Unit (PReLU) activation functions in all convolution blocks in the encoder.
4. Reshaping of the harmonic structure of the HardNet dense convolution blocks to facilitate the additional colour tensor information.

An overview of the complete training workflow for the proposed HardNet-CWS architecture is shown in Fig. 3. An overview of the proposed HardNet-CWS model architecture is shown in Fig. 4. Each of our proposed enhancements are detailed in the following subsections.

4.5.1. HardNet experimental setup

All experiments completed in the following sections used wound images and masks at 640×480 pixels. All models were trained for 100 epochs with a batch size of 5 using the AdamW optimiser with a learning rate of 0.00001, an epsilon of 0.0000001, and a weight decay of 0.01. The hardware and software configuration used for all experiments is the same as those used for the initial baseline experiments.

4.5.2. Multi-colour space tensor merging

The first adjustment to our proposed HardNet-CWS model architecture facilitates the range of additional features found in different non-RGB colour spaces. Traditionally, deep learning models that use colour medical photographs are trained and tested using images in the RGB colour space. However, recent preliminary research conducted by [56] demonstrated that combining individual colour channels from various colour spaces resulted in improved model performance on a range of chronic wound segmentation test sets. Their highest improvement was demonstrated when using the FUSC validation set as an exclusive test set, achieving increases in IoU (+0.0264) and DSC (+0.0348) when merging RGB colour channels with the Y (luminance) channel from the YCrCb colour space to form a new merged multi-channel tensor (RGB+Y). This work demonstrated that merging individual channels from non-RGB colour spaces resulted in higher performance gains when compared to merging whole colour spaces together. However, a limitation of this work is that it was only demonstrated using a simple U-Net architecture [97]. In this work, we experiment with the colour space channels that demonstrated the highest performance improvements in the prior studies completed by [56]. We complete experiments that utilise the merging of different colour channel tensors from the RGB, YCbCr, and CIELAB colour spaces. Based on the prior results from the experiments conducted by [56], we experiment by merging RGB with the Y luminance channel from the YCbCr colour space, and the ‘A’ chromaticity channel from the CIELAB colour space. We also propose an alternative representation of luminance, which we refer to as exaggerated luminance (eY), which is derived from the RGB colour space.

For the experiments which focus on merging RGB with the Y and A channels, a summary of results is shown in Table 5. Algorithm 1 shows the process of merging the RGB channels with the Y and A channels to form newly merged tensors. In terms of test results, the RGB+A, RGB+Y, and RGB+Y+A experiments all show improvements over the baseline results, with the RGB+Y+A experiment demonstrating the highest test set performance increases for test IoU (+0.0180), test DSC (+0.0241), and FNE (−0.0055).

Algorithm 1 RGB+Y+A tensor merging algorithm.

```

1: procedure TENSOR_MERGE(rgb_image)
2:   rgb_tensor  $\leftarrow$  to_tensor(rgb_image)
3:   lab  $\leftarrow$  convert_rgb_to_lab(rgb_tensor)
4:   a  $\leftarrow$  split(lab)[1]
5:   ycrcb  $\leftarrow$  convert_rgb_to_ycrcb(rgb_tensor)
6:   y  $\leftarrow$  split(ycrcb)[0]
7:   image  $\leftarrow$  merge([rgb_tensor, y, a])
8:   Return image
9: end procedure

```

To build on the prior tensor merging work completed by [56], we experiment further with the Y channel in the tensor merging operation. Our approach was to increase the difference between lighter and darker values in the Y channel by first normalising then applying a fixed exponential. We also experimented by switching the R and B coefficients during the conversion process. The process of deriving the eY channel from the RGB colour space and merging the corresponding tensors is described in Algorithm 2. For all experiments which utilise eY, we use

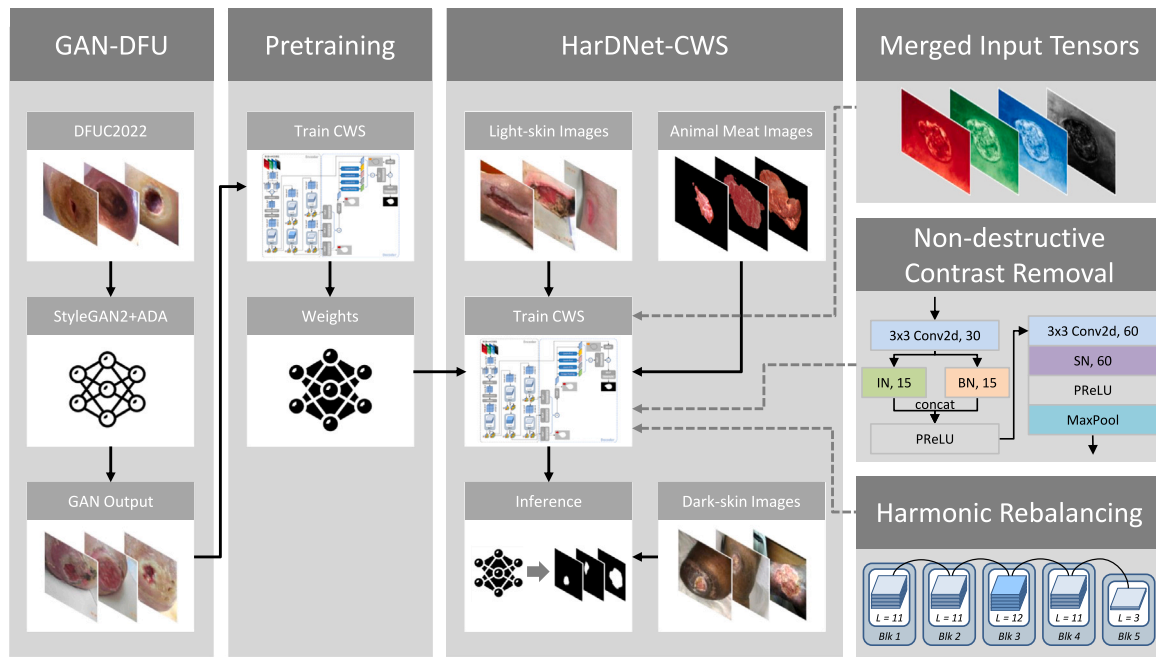


Fig. 3. Illustration showing an overview of the proposed HarDNet-CWS training and testing workflow.

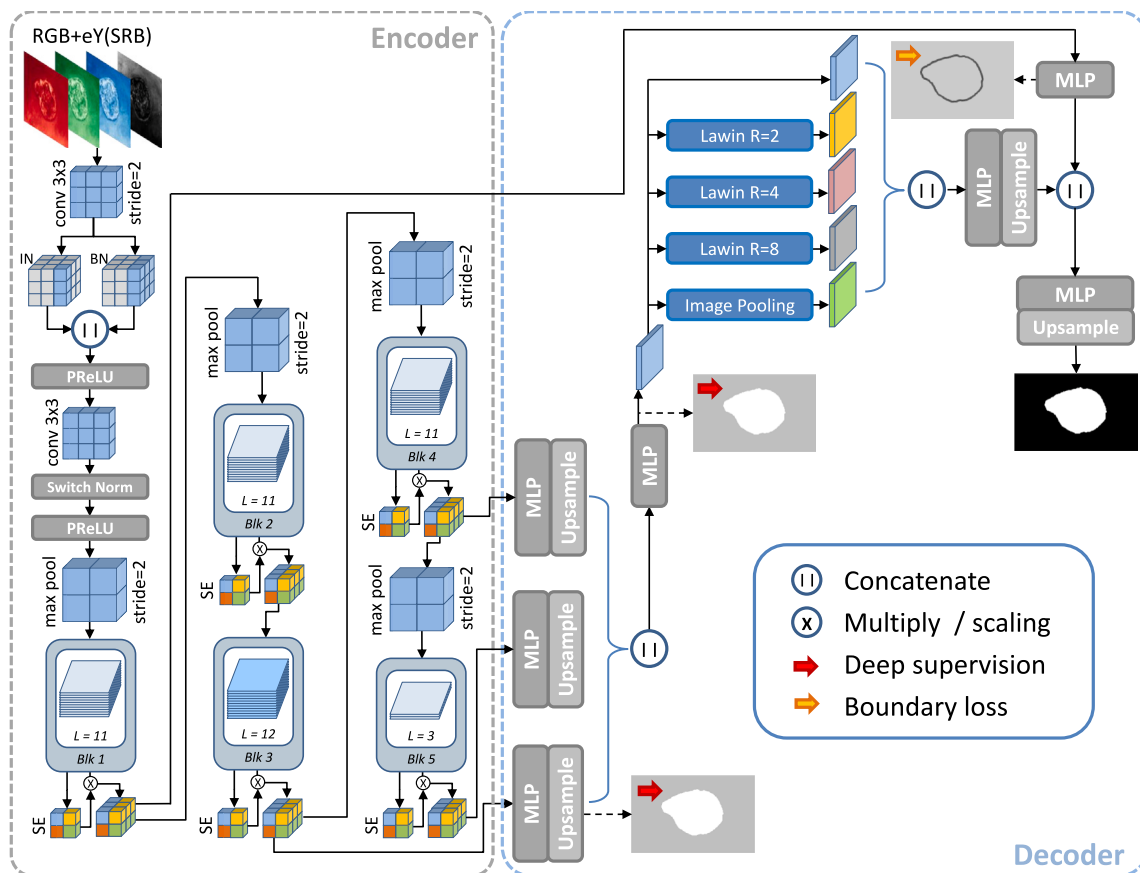


Fig. 4. Illustration of the proposed HarDNet-CWS network architecture. eY - exaggerated luminance, SRB - switched red and blue coefficients, IN - instance normalisation, BN - batch normalisation, SE - squeeze and excite, L - layers, Blk - HarDNet block, MLP - multilayer perceptron, R - patch size.

Table 5

Summary of results for the multi-colour channel tensor merging operations when merging RGB colour channels with ‘A’ chromaticity (from the CIELAB colour space) and luminance (Y channel from the YCbCr colour space).

Colour channels	Best epoch	Train IoU	Train loss	Train DSC	Val IoU	Val loss	Val DSC	Test IoU	Test DSC	FPE	FNE
RGB (baseline)	50	0.9427	0.0975	0.9694	0.6258	0.4063	0.7176	0.5350	0.6389	0.0597	0.3254
RGB+A	19	0.7315	0.3299	0.8340	0.6167	0.3942	0.7066	0.5393	0.6449	0.0421	0.3267
RGB+Y	28	0.8380	0.2301	0.9084	0.6140	0.3777	0.7060	0.5402	0.6515	0.0446	0.3580
RGB+Y+A	33	0.8745	0.1895	0.9310	0.6319	0.3581	0.7229	0.5530	0.6630	0.0508	0.3199

Table 6

Summary of results for the multi-colour channel tensor merging operations when merging RGB colour channels with exaggerated luminance (eY) and ‘A’ chromaticity using normal RGB coefficients (NC) and switched R and B coefficients (SC). Note that when deriving eY from RGB an exponent value of 5 was used for these experiments.

Colour channels	Best epoch	Train IoU	Train loss	Train DSC	Val IoU	Val loss	Val DSC	Test IoU	Test DSC	FPE	FNE
RGB (baseline)	50	0.9427	0.0975	0.9694	0.6258	0.4063	0.7176	0.5350	0.6389	0.0597	0.3254
RGB+eY (NC)	32	0.8670	0.1996	0.9267	0.6224	0.3944	0.7108	0.5422	0.6518	0.0481	0.3245
RGB+eY (SC)	32	0.8585	0.2089	0.9213	0.6232	0.3903	0.7128	0.5576	0.6654	0.0420	0.3119
RGB+eY+A (NC)	28	0.7825	0.2877	0.8719	0.6193	0.4033	0.7094	0.5436	0.6484	0.0452	0.3108
RGB+eY+A (SC)	26	0.8171	0.2554	0.8957	0.6302	0.3580	0.7201	0.5464	0.6544	0.0423	0.3427

the derivation of luminance equation (see Eq. (5)) as defined by the BT.709-4 standard as proposed by [98].

$$Y = 0.2126R + 0.7152G + 0.0722B \quad (5)$$

where R represents the red channel value, G represents the green channel value, and B represents the blue channel value.

Algorithm 2 RGB+eY tensor merging algorithm.

```

1: procedure TENSOR_MERGE( $rgb\_image$ )
2:    $rgb\_tensor \leftarrow to\_tensor(rgb\_image)$ 
3:    $r, g, b \leftarrow split(rgb\_tensor)$ 
4:    $l \leftarrow (r \times 0.0722 + g \times 0.7152 + b \times 0.2126)$ 
5:    $l \leftarrow to\_array(((l) \div max(l)) \times 255)$ 
6:    $ey \leftarrow to\_array((l^5 \div max(l^5)) \times 255)$ 
7:    $image \leftarrow merge([rgb\_tensor, ey])$ 
8:   Return  $image$ 
9: end procedure

```

Table 6 shows the results of the eY experiments, with the results compared to the baseline RGB results. The RGB+eY(SC) test results show a clear improvement over both the baseline test results and the RGB+A, RGB+Y, and RGB+Y+A results shown in Table 5. Compared to the best results from the prior experiments (see Table 5) the RGB+eY tensor merging operation with switched R and B coefficients demonstrate test set performance improvements in terms of test IoU (+0.0174), test DSC (+0.0139), FPE (−0.0026), and FNE (−0.0461).

Table 7 shows results for obtaining the optimum exponent value in the RGB+eY switched coefficient experiments. We initially selected an exponent value of 5, then experimented with values of 4 and 6. The results indicate that an exponent value of 5 provides the optimum exponent value. Fig. 5 shows two wound images from test set B for the luminance channel and the two alternate representations (eY and eY with swapped R and B coefficients). These images show a notable change in contrast between wound and non-wound regions. To the human eye, there is little discernible difference between eY and eYS-R&B, although as shown in our results (see Table 6) the latter offers test performance improvements over the former. The “Difference” column shows the difference between the eY and eYS-R&B channels, which indicates a dense concentration of features within the wound regions. The “Difference” images were produced using the absdiff function in the Python CV2 library [99].

4.5.3. Combined instance-batch normalisation

The second of our modifications utilises a combined instance and batch normalisation (IBN) layer in the first convolution block of the encoder path. The IBN activation layer improves the ability of the

encoder to extract features where contrast is still a prominent feature, found predominantly in the early layers of the encoder. In isolation, instance normalisation reduces contrast features but also reduces useful information, while batch normalisation allows for more of those features to be retained [100]. The procedure for creating an IBN layer is detailed in Algorithm 3.

Algorithm 3 Instance-batch normalisation algorithm.

```

1: procedure IBN( $channels$ )
2:    $ratio \leftarrow 0.5$ 
3:    $half \leftarrow (channels \times ratio)$ 
4:    $in \leftarrow instance\_norm(half)$ 
5:    $bn \leftarrow batch\_norm(channels - half)$ 
6:    $out \leftarrow concatenate(in, bn)$ 
7:   Return  $out$ 
8: end procedure

```

We experimented with IBN by gradually adding it to each successive convolution block in the encoder until performance started to degrade. Combining instance normalisation with batch normalisation ensures that the benefits of instance normalisation (removing contrast information [101]), are not lost while also benefiting from the effect of batch normalisation, which reduces internal covariate shift, stabilising training by reducing overfitting and improving model generalisation [102]. The integration of batch normalisation ensures that the instance normalisation component does not remove more than the contrast features. This modification to HarDNet-DFUS is inspired by the work of [100]. They demonstrated the effect of combining instance and batch normalisation in object classification and non-medical segmentation tasks. To the best of our knowledge, the use of IBN in our proposed HarDNet-CWS architecture is the first time that the method has been demonstrated in any deep learning chronic wound study.

4.5.4. Parameterised rectified linear unit

The third adjustment we make to the HarDNet-DFUS architecture is the replacement of Rectified Linear Unit (ReLU) activation layers in the encoder convolution blocks with Parametric ReLU (PReLU) activation layers. PReLU is an advanced variation of prior ReLU activation functions (ReLU and Leaky ReLU) that has been shown to improve model fitting [103]. PReLU can be used in training scenarios using backpropagation and can be optimised concurrently with other network layers. Leaky ReLU multiplies negative inputs by a nominal value, e.g. 0.022. PReLU improves on this aspect by making the nominal negative value learnable during training, allowing it to adapt more to weight and bias parameters. The mathematical expression for PReLU is shown in Eq. (6). Conditionally, if $a_i = 0$, then f becomes a ReLU

Table 7

Summary of results for the multi-colour channel tensor merging operations when merging RGB colour channels with exaggerated luminance (eY) for switched R and B coefficients using different exponent (EX) values.

Colour channels	Best epoch	Train IoU	Train loss	Train DSC	Val IoU	Val loss	Val DSC	Test IoU	Test DSC	FPE	FNE
RGB (baseline)	50	0.9427	0.0975	0.9694	0.6258	0.4063	0.7176	0.5350	0.6389	0.0597	0.3254
RGB+eY (EX = 4)	21	0.7383	0.3302	0.8405	0.6302	0.3649	0.7216	0.5427	0.6468	0.0458	0.3190
RGB+eY (EX = 5)	32	0.8585	0.2089	0.9213	0.6232	0.3903	0.7128	0.5576	0.6654	0.0420	0.3119
RGB+eY (EX = 6)	31	0.8616	0.2041	0.9234	0.6288	0.3750	0.7184	0.5559	0.6630	0.0449	0.3387

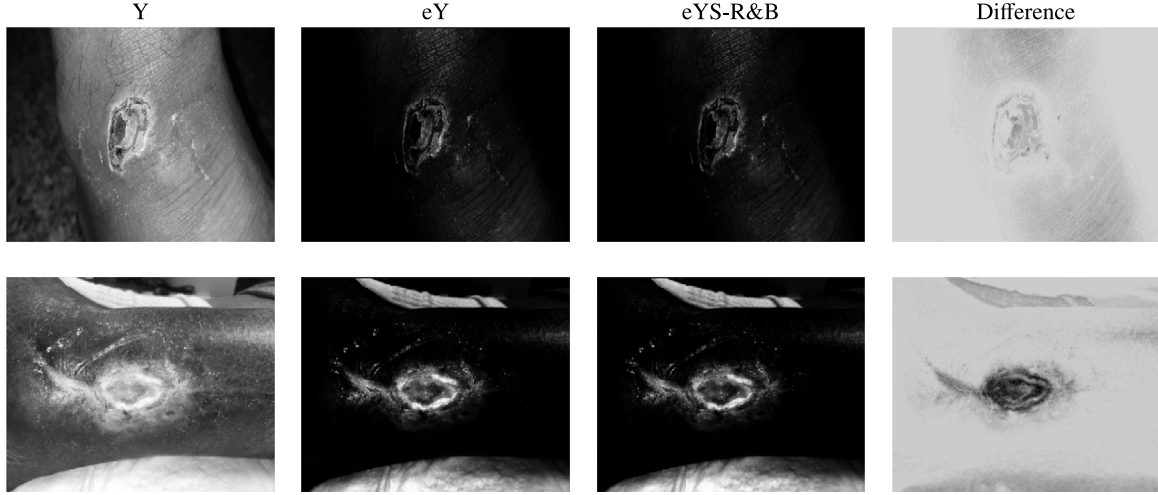


Fig. 5. Illustration of 2 cases from test set B showing the Y channel and its alternate representations. Y - luminance, eY - exaggerated luminance, eYS-R&B - exaggerated luminance with swapped R and B coefficients. Note that the Difference images show the difference in features between the eY and eYS-R&B images. The first row image is from the Alzubaidi dataset, and the second row image is from the FUSC dataset.

activation. If $a_i > 0$, then f becomes a leaky ReLU activation. If a_i is learnable, then f becomes a PReLU activation.

$$f(y_i) = \begin{cases} y_i, & \text{if } y_i > 0 \\ a_i y_i, & \text{if } y_i \leq 0 \end{cases} \quad (6)$$

where y_i is an input for the i th channel, and a_i is the learnable parameter (negative slope).

4.5.5. Switchable-normalisation

To further enhance the encoder in our proposed network architecture, we implement a Switchable-Normalisation (SN) layer in the second encoder block. As with the previous experiments using IBN, we introduced SN into all layers of the encoder and gradually removed each layer, starting from the last layer, until the optimum performance was reached. SN, originally proposed by [104], selectively learns different normalisers by using channel, layer, and minibatch values to compute means and variance statistics. SN is able to adapt to various network architecture designs, is robust to a range of batch sizes, and is not prone to hyper-parameter sensitivity as exhibited by other normalisation methods such as group normalisation. SN inherits all the benefits of instance norm, layer norm, and batch norm by learning their importance ratios during training, preventing overfitting by balancing between generalisation and feature learning. The switchable-normalisation process is summarised in Eq. (7).

$$\Phi = \{\lambda_{in}, \lambda_{ln}, \lambda_{bn}, \lambda'_{in}, \lambda'_{ln}, \lambda'_{bn}\} \quad (7)$$

where Φ is a set of learnable parameters, in represents instance normalisation, ln represents layer normalisation, and bn represents batch normalisation.

Fig. 6 shows the original block design for the stem layers of the HarDNet-DFUS encoder together with our proposed adjustments implementing IBN, PReLU, and SN.

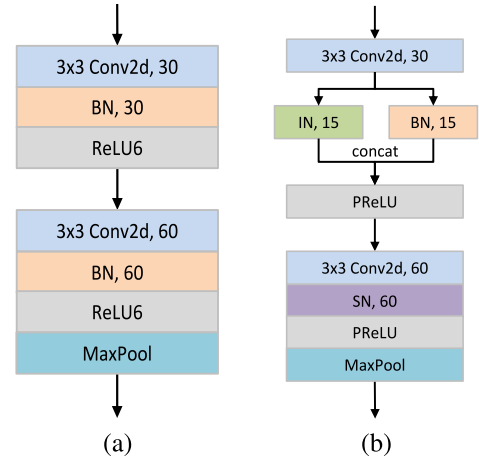


Fig. 6. Illustration of (a) the original HarDNet-DFUS convolutional block design found in the encoder stem, and (b) our enhanced block design utilising instance-batch normalisation, PReLU activation, and Switchable Normalisation. BN - batch normalisation, ReLU - rectified linear unit, IN - instance normalisation, PReLU - parametric rectified linear unit, SN - switchable normalisation.

4.5.6. Refined HarDNet block harmonic structure

The fourth refinement to our proposed HarDNet-CWS model architecture involves the adjustment of the harmonic shape found in the HarDNet convolution encoder blocks. The original block design is a pattern of increasing and decreasing sequence of convolution layers represented by each HarDNet block. In our proposed adjustment to the HarDNet blocks, we change the harmonic pattern such that the minimum and maximum layer amplitude values for the first four blocks are less pronounced. For the first four HarDNet blocks the number of layers in blocks with lower layer counts are increased, while the blocks with higher layer counts are reduced, creating a smoother harmonic

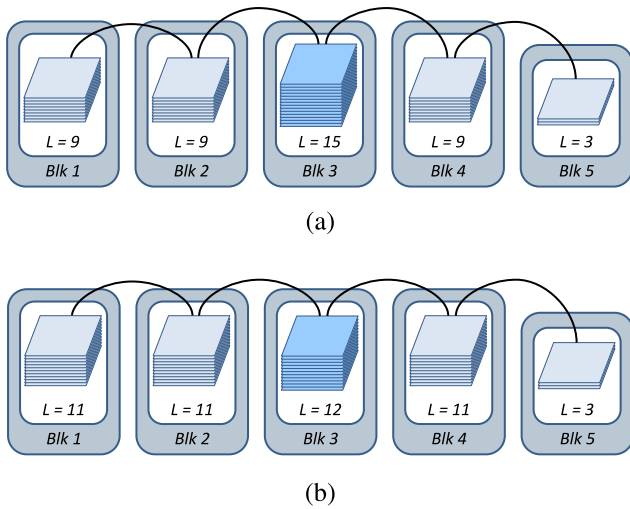


Fig. 7. Illustration of (a) the original HarDNet-DFUS harmonic block design, and (b) our proposed HarDNet-CWS harmonic block design which increases the density of the lower density blocks, and reduces the density of the higher density blocks which results in a reduction of the overall harmonic amplitude. L - number of layers in HarDNet convolution block, Blk - block number.

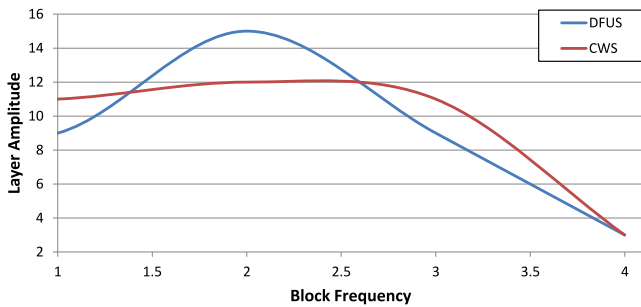


Fig. 8. Illustration comparing the waveform representations of the HarDNet blocks for the original HarDNet-DFUS architecture (blue), and our proposed HarDNet-CWS architecture (red).

pattern. This also results in an overall increase in distributed layers to facilitate the supplemental features captured from the additional eY channel tensors. Fig. 7 shows the original harmonic block design (a), and our improved harmonic block design (b). Fig. 8 shows a comparison of the block and layer patterns expressed as waveforms for the original HarDNet-DFUS and our proposed HarDNet-CWS architecture. Our experimental results indicated that the network architecture responds more to lower variations in layer counts for each HarDNet block in the encoder when trained, validated, and tested on our wound datasets. The layer amplitude for HarDNet-DFUS has a $sd = 4.2427$, while our proposed HarDNet-CWS has a layer amplitude with $sd = 3.7149$.

Table 8 shows a summary of all the proposed network architecture modifications. These results show that the highest performance increase is with the use of the CWS model trained using RGB+eY merged tensors with the proposed PReLU, IBN, SN, and HarDNet block harmonic adjustments. When using RGB+eY merged tensors with the proposed model adjustments, we observe test set performance improvements in terms of test IoU (+0.0144) and test DSC (+0.0141) when compared to using only RGB+eY merged tensors, as shown in the previous experiments.

4.6. GAN-based pretraining

[105] conducted experiments in DFU wound classification with different transfer learning scenarios. They showed that same-domain

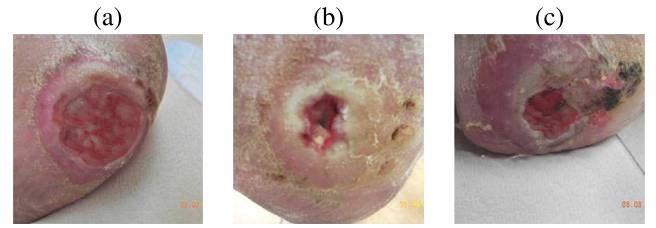


Fig. 9. Illustration of three GAN-generated DFU wounds from the 18,799 GAN-generated wound images that we used for pretraining our proposed HarDNet-CWS model.

transfer learning significantly improved model performance. [106] would later conduct DFU segmentation experiments using 4000 GAN-generated DFU wound images to improve performance of a segmentation model. In this section we experiment with a model trained and validated on a solely synthetic DFU segmentation dataset, which we then use as a pretrained model for training our proposed HarDNet-CWS model. The respective dataset, consisting of 20,000 unconditionally generated and pseudo-labelled DFU images, originates from ground-work conducted by [106] and provided for this study. Two underlying GAN-models were trained on the DFUC 2022 dataset, one on the training set and one on the training and test set. From each, 10,000 images were generated via incrementing seeds and pseudo-labelled as described in the original work. Of these a total of 18,799 samples with at least one DFU instance was selected, and samples not showing any instances were discarded. Fig. 9 shows a selection of images from the included samples, demonstrating the variety of generated representations. For model training we then split the synthetic dataset using an 80:20 ratio into a training set ($n = 15,039$) and a validation set ($n = 3760$). We then trained our best model using this data. Next, we froze the stem layers and the first HarDNet block in our model, and trained again using the trained GAN DFU model as pretrained weights. The results of this experiment are shown in Table 9. When compared to the best performing model from the previous experiments (CWS+[RGB+eY]+[PReLU+IBN+SN+Har]), the results for the test set show clear performance improvements in terms of test IoU (+0.0243), test DSC (+0.0212), FPE (−0.0032), and FNE (−0.0032).

4.7. Cross-domain weakly supervised training using animal meat dataset

In this experiment, we sourced a dataset of 363 animal meat images using Google Image Search with the Creative Commons License search option to remove copyrighted images from search results. The motivation for this experiment derives from the visual appearance of textures present in both cooked and uncooked animal meat, which we identified as being similar to those of human wounds. Given the small size of the animal meat dataset, rather than using pretraining, we include the images directly into the wound training set. Beforehand, we used our current best model to complete inference on the animal meat images and used the resulting prediction masks as ground truth. Table 10 shows the results of the experiments which introduced the animal meat dataset into the training workflow. When compared to the best performing model in the previous experiments (CWS+pre-trained), these results show clear performance improvements for test IoU (+0.0138), test DSC (+0.0154), and FNE (−0.0109). Fig. 10 shows three example masked animal meat images that we used to enhance model performance.

4.8. Augmentation and K-fold cross validation

For the final stage of training our proposed HarDNet-CWS model, we completed a 5-fold cross validation together with training augmentation and test time augmentation (TTA) to enhance model performance.

Table 8

Summary of results for the proposed model architecture improvements for HardNet-CWS. DFUS - HardNet-DFUS, CWS - HardNet-CWS, eY - exaggerated luminance, IBN - instance-batch normalisation, PR - PReLU activation function, SN - switchable normalisation, Har - harmonic block adjustment.

Model	Best epoch	Train IoU	Train loss	Train DSC	Val IoU	Val loss	Val DSC	Test IoU	Test DSC	FPE	FNE
DFUS (baseline)	50	0.9427	0.0975	0.9694	0.6258	0.4063	0.7176	0.5350	0.6389	0.0597	0.3254
CWS RGB+Y+A	33	0.8745	0.1895	0.9310	0.6319	0.3581	0.7229	0.5530	0.6630	0.0508	0.3199
CWS [RGB+Y+A]+[PReLU+IBN+SN+Har]	25	0.7903	0.2769	0.8769	0.6266	0.3572	0.7171	0.5570	0.6645	0.0417	0.3563
CWS RGB+eY	32	0.8585	0.2089	0.9213	0.6232	0.3903	0.7128	0.5576	0.6654	0.0420	0.3119
CWS [RGB+eY]+[PReLU+IBN+SN+Har]	27	0.8241	0.2483	0.9001	0.6193	0.3916	0.7082	0.5720	0.6795	0.0476	0.3132

Table 9

Summary of results showing the performance improvements for the proposed HardNet-CWS model when using the DFU GAN pretrained weights. DFUS - HardNet-DFUS, CWS - HardNet-CWS [RGB+eY]+[PReLU+IBN+SN+Har].

Model	Best epoch	Train IoU	Train loss	Train DSC	Val IoU	Val loss	Val DSC	Test IoU	Test DSC	FPE	FNE
DFUS (baseline)	50	0.9427	0.0975	0.9694	0.6258	0.4063	0.7176	0.5350	0.6389	0.0597	0.3254
CWS	27	0.8241	0.2483	0.9001	0.6193	0.3916	0.7082	0.5720	0.6795	0.0476	0.3132
CWS+pretrained	40	0.9444	0.0961	0.9704	0.6713	0.3391	0.7580	0.5963	0.7007	0.0444	0.3100

Table 10

Summary of results showing the performance improvements when introducing the animal meat dataset into the training process. BEP - best epoch, DFUS - HardNet-DFUS, CWS - HardNet-CWS [RGB+eY]+[PReLU+IBN+SN+Har], AMD - animal meat dataset.

Model	BEP	Train IoU	Train loss	Train DSC	Val IoU	Val loss	Val DSC	Test IoU	Test DSC	FPE	FNE
DFUS (baseline)	50	0.9427	0.0975	0.9694	0.6258	0.4063	0.7176	0.5350	0.6389	0.0597	0.3254
CWS+pretrained	40	0.9444	0.0961	0.9704	0.6713	0.3391	0.7580	0.5963	0.7007	0.0444	0.3100
CWS+pretrained+AMD	52	0.9509	0.0857	0.9738	0.6759	0.3213	0.7660	0.6101	0.7161	0.0456	0.2991

Table 11

Summary of results showing the performance improvements when using 5-fold cross validation (5F) and test time augmentation (TTA). BEP - best epoch, DFUS - HardNet-DFUS, CWS - HardNet-CWS [RGB+eY]+[PReLU+IBN+SN+Har], PT - pretrained, AMD - animal meat dataset.

Model	BEP	Train IoU	Train loss	Train DSC	Val IoU	Val loss	Val DSC	Test IoU	Test DSC	FPE	FNE
DFUS (baseline)	50	0.9427	0.0975	0.9694	0.6258	0.4063	0.7176	0.5350	0.6389	0.0597	0.3254
CWS+PT+AMD	52	0.9509	0.0857	0.9738	0.6759	0.3213	0.7660	0.6101	0.7161	0.0456	0.2991
CWS+PT+AMD+5F	59	0.7561	0.3049	0.8507	0.6822	0.3571	0.7775	0.6460	0.7485	0.0526	0.2672
CWS+PT+AMD+5F+TTA	59	0.7561	0.3049	0.8507	0.6822	0.3571	0.7775	0.6620	0.7610	0.0522	0.2502

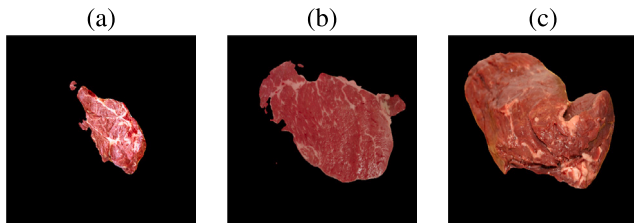


Fig. 10. Illustration of three masked animal meat images used in the weakly supervised training process to enhance performance of our HardNet-CWS model. Prior to training, ground truth masks were generated via inference using our best model.

For the training augmentation, the albumentations library [107] was utilised to generate the following: (1) centre cropping; (2) random cropping; (3) horizontal flipping; (4) vertical flipping; (5) shift scale with rotation; (6) Gaussian noise; (7) random brightness and contrast; (8) contrast limited adaptive histogram equalisation; and (9) multi-scaling. For TTA we employed horizontal and vertical flipping. The training and validation results for these experiments are summarised in Table 11. When compared to the best performing model from the previous experiments (CWS+PT+AMD), these results show clear performance improvements on the test set for the CWS+PT+AMD+5F+TTA model in terms of test IoU (+0.0519), test DSC (+0.0449), and FNE (−0.0489).

4.9. Qualitative analysis

Two clinical wound experts from two different hospitals were recruited, each with more than 10 years clinical experience, to rate the inference predictions from the HardNet-DFUS (baseline) and HardNet-CWS (proposed) models for test sets A and B using a 5-star rating

system. A rating of 1 indicates a poor quality prediction, while a rating of 5 indicates an excellent quality prediction. Raters were asked to not rate a prediction if the model failed to make any prediction where wounds were visible in the image, which is defined as a zero-star rating. If no wounds were present in an image and no prediction had been generated, then raters were asked to rate the prediction with a 5-rating. If more than one wound was present in an image, then the raters were asked to rate the overall quality of all predictions in the image. To reduce possible bias, raters were not informed of which model prediction images came from.

Statistical analysis to ascertain reliability measures taken from two clinical experts who rated the HardNet-DFUS (baseline) and HardNet-CWS (proposed) test results was completed using IBM SPSS version 28.0.1.0 (SPSS Inc., Chicago, Illinois). The analysis of the ordinal data was completed using the intra-class correlation coefficient (ICC) to obtain inter-rater reliability consistency and agreement measures. Consistency is defined as the degree to which the score of a single rater (y) can be equated to a second rater's score (x) plus a systematic error (c) (i.e., $y = x + c$). Agreement concerns the extent to which y is equal to x [108]. A two-way random effects model was used to generalise results to a population of raters from which the clinical expert raters in our study represent a sample. The mathematical expressions for ICC consistency and ICC agreement are shown in Eqs. (8) and (9) respectively.

$$ICC = \frac{MS_R - MS_E}{MS_R} \quad (8)$$

$$ICC = \frac{MS_R - MS_E}{MS_R + \frac{MS_C - MS_E}{n}} \quad (9)$$

where MS_R is the mean square for rows, MS_E is the mean square for error, MS_C is the mean square for columns, and n is the number of subjects.

Table 12

Test results for the HardNet-DFUS (baseline) and HardNet-CWS (proposed) models for test set A dark skin tone wound images that have ground truth masks.

Model	IoU	DSC	FPE	FNE
HardNet-DFUS	0.5350	0.6389	0.0597	0.3254
HardNet-CWS	0.6624	0.7610	0.0522	0.2502

ICC values are interpreted as follows: 0–0.39 indicates poor reliability; 0.4–0.74 indicates moderate reliability; 0.75–1 indicates excellent reliability [109].

5. Results

In this section we report on the results of inference using our proposed HardNet-CWS model. We present the results for two test sets: test set A which comprises 342 dark skin tone wound images and corresponding masks taken from the DFUC 2022, AZH, CWDB, and FUSC datasets; and test set B which comprises 342 dark skin tone wound images with no masks taken from the Alzubaidi, Fitzpatrick17k, FUSC, GIS-W, Medetec, Wseg, and KSUMC datasets. The test set A predictions were assessed quantitatively and qualitatively, and the test set B results were assessed qualitatively only as this test set has no ground truth masks.

5.1. Quantitative results for test set A

Test metrics for test set A inference results for the HardNet-DFUS (baseline) and HardNet-CWS (proposed) models are summarised in [Table 12](#). We observe significant improvements in terms of IoU (+0.1270), DSC (+0.1221), and FNE (−0.0752), while FPE demonstrated a more subtle improvement (−0.0075). [Fig. 11](#) shows a selection of predictions from test set A demonstrating clear improvements in segmentation performance when comparing the baseline results from the HardNet-DFUS model with the results from the proposed HardNet-CWS model. The first row shows a DFU wound on a foot exhibiting partial amputation, and shows that skin which has been miss-detected along the side of the toe with the DFUS model has not been inaccurately detected by the CWS model. This DFUS miss-detection may have been due to the darker skin on the toe, compared to the skin on the rest of the foot, which the model may have partly miss-detected as necrotic tissue. The second row shows a PRU wound on the lower-back of the torso where the CWS model has more accurately detected the edge details of the wound when compared to the DFUS prediction. This may be a result of the additional features provided by the enhanced tensor inputs in the CWS model, allowing the edge loss function to more accurately define wound boundary details. The third row shows a DFU wound on the ankle where the DFUS model prediction is more generalised and includes a significant region of miss-detected skin, and is much less accurate when compared to the CWS prediction.

5.2. Qualitative results for test sets A and B

Qualitative measures for test set A and B inference results from the HardNet-DFUS (baseline) model and HardNet-CWS (proposed) model are shown in [Table 13](#). The ICC confidence and agreement values for the HardNet-DFUS test set A predictions (confidence ICC = 0.6714, agreement ICC = 0.6717) indicate moderate reliability for the clinical ratings for this model. The ICC confidence and agreement values for the HardNet-DFUS predictions for test set B (confidence ICC = 0.7907, agreement ICC = 0.7747) indicate excellent reliability for the clinical ratings for this model. The ICC confidence and agreement values for the HardNet-CWS test set A predictions (confidence ICC = 0.6633, agreement ICC = 0.6631) indicate moderate reliability. The ICC confidence and agreement values for the HardNet-CWS predictions for test set B (confidence ICC = 0.5001, agreement ICC = 0.4992) indicate

Table 13

Measures derived from expert rater quality assessment of test sets A and B inference results for the HardNet-DFUS (baseline) and HardNet-CWS (proposed) model. ICC - intra-class correlation coefficient, Co - consistency, Ag - agreement, LB - lower bound, UB - upper bound, CI - confidence interval.

Test set	Seg model	Type	ICC	LB95%CI	UB95%CI
A	DFUS	Co	0.6714	0.5935	0.7343
A	DFUS	Ag	0.6717	0.5940	0.7346
B	DFUS	Co	0.7907	0.7411	0.8308
B	DFUS	Ag	0.7749	0.6986	0.8287
A	CWS	Co	0.6633	0.5835	0.7278
A	CWS	Ag	0.6631	0.5834	0.7276
B	CWS	Co	0.5001	0.3817	0.5959
B	CWS	Ag	0.4992	0.3809	0.5949

Table 14

Summary of percentage improvements in terms of 5 star ratings (a) and 4–5 star ratings (b) for the HardNet-CWS (proposed) model when compared to the HardNet-DFUS (baseline) model.

Test set	Rater	DFUS rating	CWS rating	Improvement %
A	1	90.06% ^a	93.57% ^a	3.51%
A	2	89.47% ^a	92.98% ^a	3.51%
B	1	70.47% ^a	90.94% ^a	20.47%
B	2	88.30% ^a	92.04% ^a	3.74%
A	1	92.98% ^b	96.49% ^b	3.51%
A	2	92.39% ^b	96.20% ^b	3.81%
B	1	86.84% ^b	96.79% ^b	9.95%
B	2	91.81% ^b	95.87% ^b	4.06%

moderate reliability. Overall, the ICC reliability measures for the DFUS (baseline) model predictions indicate moderate to excellent reliability, while moderate reliability is demonstrated for the CWS (proposed) model. For the CWS ICC test set A reliability measures, 311 ratings exactly matched, while 19 ratings varied by 1. For the CWS ICC test set B reliability measures, 308 ratings exactly matched, while 20 ratings varied by 1. These results indicate that the majority of ratings between raters matched exactly, or had a difference of no more than 1. [Fig. 12](#) shows a representative rating for each of the 0–5 star ratings for test sets A and B.

To provide further insights into the clinician prediction ratings, we conducted a relative distribution analysis. A summary of the distribution analysis for the DFUS predictions is shown in [Fig. 13](#). These results indicate that for the DFUS (baseline) results, both raters consistently rated the predictions highly, within the 4–5 star range: test set A for rater 1 = 92.98%, test set A for rater 2 = 92.39%, test set B for rater 1 = 86.84%, test set B for rater 2 = 91.81%, test sets A and B for rater 1 = 89.91%, and test sets A and B for rater 2 = 92.11%.

The distribution analysis for the CWS predictions is shown in [Fig. 14](#). These results indicate that for the CWS (proposed) model, both raters consistently rated the predictions highly, within the 4–5 star range: test set A for rater 1 = 96.49%, test set A for rater 2 = 96.20%, test set B for rater 1 = 96.79%, test set B for rater 2 = 95.87%, test sets A and B for rater 1 = 96.64%, and test sets A and B for rater 2 = 95.62%.

We observe that for both test sets and both raters, the CWS (proposed) predictions demonstrated higher scores than the DFUS (baseline) predictions in terms of expert qualitative assessment. A summary of the improvements demonstrated by the CWS (proposed) model based on expert qualitative assessment is shown in [Table 14](#) for 5 star ratings and for 4–5 star ratings. We observe that the number of 5 star ratings for rater 1 on test set B is significantly lower than the other 5 star ratings for this model. However, the difference is much less pronounced when taking into account the 4–5 star ratings, meaning that the discrepancy is mostly due to a difference of 1 star between raters.

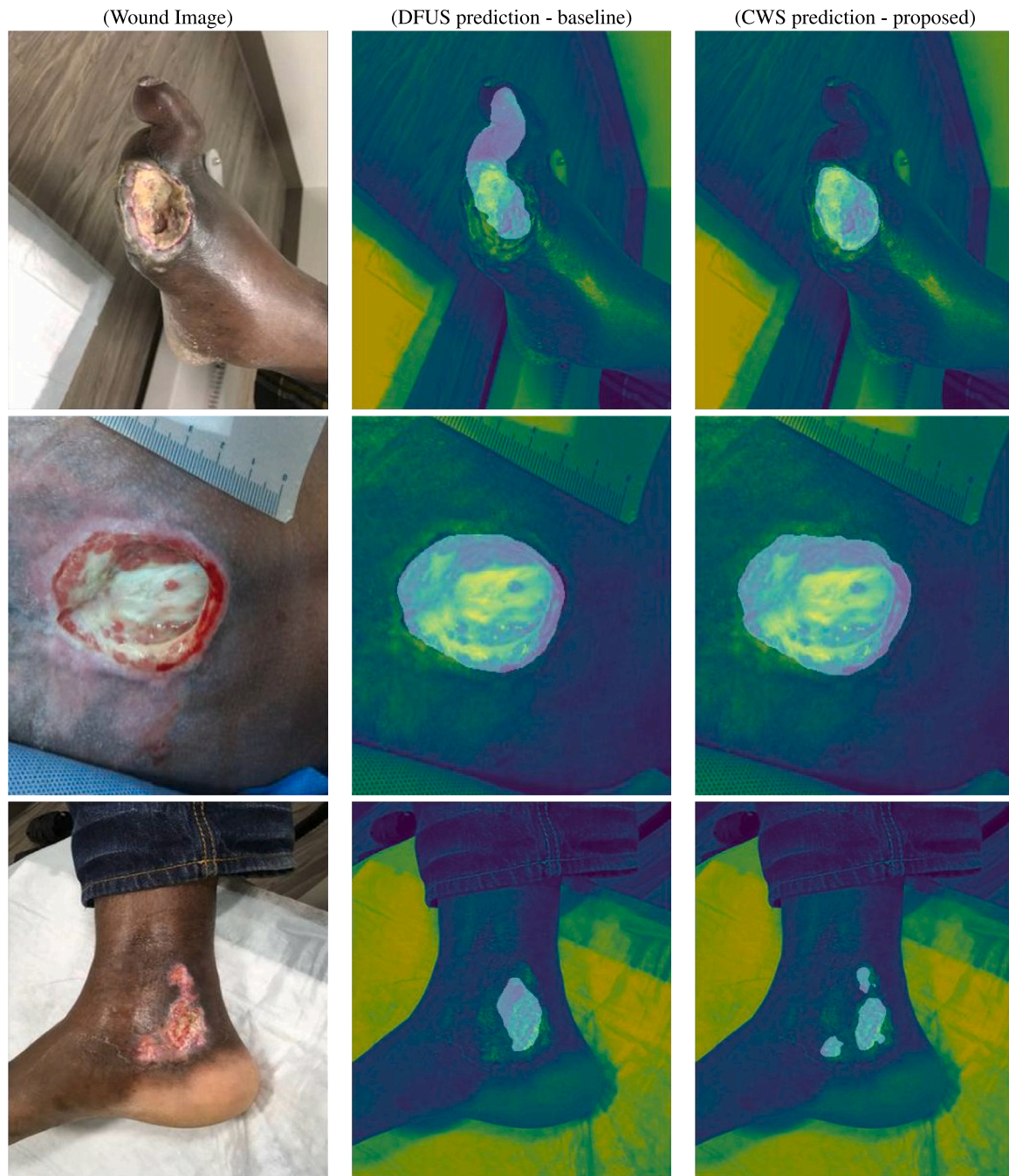


Fig. 11. Illustration of a selection of wound segmentation predictions from test set A for the HarDNet-DFUS (baseline) and HarDNet-CWS (proposed) models. The first row shows a DFU wound on a foot exhibiting partial amputation, the second row shows a PRU wound on the lower back of the torso, and the third row shows a DFU wound on the ankle. The first and third row images are from the FUSC dataset, and the second row image is from the CWDB dataset. Note that images have been cropped for illustrative purposes.

5.3. Test set images with blank masks

During testing with test set A, we observed a number of cases where the ground truth masks comprised only of black pixels, indicating that there were no wound regions present in the corresponding images. However, qualitative results obtained from clinicians showed that some of these cases had in fact been labelled incorrectly. We identified 14 cases in test set A that were sourced from the AZH ($n = 4$), FUSC ($n = 9$), and DFUC 2022 ($n = 1$) datasets where wounds were clearly present in the images, but the corresponding masks comprised of only black pixels. The total number of incorrectly labelled blank masks represents $\approx 4\%$ of the test set total (342 images/masks), indicating that the reported metrics in Tables 5 to 12 are likely to be under-estimates.

6. Discussion

This work focuses primarily on subjective measures derived from expert assessment of model predictions - a facet which is absent from almost all chronic wound deep learning research. Our experiment results indicate significant disparities between the quantitative lab based results and the qualitative results obtained from clinical expert ratings for both baseline (HarDNet-DFUS) and proposed (HarDNet-CWS) models. However, the results for our proposed HarDNet-CWS model show clear performance improvements in terms of lab based metrics and expert qualitative assessment.

The reliability measures obtained from both clinical expert raters for test sets A and B indicate that reliability is moderate to excellent for the

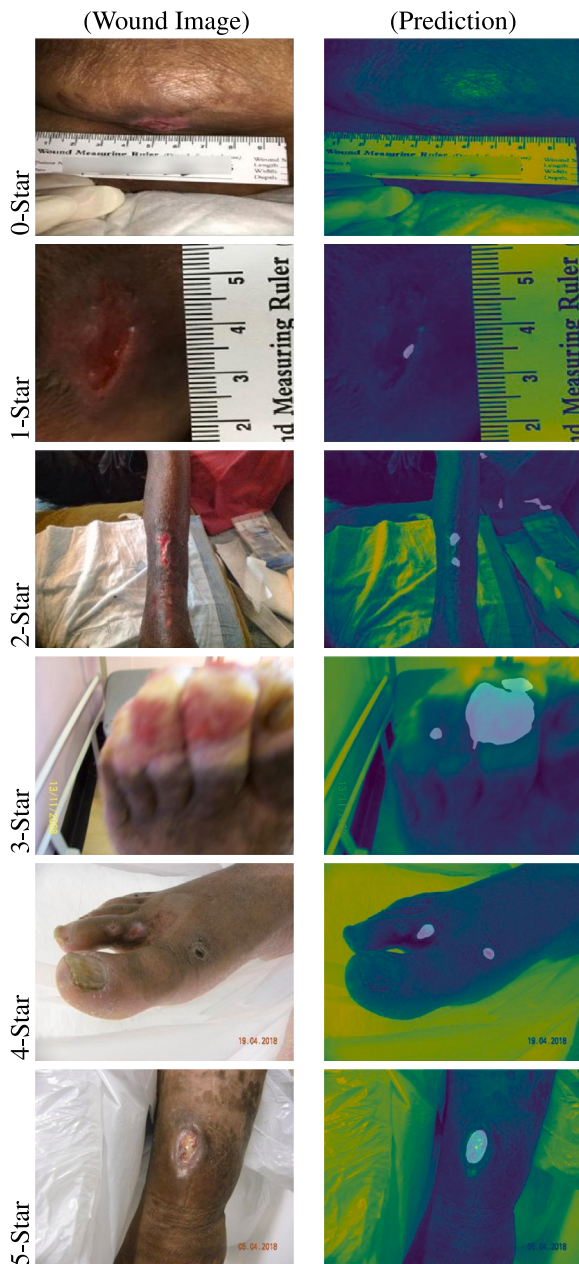


Fig. 12. Illustration of a representative expert clinical rating for each of the 0-5 star ratings for test sets A and B. The 0-star example shows a small wound that was not detected. The 1-star example shows that only a very small region of the wound has been detected. The 2-star example shows that although most of the wound has been detected, non-wound regions have also been miss-detected. The 3-star example shows that more than half of the wound regions have been accurately detected, however, there are still significant wound regions that have not been detected. The 4-star example shows that 2 of the 3 wound regions have been accurately detected, although the smallest wound region on the left side has not been detected. The 5-star example shows a highly accurate prediction of a wound region.

baseline model, and is moderate for the proposed model. However, a further analysis of these results shows that for the proposed model, 311 of 342 5 star ratings matched between raters for test set A, and 308 of 342 5 star ratings matched between raters for test set B. Further, 19 ratings for test set A varied by only 1 star, and for test set B 20 ratings varied by only 1 star. For test set A, a total of 330 of 342 ratings either matched or differed by only 1 star, and for test set B a total of 328 of 342 ratings matched or differed by only 1 star. We therefore suggest that when taking into account that the majority of

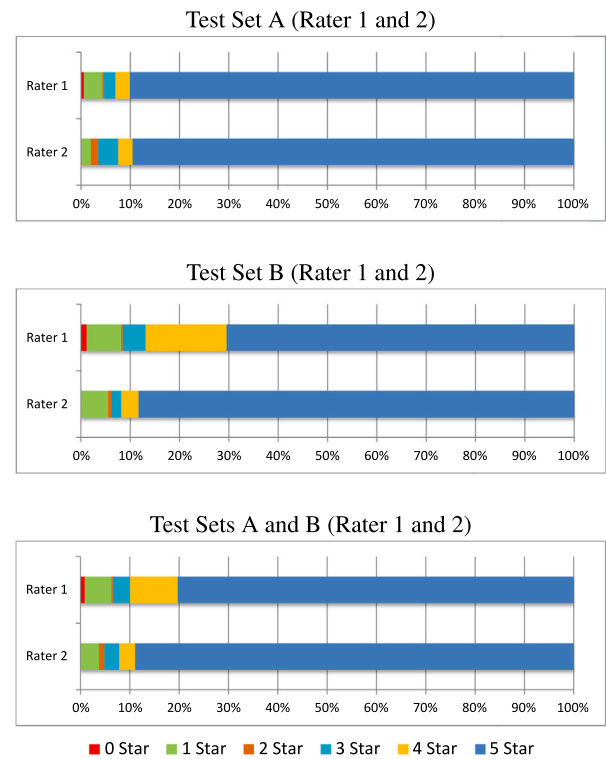


Fig. 13. Relative distribution of clinical ratings for test sets A and B DFUS (baseline) model predictions.

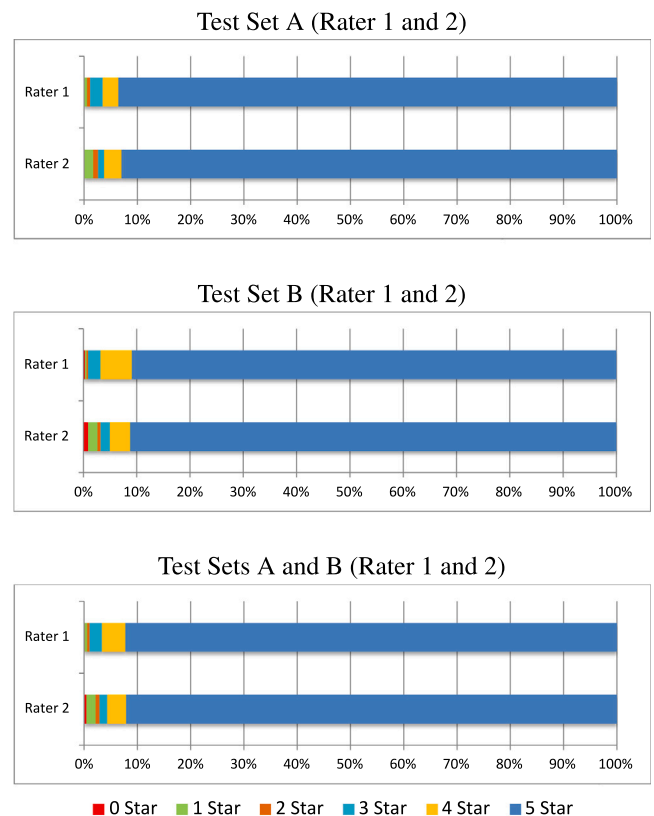


Fig. 14. Relative distribution of clinical ratings for test sets A and B CWS (proposed) model predictions.

expert ratings (>95%) either matched or differed by only 1 star, these results should be considered to demonstrate generally excellent levels of agreement.

Our proposed model was trained and validated on chronic wound images taken from patients with lighter skin, while the two test sets comprised only wound images acquired from patients with darker skin tones. We observe that the validation results for our best performing model on test set A (CWS+PT+AMD+5F+TTA - see Table 11) are marginally higher when compared to the IoU and DSC test results: +0.0202 val IoU compared to test IoU, +0.0165 val DSC compared to test DSC. These results may be evidence that models trained only on lighter skin wound images may find inference challenging on darker skin wound images. However, in the absence of qualitative comparisons between the validation and test inference results, and taking into account the significant disparity between the lab based metrics and the expert qualitative results, we suggest that the differences in validation (lighter skin) and test (darker skin) results may not provide a complete assessment of the model's true ability.

A limitation of this work is that the lab based metrics are assessed on a more fine-grained continuous scale (0–0.1), while the qualitative measures are measured on a 0–5 star ordinal scale. Future work might focus on a more fine-grained approach to qualitative measures, although we suggest that our results give a good general indication of the qualitative aspects of model predictions.

The colour aspects of deep learning research involving the use of medical colour imaging is relatively under-explored. Colour imaging provides an enhanced visualisation of dermatological surface and sub-surface structures which present novel challenges. This is especially pertinent in the deep learning domain, as most methods focus on single-channel images, which are generally less applicable to multi-colour channel domains [110]. In this paper, we make an attempt to direct focus on this aspect with the use of manipulated multi-colour space tensors and a corresponding modified hybrid transformer network architecture that facilitates the additional colour information. Our experiments seem to indicate that there may be additional features in different colour spaces, which the model is able to learn from when such colour space data is merged into single tensors. Our future work will continue to explore the colour aspects of medical wound photographs when training deep learning models.

Our results indicate that there may be a limited capacity for lab-based accuracy metrics when using the current publicly available datasets. We posit that this is largely due to variability in segmentation labelling. This is especially pertinent in the case of chronic wound labelling, which has been shown to be highly variable and subjective [111]. The observed disparity between DSC/IoU and expert subjective ratings for model predictions in our study indicate that the lab-based metrics are only providing part of the picture in deep learning assessment.

Recent studies, such as those conducted by [112], have highlighted a disparity in laboratory results obtained from deep learning models and results obtained in real-world scenarios. To address this issue, our study has an increased emphasis on presenting results from a qualitative analysis of the model predictions obtained in our wound segmentation experiments. The measures derived from our qualitative analysis clearly show that clinician ratings of model predictions are significantly more favourable when compared to the lab-based metrics.

The test sets we used in our main experiments, comprising only darker skin tones, were relatively small compared to most test sets used in deep learning studies. However, this limitation is due to the number of publicly available chronic wound images with ground truth masks, and the limited available time of our clinical collaborators who provided the expert assessment of model predictions. Despite these limitations, the present work presents the most extensive qualitative study so far in chronic wound segmentation.

This work represents the first study to identify that animal meat images can be used to enhance the performance of a chronic wound

segmentation model. Using just 363 animal meat images, with weak supervision, we were able to improve model performance by 0.0141 for test DSC and 0.0144 for test IoU. Animal meat images are significantly easier to obtain than chronic wound images, and require no ethical approval to collect. Furthermore, it may be of interest to experiment with GANs that can generate additional meat images, and to experiment to see how much further such images can be used to boost chronic wound model performance. The number of publicly available chronic wound images with corresponding ground truth segmentation masks is notably limited in deep learning terms (<10,000). If animal meat images can improve model performance further, then this may be a way to at least partly negate the difficult problem of wound image acquisition from medical settings. We strongly encourage other researchers working in chronic wound deep learning studies, especially those working in localisation and segmentation, to experiment with such images.

This work is motivated by the development of new technologies that will allow for the remote detection and monitoring of chronic wounds in home settings. Patients living in remote locations have been shown to have worse outcomes when compared to those living in urban areas. The development of new remote monitoring solutions using deep learning techniques may provide a solution to help reduce such disparities [113]. Such technologies have the potential to reduce the number of patient hospital visits, reducing nosocomial infections. The viability of deep learning detection systems within medical settings has been demonstrated for chronic wounds [40]. However, further clinical evaluations are required in larger studies to confirm model effectiveness across a more diverse range of skin tones. Such studies will be vital to identify where shortfalls exist in current segmentation models.

Strategic approaches to preprocessing methods when training deep learning models for chronic wounds have been shown to be highly effective, as per recent work completed by [114]. This work demonstrates the importance of careful targeting of preprocessing methods for different wound types. Our future work will be guided by these methods to attempt to further improve network performance.

Future work will focus on models that utilise multi-modal data which will include additional clinical information collected from patient records. These data will include details of infection, ischemia, neuropathy, and other clinical measures such as patient age, ethnicity, and blood type. Work is currently underway with our clinical collaborators to collect the required patient data. Prior studies in similar research domains have shown that multi-modality in training workflows can assist in improvements to model accuracy [115]. Using patient IDs linked to dataset images will allow us to reduce the number of cases which are currently spread across training and test sets, reducing potential biases. We will also expand our work to investigate instance segmentation of wound and periwound to determine if features from surrounding wound tissue can help to improve segmentation and classification accuracy.

We note that there are currently no established standards for the accepted levels of accuracy in chronic wound localisation and segmentation. In general, IoU thresholds of 0.50 and 0.75 are most commonly used [116]. However, these measures may differ depending on the research domain. The disparities observed in the present study between lab based metrics and qualitative measures highlight this issue further. We propose that future work should investigate the formulation of accuracy and evaluation standards for chronic wounds via an international consortium of clinical and deep learning experts. The clinical labelling of our datasets reveals that labelling amongst clinicians can be highly variable, a problem which occurs frequently in wound image datasets [117]. Establishing internationally agreed standards may help to improve the accuracy of future models. This is especially pertinent at this stage in the evolution of deep learning models trained using chronic wound datasets, whereby the number of publicly available datasets continues to grow.

Our research group is currently in the process of capturing video recordings of chronic wounds in medical settings, which we intend to use in future studies. Videos of wounds, captured at different angles, would allow for the capture of additional spatial data that may be able to improve the accuracy of predictive models and could be especially useful in the automatic assessment of wound healing over time. Short video clips would be straight forward to capture and analyse using the mobile and cloud frameworks developed in our prior wound studies [27,40].

7. Conclusion

In this work we proposed a novel harmonic densely connected hybrid transformer network architecture utilising multi-colour space tensor merging. We conduct the most comprehensive reliability study to date in chronic wound segmentation using 684 cases to obtain inter-rater reliability measures. A total of 13 datasets were used to train and test our proposed segmentation model. Our proposed model demonstrates significant improvements over the baseline model in terms of lab based metrics (+0.1270 for IoU, +0.1221 for DSC) and in terms of expert qualitative assessment (up to 20% when using a 5 star rating method). For the first time, we demonstrate the ability of a model trained only on patients with lighter skin tones to segment wounds on patients with darker skin tones in an effort to address the issue of biases inherent in many chronic wound deep learning studies. We also demonstrate performance improvements using GAN-generated wound images and an animal meat dataset in the training workflow. The aim of our work is to utilise and build upon state-of-the-art advances in the field to address the problem of accurate chronic wound segmentation and to bring these advances closer to the patients who need them most. Improved methods of chronic wound segmentation will allow for the automated monitoring of chronic wounds over time to determine healing status, requiring fewer hospital visits, and improving patient care.

CRediT authorship contribution statement

Bill Cassidy: Writing – review & editing, Writing – original draft, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Christian McBride:** Writing – review & editing, Writing – original draft, Software, Methodology, Formal analysis, Conceptualization. **Connah Kendrick:** Writing – review & editing, Writing – original draft, Supervision, Software. **Neil D. Reeves:** Writing – review & editing, Writing – original draft, Supervision, Project administration. **Joseph M. Pappachan:** Writing – review & editing, Writing – original draft, Supervision, Project administration, Formal analysis, Data curation. **Cornelius J. Fernandez:** Writing – review & editing, Writing – original draft, Validation, Formal analysis. **Elias Chacko:** Writing – review & editing, Writing – original draft, Validation, Formal analysis. **Raphael Brüngel:** Writing – review & editing, Writing – original draft, Software, Methodology, Data curation, Conceptualization. **Christoph M. Friedrich:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Data curation, Conceptualization. **Metib Alotaibi:** Writing – review & editing, Writing – original draft, Project administration, Formal analysis, Data curation. **Abdullah Abdulaziz AlWabel:** Writing – review & editing, Writing – original draft, Project administration, Data curation. **Mohammad Alderwish:** Writing – review & editing, Writing – original draft, Project administration, Formal analysis, Data curation. **Kuan-Ying Lai:** Writing – review & editing, Writing – original draft, Software, Methodology. **Moi Hoon Yap:** Writing – review & editing, Writing – original draft, Validation, Supervision, Resources, Project administration, Formal analysis, Data curation.

Ethical approval and informed consent

The KSUMC dataset used in this study was obtained with ethical approval from King Saud University Medical City, Saudi Arabia (REF: 24/1159/IRB, 1st March 2024). The Diabetic Foot Ulcer Challenge 2022 dataset used in our experiments was obtained with ethical approval from Lancashire Teaching Hospitals NHS Foundation Trust (REF: SE-281, 30th November 2021). All data acquisition methods were performed in accordance with the guidelines and regulations set out by the King Saud University Institutional Review Board, the NHS Research Ethics Committee, and the Health Research Authority (UK). Written informed consent was obtained from all participating patients. The remaining datasets used in this study are publicly available for research purposes and can be requested from their respective owners.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We would like to thank clinicians at the King Saud University Medical City, Saudi Arabia for granting permission to use the KSUMC chronic wound dataset in our experiments. We would also like to thank clinicians at the following UK NHS hospitals for providing valuable clinical feedback: Lancashire Teaching Hospitals NHS Foundation Trust, UK; United Lincolnshire Hospitals NHS Trust, UK; Jersey General Hospital, Jersey. Raphael Brüngel was partially funded by a PhD grant from the University of Applied Sciences and Arts Dortmund, Dortmund, Germany.

References

- [1] J. Moura, J. Rodrigues, M. Gonçalves, C. Amaral, M. Lima, E. Carvalho, Imbalance in T-cell differentiation as a biomarker of chronic diabetic foot ulceration, *Cell. Mol. Immunol.* (2019) 1–2.
- [2] K. Khunti, L.J. Gray, T. Skinner, M.E. Carey, K. Realf, H. Dallosso, H. Fisher, M. Campbell, S. Heller, M.J. Davies, Effectiveness of a diabetes education and self management programme (DESMOND) for people with newly diagnosed type 2 diabetes mellitus: three year follow-up of a cluster randomised controlled trial in primary care, *BMJ* 344 (2012) <http://dx.doi.org/10.1136/bmj.e2333>, URL: <https://www.bmj.com/content/344/bmj.e2333>.
- [3] M.J. Davies, L.J. Gray, D. Ahrabian, M. Carey, A. Farooqi, A. Gray, S. Goldby, S. Hill, K. Jones, J. Leal, K. Realf, T. Skinner, B. Stribling, J. Troughton, T. Yates, K. Khunti, Research highlights the challenges of preventing diabetes with group education sessions, *Diabetes Metab. Horm.* (2017) <http://dx.doi.org/10.3310/signal-000396>, URL: <https://evidence.nihr.ac.uk/alert/research-highlights-the-challenges-of-preventing-diabetes-with-group-education-sessions/>.
- [4] K. Ong, L. Stafford, S. McLaughlin, E. Boyko, S. Vollset, A. Smith, B. Dalton, J. Duprey, J. Cruz, H. Hagins, P. Lindstedt, A. Aali, Y. Habtegiorgis, M. Dagne, M. Abbasian, Z. Abbasi-Kangevari, M. Abbasi-Kangevari, S. Abd Elhafeez, R. Abd-Rabu, T. Vos, Global, regional, and national burden of diabetes from 1990 to 2021, with projections of prevalence to 2050: a systematic analysis for the Global Burden of Disease Study 2021, *Lancet* (2023) [http://dx.doi.org/10.1016/S0140-6736\(23\)01301-6](http://dx.doi.org/10.1016/S0140-6736(23)01301-6).
- [5] S. Klein, A. Gastaldelli, H. Yki-Järvinen, P.E. Scherer, Why does obesity cause diabetes? *Cell Metab.* 34 (1) (2022) 11–20, <http://dx.doi.org/10.1016/j.cmet.2021.12.012>, URL: <https://www.sciencedirect.com/science/article/pii/S1550413121006318>.
- [6] B. Petersen, W. Linde-Zwirble, T.-W. Tan, G. Rothenberg, S. Salgado, J. Bloom, D. Armstrong, Higher rates of all-cause mortality and resource utilization during episodes-of-care for diabetic foot ulceration, *Diabetes Res. Clin. Pract.* (2022) <http://dx.doi.org/10.1016/j.diabres.2021.109182>.
- [7] D.A. Jenkins, S. Mohamed, J.K. Taylor, N. Peek, S.N. van der Veer, Potential prognostic factors for delayed healing of common, non-traumatic skin ulcers: A scoping review, *Int. Wound J.* 16 (3) (2019) 800–812, <http://dx.doi.org/10.1111/iwj.13100>.
- [8] P. Zhang, J. Lu, Y. Jing, S. Tang, D. Zhu, Y. Bi, Global epidemiology of diabetic foot ulceration: a systematic review and meta-analysis, *Ann. Med.* 49 (2) (2017) 106–116, <http://dx.doi.org/10.1080/07853890.2016.1231932>, PMID: 27585063.

- [9] E. Eriksson, P. Liu, G. Schultz, M. Martins-Green, R. Tanaka, D. Weir, L. Gould, D. Armstrong, G. Gibbons, R. Wolcott, O. Olutoye, R. Kirsner, G. Gurtner, Chronic wounds: Treatment consensus, *Wound Repair Regen.* 30 (2022) <http://dx.doi.org/10.1111/wrr.12994>.
- [10] P. Franks, J. Barker, M. Collier, G. Gethin, E. Haesler, A. Jawien, S. Lächli, G. Mosti, S. Probst, C. Weller, Management of patients with venous leg ulcers: Challenges and current best practice, *J. Wound Care* 25 (2016) S1–S67, <http://dx.doi.org/10.12968/jowc.2016.25.Sup6.S1>.
- [11] J. Mader, W. Haas, F. Aberer, B. Boulgaropoulos, P.M. Baumann, M. Pandis, K. Horvath, F. Aziz, G. Köhler, T. Pieber, J. Plank, H. Sourij, Patients with healed diabetic foot ulcer represent a cohort at highest risk for future fatal events, *Sci. Rep.* 9 (2019) <http://dx.doi.org/10.1038/s41598-019-46961-8>.
- [12] X.-f. Xiong, L. Wei, Y. Xiao, Y.-C. Han, J. Yang, H. Zhao, M. Yang, L. Sun, Family history of diabetes is associated with diabetic foot complications in type 2 diabetes, *Sci. Rep.* 10 (2020) 17056, <http://dx.doi.org/10.1038/s41598-020-74071-3>.
- [13] R. Costa, N. Cardoso, R. Procópio, T. Navarro, A. Dardik, L. Cisneros, Diabetic foot ulcer carries high amputation and mortality rates, particularly in the presence of advanced age, peripheral artery disease and anemia, *Diabetes Metab. Syndr.: Clin. Res. Rev.* 11 (2017) <http://dx.doi.org/10.1016/j.dsx.2017.04.008>.
- [14] E. Vainieri, R. Ahluwalia, H. Slim, D. Walton, C. Manu, S. Taori, J. Wilkins, D. Huang, M. Edmonds, H. Rashid, V. Kavarthapu, P. Vas, Outcomes after emergency admission with a diabetic foot attack indicate a high rate of healing and limb salvage but increased mortality: 18-month follow-up study, *Exp. Clin. Endocrinol. Diabetes* (2020) <http://dx.doi.org/10.1055/a-1322-4811>.
- [15] R. Renner, C. Erfurt-Berge, Depression and quality of life in patients with chronic wounds: ways to measure their influence and their effect on daily life, *Chronic Wound Care Manag. Res.* 4 (2017) 143–151, <http://dx.doi.org/10.2147/CWCMR.S124917>.
- [16] M. Polikandrioti, G. Vasilopoulos, I. Koutelekos, G. Panoutsopoulos, G. Geroianni, V. Alikari, E. Dousis, A. Zartaloudi, Depression in diabetic foot ulcer: Associated factors and the impact of perceived social support and anxiety on depression, *Int. Wound J.* 17 (4) (2020) 900–909, <http://dx.doi.org/10.1111/iwj.13348>, URL: <https://onlinelibrary.wiley.com/doi/abs/10.1111/iwj.13348>.
- [17] M.M. Iversen, G.S. Tell, B. Espehaug, K. Midtthjell, M. Graue, B. Rokne, L.I. Berge, T. Østbye, Is depression a risk factor for diabetic foot ulcers? 11-years follow-up of the Nord-Trøndelag Health Study (HUNT), *J. Diabetes Complicat.* 29 (1) (2015) 20–25, <http://dx.doi.org/10.1016/j.jdiacomp.2014.09.006>.
- [18] M. Iversen, J. Igland, H. Smith-Strøm, T. Østbye, G. Tell, S. Skeie, J. Cooper, M. Peyrot, M. Graue, Effect of a telemedicine intervention for diabetes-related foot ulcers on health, well-being and quality of life: secondary outcomes from a cluster randomized controlled trial (DiaFoTo), *BMC Endocr. Disord.* 20 (2020) <http://dx.doi.org/10.1186/s12902-020-00637-x>.
- [19] C.K. Sen, Human wound and its burden: Updated 2020 compendium of estimates, *Adv. Wound Care* 10 (5) (2021) 281–292, <http://dx.doi.org/10.1089/wound.2021.0026>, PMID: 33733885.
- [20] F. Petrone, A. Giribono, L. Massini, L. Pietrangelo, I. Magnifico, U. Bracale, R. Di Marco, R. Bracale, G. Petronio Petronio, Retrospective observational study on microbial contamination of ulcerative foot lesions in diabetic patients, *Microbiol. Res.* 12 (2021) <http://dx.doi.org/10.3390/microbiolres12040058>.
- [21] H. Rathur, A. Boulton, The neuropathic diabetic foot, *Nat. Clin. Pr. Endocrinol. Metab.* 3 (2007) 14–25.
- [22] M.S. Bader, Diabetic foot infection, *Am. Fam. Physician* 78 (1) (2008).
- [23] A.J. Boulton, L. Vileikyte, G. Ragnarson-Tennvall, J. Apelqvist, The global burden of diabetic foot disease, *Lancet* 366 (9498) (2005) 1719–1724, [http://dx.doi.org/10.1016/S0140-6736\(05\)67698-2](http://dx.doi.org/10.1016/S0140-6736(05)67698-2), URL: <https://www.sciencedirect.com/science/article/pii/S0140673605676982>.
- [24] J. Van Netten, D. Clark, P. Lazzarini, M. Janda, L. Reed, The validity and reliability of remote diabetic foot ulcer assessment using mobile phone images, *Sci. Rep.* 7 (2017) <http://dx.doi.org/10.1038/s41598-017-09828-4>.
- [25] J. Apelqvist, J. Larsson, C.-D. Agardh, Long-term prognosis for diabetic patients with foot ulcers, *J. Intern. Med.* 233 (6) (1993) 485–491, <http://dx.doi.org/10.1111/j.1365-2796.1993.tb01003.x>.
- [26] J. Larsson, C.-D. Agardh, J. Apelqvist, A. Stenström, Long term prognosis after healed amputation in patients with diabetes, *Clin. Orthop. Relat. Res.* 350 (1998) 149–158, <http://dx.doi.org/10.1097/00003086-199805000-00021>.
- [27] B. Cassidy, N.D. Reeves, J.M. Pappachan, N. Ahmad, S. Haycocks, D. Gillespie, M. Yap, A cloud-based deep learning framework for remote detection of diabetic foot ulcers, *IEEE Pervasive Comput.* (01) (2022) 1–9, <http://dx.doi.org/10.1109/MPRV.2021.3135686>.
- [28] N.D. Reeves, B. Cassidy, C.A. Abbott, M.H. Yap, Chapter 7 - Novel technologies for detection and prevention of diabetic foot ulcers, in: A. Gefen (Ed.), *The Science, Etiology and Mechanobiology of Diabetes and its Complications*, Academic Press, 2021, pp. 107–122, <http://dx.doi.org/10.1016/B978-0-12-821070-3.00007-6>, URL: <https://www.sciencedirect.com/science/article/pii/B9780128210703000076>.
- [29] J.M. Pappachan, B. Cassidy, C.J. Fernandez, V. Chandrabalan, M.H. Yap, The role of artificial intelligence technology in the care of diabetic foot ulcers: the past, the present, and the future, *World J. Diabetes* 13 (2022) 1131–1139, <http://dx.doi.org/10.4239/wjcd.v13.i12.1131>.
- [30] K. Yammine, M. Estephan, Telemedicine and diabetic foot ulcer outcomes. A meta-analysis of controlled trials, *Foot* (2021) <http://dx.doi.org/10.1016/j.foot.2021.101872>, URL: <https://www.sciencedirect.com/science/article/pii/S0958259221000985>.
- [31] A. Esteve, B. Kuprel, R.A. Novoa, J. Ko, S.M. Swetter, H.M. Blau, et al., Dermatologist-level classification of skin cancer with deep neural networks, *Nature* 542 (7639) (2017) 115–118.
- [32] T.J. Brinker, A. Hekler, A.H. Enk, J. Klode, A. Hauschild, C. Berking, B. Schilling, S. Haferkamp, D. Schadendorf, S. Fröhling, J.S. Utikal, C. von Kalle, A convolutional neural network trained with dermoscopic images performed on par with 145 dermatologists in a clinical melanoma image classification task, *Eur. J. Cancer* 111 (2019) 148–154, <http://dx.doi.org/10.1016/j.ejca.2019.02.005>, URL: <http://www.sciencedirect.com/science/article/pii/S0959804919301443>.
- [33] T.J. Brinker, A. Hekler, A.H. Enk, C. Berking, S. Haferkamp, A. Hauschild, M. Weichenthal, J. Klode, D. Schadendorf, T. Holland-Letz, C. von Kalle, S. Fröhling, B. Schilling, J.S. Utikal, Deep neural networks are superior to dermatologists in melanoma image classification, *Eur. J. Cancer* 119 (2019) 11–17, <http://dx.doi.org/10.1016/j.ejca.2019.05.023>, URL: <http://www.sciencedirect.com/science/article/pii/S0959804919303491>.
- [34] Y. Fujisawa, Y. Otomo, Y. Ogata, Y. Nakamura, R. Fujita, Y. Ishitsuka, R. Watanabe, N. Okiyama, K. Ohara, M. Fujimoto, Deep-learning-based, computer-aided classifier developed with a small dataset of clinical images surpasses board-certified dermatologists in skin tumour diagnosis, *Br. J. Dermatol.* 180 (2) (2019) 373–381, <http://dx.doi.org/10.1111/bjd.16924>.
- [35] T.C. Pham, V.D. Hoang, C.T. Tran, M.S.K. Luu, D.A. Mai, A. Doucet, C.M. Luong, Improving binary skin cancer classification based on best model selection method combined with optimizing full connected layers of Deep CNN, in: 2020 International Conference on Multimedia Analysis and Pattern Recognition, MAPR, 2020, pp. 1–6, <http://dx.doi.org/10.1109/MAPR49794.2020.9237778>.
- [36] S. Jinnai, N. Yamazaki, Y. Hirano, Y. Sugawara, Y. Ohe, R. Hamamoto, The development of a skin cancer classification system for pigmented skin lesions using deep learning, *Biomolecules* 10 (8) (2020).
- [37] H. Haenssle, J. Winkler, C. Fink, F. Toberer, A. Enk, W. Stolz, T. Kränke, R. Hofmann-Wellenhof, H. Kittler, P. Tschandl, C. Rosendahl, A. Lallas, A. Blum, M. Abassi, L. Thomas, I. Tromme, A. Rosenberger, M. Bachelier, S. Bajaj, P. Zukervar, Skin lesions of face and scalp - Classification by a market-approved convolutional neural network in comparison with 64 dermatologists, *Eur. J. Cancer* 144 (2021) 192–199, <http://dx.doi.org/10.1016/j.ejca.2020.11.034>.
- [38] P. Sheehan, P. Jones, A. Caselli, J. Giurini, A. Veves, Percent change in wound area of diabetic foot ulcers over a 4-week period is a robust predictor of complete healing in a 12-week prospective trial, *Diabetes Care* 26 (2003) 1879–1882, <http://dx.doi.org/10.2337/diacare.26.6.1879>.
- [39] M. Goyal, N. Reeves, S. Rajbhandari, M.H. Yap, Robust methods for real-time diabetic foot ulcer detection and localization on mobile devices, *IEEE J. Biomed. Heal. Inform.* (2018).
- [40] B. Cassidy, M. Hoon Yap, J.M. Pappachan, N. Ahmad, S. Haycocks, C. O'Shea, C.J. Fernandez, E. Chacko, K. Jacob, N.D. Reeves, Artificial intelligence for automated detection of diabetic foot ulcers: A real-world proof-of-concept clinical evaluation, *Diabetes Res. Clin. Pract.* 205 (2023) 110951, <http://dx.doi.org/10.1016/j.diabres.2023.110951>, URL: <https://www.sciencedirect.com/science/article/pii/S0168822723007143>.
- [41] M. Goyal, M.H. Yap, N.D. Reeves, S. Rajbhandari, J. Spragg, Fully convolutional networks for diabetic foot ulcer segmentation, in: 2017 IEEE International Conference on Systems, Man, and Cybernetics, SMC, 2017, pp. 618–623, <http://dx.doi.org/10.1109/SMC.2017.8122675>.
- [42] C. Wang, D.M. Anisuzzaman, V. Williamson, M.K. Dhar, B. Rostami, J. Niezgoda, S. Gopalakrishnan, Z. Yu, Fully automatic wound segmentation with deep convolutional neural networks, *Sci. Rep.* 10 (1) (2020) 1–9, <http://dx.doi.org/10.1038/s41598-020-78799-w>.
- [43] C. Wang, A. Mahbod, I. Ellinger, A. Galdan, S. Gopalakrishnan, J. Niezgoda, Z. Yu, FUSeg: The foot ulcer segmentation challenge, 2022, <http://dx.doi.org/10.48550/ARXIV.2201.00414>, URL: <https://arxiv.org/abs/2201.00414>.
- [44] A. Mahbod, R. Ecker, I. Ellinger, Automatic foot ulcer segmentation using an ensemble of convolutional neural networks, 2021, arXiv preprint arXiv:2109.01408.
- [45] G. Scebbia, J. Zhang, S. Catanzaro, C. Mihai, O. Distler, M. Berli, W. Karlen, Detect-and-segment: a deep learning approach to automate wound image segmentation, 2021, arXiv preprint arXiv:2111.01590.
- [46] S. Yang, J. Park, H. Lee, S. Kim, B.-U. Lee, K.-Y. Chung, B. Oh, Sequential change of wound calculated by image analysis using a color patch method during a secondary intention healing, *PLOS ONE* 11 (2016) <http://dx.doi.org/10.1371/journal.pone.0163092>.
- [47] B. Cassidy, C. Kendrick, N. Reeves, J. Pappachan, C. O'Shea, D. Armstrong, M.H. Yap, Diabetic foot ulcer grand challenge 2021: Evaluation and summary, 2022, pp. 90–105, <http://dx.doi.org/10.1007/978-3-030-94907-5.7>.
- [48] S. Thomas, Medetec, 2014, URL: <http://www.medetec.co.uk/index.html>. last access: 08/11/21.

- [49] T.-Y. Liao, C.-H. Yang, Y.-W. Lo, K.-Y. Lai, P.-H. Shen, Y.-L. Lin, HardNet-DFUS: An enhanced harmonically-connected network for diabetic foot ulcer image segmentation and colonoscopy polyp segmentation, 2022, <http://dx.doi.org/10.48550/ARXIV.2209.07313>, URL: <https://arxiv.org/abs/2209.07313>.
- [50] C.-H. Huang, H.-Y. Wu, Y.-L. Lin, HardNet-MSEG: A simple encoder-decoder polyp segmentation neural network that achieves over 0.9 mean dice and 86 FPS, 2021, <http://dx.doi.org/10.48550/arXiv.2101.07172>, ArXiv Preprint [arXiv:2101.07172v2](https://arxiv.org/abs/2101.07172v2).
- [51] P. Chao, C.-Y. Kao, Y. Ruan, C.-H. Huang, Y.-L. Lin, HardNet: A low memory traffic network, 2019, pp. 3551–3560, <http://dx.doi.org/10.1109/ICCV.2019.00365>.
- [52] D. Ramachandram, J. Ramírez-GarcíaLuna, R. Fraser, M.A. Martínez-Jiménez, J. Arriaga-Caballero, J. Allport, Improving objective wound assessment: Fully-automated wound tissue segmentation using deep learning on mobile devices, JMIR Mhealth Uhealth 10 (2022) <http://dx.doi.org/10.2196/36977>.
- [53] M. Swerdlow, O. Guler, R. Yaakov, D.G. Armstrong, Simultaneous segmentation and classification of pressure injury image data using mask-RCNN, Comput. Math. Methods Med. 2023 (2023) 1–7, <http://dx.doi.org/10.1155/2023/3858997>.
- [54] S. Gowda, C. Yuan, ColorNet: Investigating the importance of color spaces for image classification, 2019, pp. 581–596, http://dx.doi.org/10.1007/978-3-030-20870-7_36.
- [55] P. Simon, B. Uma, DeepLumina: A method based on deep features and luminance information for color texture classification, Comput. Intell. Neurosci. 2022 (2022) 1–16, <http://dx.doi.org/10.1155/2022/9510987>.
- [56] C. McBride, B. Cassidy, C. Kendrick, N.D. Reeves, J.M. Pappachan, M.H. Yap, Multi-colour space channel selection for improved chronic wound segmentation, in: 2024 IEEE International Symposium on Biomedical Imaging, ISBI, 2024, pp. 1–5, <http://dx.doi.org/10.1109/ISBI56570.2024.10635155>.
- [57] D. Wen, S.M. Khan, A.J. Xu, H. Ibrahim, L. Smith, J. Caballero, L. Zepeda, C. de Blas Perez, A.K. Denniston, X. Liu, R.N. Martin, Characteristics of publicly available skin cancer image datasets: a systematic review, Lancet Digit. Heal. (2021) [http://dx.doi.org/10.1016/S2589-7500\(21\)00252-1](http://dx.doi.org/10.1016/S2589-7500(21)00252-1), URL: <https://www.sciencedirect.com/science/article/pii/S2589750021002521>.
- [58] M. Benčević, M. Habijan, I. Galić, D. Babin, A. Pižurica, Understanding skin color bias in deep learning-based skin lesion segmentation, Comput. Methods Programs Biomed. 245 (2024) 108044, <http://dx.doi.org/10.1016/j.cmpb.2024.108044>, URL: <https://www.sciencedirect.com/science/article/pii/S0169260724000403>.
- [59] I. Dipto, B. Cassidy, C. Kendrick, N. Reeves, J. Pappachan, V. Chandrabalan, M.H. Yap, Quantifying the effect of image similarity on diabetic foot ulcer classification, 2023, pp. 1–18, http://dx.doi.org/10.1007/978-3-031-26354-5_1.
- [60] B. Akkoca-Gazioğlu, M. Kamasak, Effects of objects and image quality on melanoma classification using Deep Neural Networks, 2020, <http://dx.doi.org/10.21203/rs.3.rs-35907/v1>.
- [61] B. Cassidy, C. Kendrick, A. Brodzicki, J. Jaworek-Korjakowska, M.H. Yap, Analysis of the ISIC image datasets: Usage, benchmarks and recommendations, Med. Image Anal. (2021) <http://dx.doi.org/10.1016/j.media.2021.102305>, URL: <https://www.sciencedirect.com/science/article/pii/S1361841521003509>.
- [62] R. Daneshjoui, C. Barata, B. Betz-Stablein, M.E. Celebi, N. Codella, M. Combalia, P. Gutera, D. Gutman, A. Halpern, B. Helba, H. Kittler, K. Kose, K. Liopyris, J. Malvehy, H.S. Seog, H.P. Soyer, E.R. Tkaczyk, P. Tschandl, V. Rotemberg, Checklist for evaluation of image-based artificial intelligence reports in dermatology: CLEAR dermat consensus guidelines from the international skin imaging collaboration artificial intelligence working group, JAMA Dermatol. (2021) <http://dx.doi.org/10.1001/jamadermatol.2021.4915>.
- [63] J. Winkler, K. Sies, C. Fink, F. Toberer, A. Enk, M. Abassi, T. Fuchs, H. Haenssle, Association between different scale bars in dermoscopic images and diagnostic performance of a market-approved deep learning convolutional neural network for melanoma recognition, Eur. J. Cancer 145 (2021) 146–154, <http://dx.doi.org/10.1016/j.ejca.2020.12.010>.
- [64] J. Jaworek-Korjakowska, A. Wojcicka, D. Kucharski, A. Brodzicki, C. Kendrick, B. Cassidy, M.H. Yap, Skin_Hair Dataset: Setting the Benchmark for Effective Hair Inpainting Methods for Improving the Image Quality of Dermoscopic Images, Springer, 2023, pp. 167–184, http://dx.doi.org/10.1007/978-3-031-25069-9_12.
- [65] S.W. Pewton, B. Cassidy, C. Kendrick, M.H. Yap, Dermoscopic dark corner artifacts removal: Friend or foe? Comput. Methods Programs Biomed. 244 (2024) 107986, <http://dx.doi.org/10.1016/j.cmpb.2023.107986>, URL: <https://www.sciencedirect.com/science/article/pii/S0169260723006521>.
- [66] B. Cassidy, N.D. Reeves, J.M. Pappachan, D. Gillespie, C. O'Shea, S. Rajbhandari, A.G. Maiya, E. Frank, A.J.M. Boulton, D.G. Armstrong, B. Najafi, J. Wu, R.S. Kochhar, M.H. Yap, The DFUC 2020 dataset: Analysis towards diabetic foot ulcer detection, TouchREVIEWS Endocrinol. 17 (2021) 5–11, <http://dx.doi.org/10.17925/EE.2021.17.1.5>, URL: <https://www.touchendocrinology.com/diabetes/journal-articles/the-dfuc-2020-dataset-analysis-towards-diabetic-foot-ulcer-detection/1>.
- [67] M.H. Yap, B. Cassidy, J.M. Pappachan, C. O'Shea, D. Gillespie, N.D. Reeves, Analysis towards classification of infection and ischaemia of diabetic foot ulcers, in: 2021 IEEE EMBS International Conference on Biomedical and Health Informatics, BHI, 2021, pp. 1–4, <http://dx.doi.org/10.1109/BHI50953.2021.9508563>.
- [68] C. Kendrick, B. Cassidy, J.M. Pappachan, C. O'Shea, C.J. Fernandez, E. Chacko, K. Jacob, N.D. Reeves, M.H. Yap, Translating clinical delineation of diabetic foot ulcers into machine interpretable segmentation, in: M.H. Yap, C. Kendrick, R. Brüngel (Eds.), Diabetic Foot Ulcers Grand Challenge, Springer Nature Switzerland, Cham, 2025, pp. 1–14.
- [69] M.H. Yap, R. Hachiuma, A. Alavi, R. Brüngel, B. Cassidy, M. Goyal, H. Zhu, J. Rückert, M. Olshansky, X. Huang, H. Saito, S. Hassanpour, C.M. Friedrich, D.B. Ascher, A. Song, H. Kajita, D. Gillespie, N.D. Reeves, J.M. Pappachan, C. O'Shea, E. Frank, Deep learning in diabetic foot ulcers detection: A comprehensive evaluation, Comput. Biol. Med. 135 (2021) 104596, <http://dx.doi.org/10.1016/j.compbiomed.2021.104596>, URL: <https://www.sciencedirect.com/science/article/pii/S0010482521003905>.
- [70] M.H. Yap, C. Kendrick, N. Reeves, M. Goyal, J. Pappachan, B. Cassidy, Development of diabetic foot ulcer datasets: An overview, 2022, pp. 1–18, http://dx.doi.org/10.1007/978-3-030-94907-5_1.
- [71] M.H. Yap, B. Cassidy, M. Byra, T. yu Liao, H. Yi, A. Galdran, Y.-H. Chen, R. Brüngel, S. Koitka, C.M. Friedrich, Y. wen Lo, C. hui Yang, K. Li, Q. Lao, M.A.G. Ballester, G. Carneiro, Y.-J. Ju, J.-D. Huang, J.M. Pappachan, N.D. Reeves, V. Chandrabalan, D. Dancy, C. Kendrick, Diabetic foot ulcers segmentation challenge report: Benchmark and analysis, Med. Image Anal. 94 (2024) 103153, <http://dx.doi.org/10.1016/j.media.2024.103153>, URL: <https://www.sciencedirect.com/science/article/pii/S1361841524000781>.
- [72] L. Alzubaidi, M.A. Fadhel, S.R. Oleiwi, O. Al-Shamma, J. Zhang, DFU_QUTNet: Diabetic foot ulcer classification using novel deep convolutional neural network, Multimedia Tools Appl. 79 (21–22) (2020) 15655–15677, <http://dx.doi.org/10.1007/s11042-019-07820-w>.
- [73] C. Wang, A. Mahbod, I. Ellinger, A. Galdran, S. Gopalakrishnan, J. Niezgoda, Z. Yu, FUSeg: The foot ulcer segmentation challenge, Information 15 (2024) 140, <http://dx.doi.org/10.3390/info15030140>.
- [74] M. Groh, C. Harris, L. Soenksen, F. Lau, R. Han, A. Kim, A. Koochek, O. Badri, Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 1820–1828.
- [75] M. Kręćhwest, J. Czajkowska, A. Wijata, J. Juszczyk, B. Pyciński, M. Biesok, M. Rudzki, J. Majewski, J. Kostecki, E. Pietka, Chronic wounds multimodal image database, Comput. Med. Imaging Graph. 88 (2021) <http://dx.doi.org/10.1016/j.compmedimag.2020.101844>, URL: <https://www.sciencedirect.com/science/article/pii/S0895611120301397>.
- [76] T.A. Pereira, R.C. Popim, L.A. Passos, D.R. Pereira, C.R. Pereira, J.P. Papa, ComplexWoundDB: A database for automatic complex wound tissue categorization, in: 2022 29th International Conference on Systems, Signals and Image Processing, IWSSIP, Vol. CFP2255E-ART, 2022, pp. 1–4, <http://dx.doi.org/10.1109/IWSSIP55020.2022.9854419>.
- [77] S.R. Oota, V. Rowtula, S. Mohammed, M. Liu, M. Gupta, WSNet: Towards an effective method for wound image segmentation, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, WACV, 2023, pp. 3234–3243.
- [78] A. Dutta, A. Gupta, A. Zissermann, VGG image annotator (VIA), 2016, <http://www.robots.ox.ac.uk/~vgg/software/via/>. Version: 2.0.11, Accessed: 21st April 2022.
- [79] A. Dutta, A. Zisserman, The VIA annotation software for images, audio and video, in: Proceedings of the 27th ACM International Conference on Multimedia, MM '19, ACM, New York, NY, USA, 2019, <http://dx.doi.org/10.1145/3343031.3350535>.
- [80] D.-J. Kroon, Snake: Active contour, 2022, Online. URL: <https://www.mathworks.com/matlabcentral/fileexchange/28149-snake-active-contour>.
- [81] D. Jha, P.H. Smedsrud, M.A. Riegler, D. Johansen, T.D. Lange, P. Halvorsen, H. D. Johansen, ResUNet++: An advanced architecture for medical image segmentation, in: Proceedings of the IEEE International Symposium on Multimedia, ISM, 2019, pp. 225–230.
- [82] 4uiiur21, PyTorch implementation of UNet++ (Nested U-Net), 2020, <https://github.com/4uiiur21/pytorch-nested-unet>. (Accessed 30 March 2023).
- [83] S. Czekalski, Attention U-Net, 2020, https://github.com/sfczekalski/attention_unet. Accessed: 30th March 2023.
- [84] Q. Xu, Z. Ma, N. He, W. Duan, DCSAU-Net: A deeper and more compact split-attention U-Net for medical image segmentation, Comput. Biol. Med. 154 (2023) 106626, <http://dx.doi.org/10.1016/j.compbiomed.2023.106626>, URL: <https://www.sciencedirect.com/science/article/pii/S0010482523000914>.
- [85] S. Jin, S. Yu, J. Peng, H. Wang, Y. Zhao, A novel medical image segmentation approach by using multi-branch segmentation network based on local and global information synchronous learning, Sci. Rep. 13 (2023) <http://dx.doi.org/10.1038/s41598-023-33357-y>.
- [86] Y. Li, Human segmentation in Pytorch, 2023, <https://github.com/cavalleria/humanseg.pytorch>. (Accessed 12 March 2023).
- [87] B. Mayali, Pretrained backbones with UNet, 2023, <https://github.com/mberkay0/pretrained-backbones-unet>. (Accessed 30 March 2023).
- [88] J.M.J. Valanarasu, V.M. Patel, UNeXt: MLP-based rapid medical image segmentation network, 2022, <http://dx.doi.org/10.48550/arXiv.2203.04967>, arXiv preprint [arXiv:2203.04967](https://arxiv.org/abs/2203.04967).

- [89] T.-Y. Liao, C.-H. Yang, Y.-W. Lo, K.-Y. Lai, P.-H. Shen, Y.-L. Lin, HardNet-DFUS: An enhanced harmonically-connected network for diabetic foot ulcer image segmentation and colonoscopy polyp segmentation, 2022, <http://dx.doi.org/10.48550/arXiv.2209.07313>, arXiv preprint arXiv:2209.07313.
- [90] H. Zhu, B. Chen, C. Yang, Understanding why ViT trains badly on small datasets: An intuitive perspective, 2023, <http://dx.doi.org/10.48550/arXiv.2302.03751>, arXiv preprint arXiv:2302.03751.
- [91] H. Gani, M. Naseer, M. Yaqub, How to train vision transformer on small-scale datasets? 2022, <http://dx.doi.org/10.48550/arXiv.2210.07240>, arXiv preprint arXiv:2210.07240.
- [92] T. Fitzpatrick, The validity and practicality of sun-reactive skin types I through VI, *Arch. Dermatol.* 124 (1988) 869–871, <http://dx.doi.org/10.1001/archderm.124.6.869>.
- [93] I. Tolstikhin, N. Houlsby, A. Kolesnikov, L. Beyer, X. Zhai, T. Unterthiner, J. Yung, A. Steiner, D. Keysers, J. Uszkoreit, M. Lucic, A. Dosovitskiy, MLP-mixer: An all-MLP architecture for vision, 2021, <http://dx.doi.org/10.48550/arXiv.2105.01601>, arXiv preprint arXiv:2105.01601.
- [94] K. He, X. Zhang, S. Ren, J. Sun, Spatial pyramid pooling in deep convolutional networks for visual recognition, *IEEE Trans. Pattern Anal. Mach. Intell.* 37 (2014) <http://dx.doi.org/10.1109/TPAMI.2015.2389824>.
- [95] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, Z. Tu, Deeply-supervised nets, in: *Artificial Intelligence and Statistics*, PMLR, 2015, pp. 562–570.
- [96] D. Morales-Brotons, T. Vogels, H. Hendriks, Exponential moving average of weights in deep learning: Dynamics and benefits, *Trans. Mach. Learn. Res.* (2024) URL: <https://openreview.net/forum?id=2M9CUnYnBA>.
- [97] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2015, pp. 234–241.
- [98] International Telecommunication Union, Recommendation BT.709-4, 2000, Online. URL: <https://www.itu.int/rec/R-REC-BT.709-4-200003-S/en>.
- [99] G. Bradski, The OpenCV library, Dr. Dobb's J. Softw. Tools (2000) URL: <https://github.com/opencv/opencv-python>.
- [100] X. Pan, P. Luo, J. Shi, X. Tang, Two at Once: Enhancing Learning and Generalization Capacities via IBN-Net: 15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part IV, 2018, pp. 484–500, http://dx.doi.org/10.1007/978-3-030-01225-0_29.
- [101] D. Ulyanov, A. Vedaldi, V. Lempitsky, Instance normalization: The missing ingredient for fast stylization, 2017, <http://dx.doi.org/10.48550/arXiv.1607.08022>.
- [102] S. Ioffe, C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, 2015, <http://dx.doi.org/10.48550/arXiv.1502.03167>, arXiv preprint arXiv:1502.03167.
- [103] K. He, X. Zhang, S. Ren, J. Sun, Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification, in: *IEEE International Conference on Computer Vision, ICCV 2015*, Vol. 1502, 2015, <http://dx.doi.org/10.1109/ICCV.2015.123>.
- [104] P. Luo, R. Zhang, J. Ren, Z. Peng, J. Li, Switchable normalization for learning-to-normalize deep representation, *IEEE Trans. Pattern Anal. Mach. Intell.* 43 (2) (2021) 712–728, <http://dx.doi.org/10.1109/TPAMI.2019.2932062>.
- [105] L. Alzubaidi, M.A. Fadhel, O. Al-Shamma, J. Zhang, J. Santamaría, Y. Duan, S. R. Olewi, Towards a better understanding of transfer learning for medical imaging: A case study, *Appl. Sci.* 10 (13) (2020) <http://dx.doi.org/10.3390/app10134523>, URL: <https://www.mdpi.com/2076-3417/10/13/4523>.
- [106] R. Brüngel, S. Koitka, C.M. Friedrich, Unconditionally generated and pseudo-labeled synthetic images for diabetic foot ulcer segmentation dataset extension, in: *Diabetic Foot Ulcers Grand Challenge: Third Challenge, DFUC 2022, Held in Conjunction with MICCAI 2022, Singapore, September 22, 2022, Proceedings*, Springer, 2023, pp. 65–79.
- [107] A. Buslaev, V.I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, A.A. Kalinin, Albumentations: Fast and flexible image augmentations, *Information* 11 (2) (2020) <http://dx.doi.org/10.3390/info11020125>, URL: <https://www.mdpi.com/2078-2489/11/2/125>.
- [108] T. Koo, M. Li, A guideline of selecting and reporting intraclass correlation coefficients for reliability research, *J. Chiropr. Med.* 15 (2016) <http://dx.doi.org/10.1016/j.jcm.2016.02.012>.
- [109] J.L. Fleiss, Reliability of measurement, in: *The Design and Analysis of Clinical Experiments*, John Wiley & Sons, Ltd, 1999, pp. 1–32, <http://dx.doi.org/10.1002/9781118032923.ch1>, chapter 1. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118032923.ch1>.
- [110] M.E. Celebi, C. Barata, A. Halpern, P. Tschandl, M. Combalia, Y. Liu, Guest editorial: Image analysis in dermatology, *Med. Image Anal.* 79 (2022) 102468, <http://dx.doi.org/10.1016/j.media.2022.102468>.
- [111] D. Ramachandram, J.L. Ramirez-GarciaLuna, R.D.J. Fraser, M.A. Martínez-Jiménez, J.E. Arriaga-Caballero, J. Allport, Fully automated wound tissue segmentation using deep learning on mobile devices: Cohort study, *JMIR Mhealth Uhealth* 10 (4) (2022) e36977, <http://dx.doi.org/10.2196/36977>.
- [112] M. Combalia, N. Codella, V. Rotemberg, C. Carrera, S. Dusza, D. Gutman, B. Helba, H. Kittler, N. Kurtansky, K. Liopyris, M. Marchetti, S. Podlipnik, S. Puig, C. Rinner, P. Tschandl, J. Weber, A. Halpern, J. Malvehy, Validation of artificial intelligence prediction models for skin cancer diagnosis using dermoscopy images: the 2019 International Skin Imaging Collaboration Grand Challenge, *Lancet Digit. Heal.* 4 (2022) e330–e339, [http://dx.doi.org/10.1016/S2589-7500\(22\)00021-8](http://dx.doi.org/10.1016/S2589-7500(22)00021-8).
- [113] A. Drovandi, S. Wong, L. Seng, B. Crowley, C. Alahakoon, J. Banwait, M. Fernando, J. Golledge, Remotely delivered monitoring and management of diabetes-related foot disease: An overview of systematic reviews, *J. Diabetes Sci. Technol.* (2021) <http://dx.doi.org/10.1177/19322968211012456>.
- [114] N.C. Okafor, B. Cassidy, C. O'Shea, J.M. Pappachan, The Effect of Image Preprocessing Algorithms on Diabetic Foot Ulcer Classification, Springer, Cham, 2024, pp. 336–352, http://dx.doi.org/10.1007/978-3-031-66958-3_25.
- [115] J. Jaworek-Korjakowska, A. Brodzicki, B. Cassidy, C. Kendrick, M.H. Yap, Interpretability of a deep learning based approach for the classification of skin lesions into main anatomic body sites, *Cancers* 13 (23) (2021) <http://dx.doi.org/10.3390/cancers13236048>, URL: <https://www.mdpi.com/2072-6694/13/23/6048>.
- [116] R. Padilla, W.L. Passos, T.L.B. Dias, S.L. Netto, E.A.B. da Silva, A comparative analysis of object detection metrics with a companion open-source toolkit, *Electronics* 10 (3) (2021) <http://dx.doi.org/10.3390/electronics10030279>, URL: <https://www.mdpi.com/2079-9292/10/3/279>.
- [117] R.S. Howell, H.H. Liu, A.A. Khan, J.S. Woods, L.J. Lin, M. Saxena, H. Saxena, M. Castellano, P. Petrone, E. Slone, E.S. Chiu, B.M. Gillette, S.A. Gorenstein, Development of a method for clinical evaluation of artificial intelligence-based digital wound assessment tools, *JAMA Netw. Open* 4 (5) (2021) <http://dx.doi.org/10.1001/jamanetworkopen.2021.7234>.