IEEE Open Journal of Engineering in Medicine and Biology

Technology

Chronic Wound Image Augmentation and Assessment using Semi-Supervised Progressive Multi-Granularity EfficientNet

Ziyang Liu, Emmanuel Agu, Peder Pedersen, Clifford Lindsay, Bengisu Tulu, Diane Strong

***Abstract*—*Goal:* Augment a small, imbalanced, wound dataset by using semi-supervised learning with a secondary dataset. Then utilize the augmented wound dataset for deep learning-based wound assessment.**

***Methods:* The clinically-validated Photographic Wound Assessment Tool (PWAT) scores eight wound attributes: Size, Depth, Necrotic Tissue Type, Necrotic Tissue Amount, Granulation Tissue type, Granulation Tissue Amount, Edges, Periulcer Skin Viability to comprehensively assess chronic wound images. A small corpus of 1639 wound images labeled with ground truth PWAT scores was used as reference. A Semi-Supervised learning and Progressive Multi-Granularity training mechanism were used to leverage a secondary corpus of 9870 unlabeled wound images. Wound scoring utilized the EfficientNet Convolutional Neural Network on the augmented wound corpus.**

***Results:* Our proposed Semi-Supervised PMG EfficientNet (SS-PMG-EfficientNet) approach estimated all 8 PWAT sub-scores with classification accuracies and F1 scores of about 90% on average, and outperformed a comprehensive list of baseline models and had a 7% improvement over the prior state-of-the-art (without data augmentation). We also demonstrate that synthetic wound image generation using Generative Adversarial Networks (GANs) did not improve wound assessment.**

***Conclusions:* Semi-supervised learning on unlabeled wound images in a secondary dataset achieved impressive performance for deep learning-based wound grading.**

***Index Terms*—Chronic wounds, data imbalance, data augmentation, Neural Networks, smartphone assessment.**

***Impact Statement*- Our envisioned smartphone wound assessment system can reduce the significant burden that manual wound grading imposes on wound care nurses.**

I. INTRODUCTION

MOTIVATION: More than 6.5 million people in the US have chronic wounds (or approximately 2% of the population) [1]. Chronic wounds are often painful and are prevalent in the elderly population [2] [3], which costs the healthcare system over $25 billion annually [4]. In order to heal properly, chronic wounds require proper treatment including cleaning, debridement, changing of dressings and using antibiotics [5]. Without proper care, such wounds may become infected [6] or cause limbs to be amputated. The number of chronic wounds is large and growing, increasing the need for more efficient chronic wound care especially information technology solutions that assist the work of medical personnel and reduce the cost of care. Additional background and detailed descriptions of various types of chronic wounds can be found in the Supplementary Materials.

A. Background of our research

Smartphone-based image analyses provide a new method for remote wound assessment [7] [8] [9] [10]. Since 2011, our group has been researching and developing the Smartphone Wound Analysis and Decision-Support (SmartWAnDS) [11] [12] [13] [14] since 2011. SmartWAnDS analyzed the smartphone captured chronic wound images autonomously and provide wound care recommendations to patients and their caregivers. The SmartWAnDS system can provide standardized feedback on wounds for patients when they are at home between hospital visits and engage patients in the care of their wounds It can also support the work of wound nurses with care recommendations when they are in remote locations and wound doctors are unavailable temporarily. The recommended wound care is based on its current status and healing progress since the preceding examination. Consequently, it is necessary to grade the wound before making treatment decisions. The research described in this paper focuses on the SmartWAnDS module that uses deep learning to autonomously grade the wound's healing status based on its visual appearance in a smartphone image.

In collaboration with wound experts (1 wound doctor and 1 wound nurse) who labeled all images, our group created *WoundNet*, a chronic wound image dataset with 1639 chronic wound images totally, as mentioned in our previous research [15]. *WoundNet* contains four types of wounds: diabetic foot ulcers, pressure ulcers, vascular ulcers and surgical wounds, which are the most common types seen by wound experts at hospitals [16]. Table Ia summarizes the statistics of the four wound types in *WoundNet*. Example images of diabetic, venous, arterial and pressure ulcers are shown in Fig. 1 (a).

This project is funded by the National Institutes of Health under grant number (1R01EB025801).

Corresponding author: Emmanuel Agu. Ziyang Liu and Emmanuel Agu are with the Computer Science Department, Worcester Polytechnic Institute, Worcester, MA, USA. Peder Pedersen is with the Electrical and Computer Engineering Department, Worcester Polytechnic Institute, Worcester, MA, USA. Clifford Lindsay is with the Department of Radiology, University of Massachusetts Medical School, Worcester, MA, USA. Bengisu Tulu and Diane Strong are with the Foisie Business School, Worcester Polytechnic Institute, Worcester, MA, USA. (e-mail: zliu10@wpi.edu, emmanuel@wpi.edu, pedersen@wpi.edu, clifford.lindsay@umassmed.edu, bengisu@wpi.edu, dstrong@wpi.edu)

This article has supplementary materials.

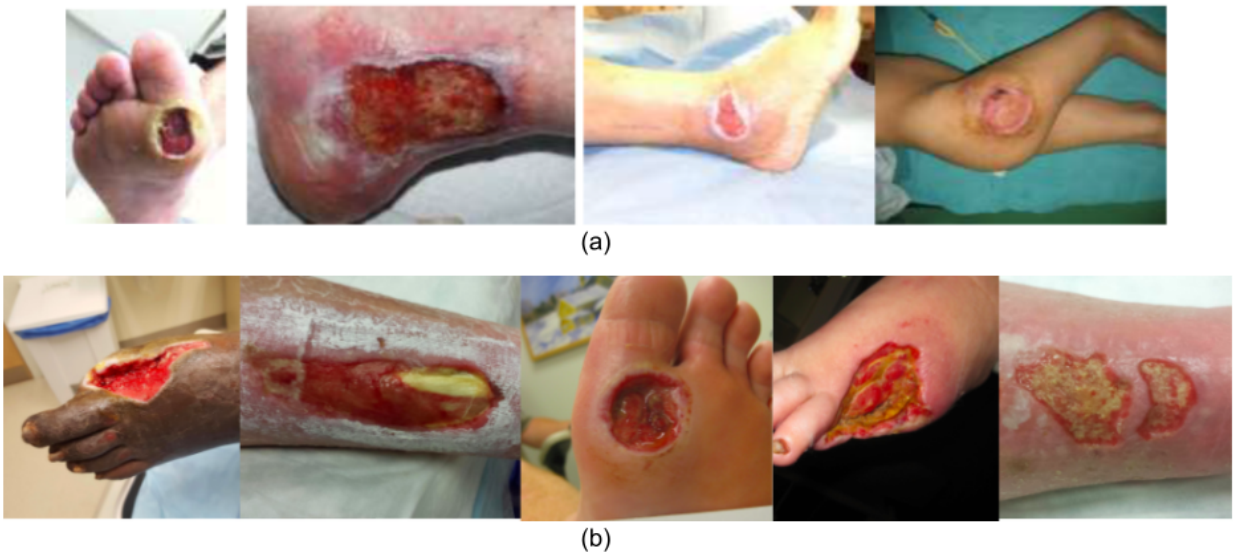


Fig. 1. (a) Examples of diabetic, venous, arterial and pressure ulcers wound types (left to right) (b) Example wound images corresponding to PWAT Necrotic Amount scores 0 (left) to 4 (right). The target sub-classes (Necrotic amount scores) can appear quite similar visually, posing a fine-grained image classification problem

Prior work including ours [15] has explored using machine and deep learning models to grade wound healing status. As ground truth healing assessment scores that machine learning models can predict as target labels, each wound in *WoundNet* was comprehensively graded using the Photographic Wound Assessment Tool (PWAT), a clinically validated wound grading rubric [17] [18] [19]. The PWAT evaluates eight attributes of wounds [19] from an image: 1) Size 2) Depth 3) Necrotic Tissue Type 4) Necrotic Tissue Amount 5) Granulation Tissue Type 6) Granulation Tissue Amount 7) Edges and 8) Skin viability. Each PWAT sub-score grades a single wound attribute with a score of 0 (best), 1, 2, 3 or 4 (worst) and higher scores indicate a worse wound condition. All 8 PWAT sub-scores are summed to generate a total PWAT wound score (max = 32). The PWAT sub-scores for Necrotic Tissue Type, Necrotic Tissue Amount, Granulation Tissue Type and Granulation Tissue Amount are abbreviated as Nec Type, Nec Amount, Gran Type and Gran Amount respectively in this paper. A table with detailed descriptions of each PWAT sub-score and their corresponding grading criteria can be found in the Supplementary Materials.

(a) Statistics of types of wounds in *WoundNet* dataset

Wound Types	Numbers of Images
1. Diabetic Foot Ulcers	121
2. Pressure Ulcers	13
3. Vascular Ulcers	1349
4. Surgical Wounds	156

(b) Statistics of PWAT sub-scores of images in *WoundNet* dataset

PWAT subscore	0	1	2	3	4
1. Size	141	393	537	337	231
2. Depth	134	135	814	378	178
3. Necrotic Type	509	472	273	76	309
4. Necrotic Amount	347	160	150	290	692
5. Granulation Type	143	319	626	95	456
6. Granulation Amount	137	77	194	342	889
7. Edges	142	311	1035	131	20
8. Skin	435	1061	143	0	0

TABLE I. Statistics of types of wounds and PWAT sub-scores of images in *WoundNet* dataset

B. Problem

Due to the high cost associated with collecting medical datasets and variability in the occurrence of various wound severities, many medical datasets are small and imbalanced, which presents a challenge to machine and deep learning. Labeling is manual and often has to be done with experts whose time is expensive. As shown in Table Ib, the number of images corresponding to several PWAT sub-scores in *WoundNet* was inadequate and the distribution of sub-scores was imbalanced. This presented a challenge to deep learning wound assessment model development, and prevented clinically usable classification performance from being performance achieved.

Table II summarizes related work. Prior research on assessing PWAT wound attributes from images [20] [21] [22] [23] [24] [25] [26] [27] [28] [29] typically only assessed a few wound attributes instead of assessing all clinically important attributes of wounds comprehensively. Prior research that explored data augmentation of wound images [30] [31] [32] [33] [34] [35] applied both traditional data augmentation techniques or GAN-based methods. Prior research on other medical image problems used data augmentation methods [36] [37] [38] [39] [40] [41] including GANs methods and transform models, which improved the performance of machine learning models. The previous state-of-the-art neural networks model for PWAT wound assessment was our Patch Attention DenseNet [15] but it achieved only about 82% in accuracy and F1-score,

Medical imaging research assessing PWAT wound attributes						
Related research	PWAT subscore	machine learning method	task	wound type	dataset size	results
Chino et al. 2020 [20]	1. Size	deep neural network	segment the wound; estimate the size	venous and arterial ulcer	446	estimate wound area in cm2 with error of 14%
Spinczyk et al. 2017 [21]		triangulation technique	wound 3D surface reconstruction	not specified	10 patient	measure wound area with error of 11%
Hsu et al. 2017 [22]	3. Nec Type 4. Nec Amount 5. Gran Type 6. Gran Amount	clustering method, SVM	detect necrotic tissue	postsurgical wound	42	detection accuracy 95.23%
Blanco et al. 2020 [23]		superpixel-driven deep learning approach	segmenting Necrotic and Granulation, and wounded area	arterial and venous ulcers	217	spot wounded tissues with AUC = 0.986
Godeiro et al. 2018 [24]		deep neural network	classifying Necrotic and Granulation	chronic wounds	30	tissue classification accuracy 96%
Nejati et al. 2018 [25]		deep neural network, SVM	classifying Necrotic and Granulation	chronic wounds	350	tissue classification accuracy 86.4%
Hsu et al. 2019 [26]		robust image segmentation, SVM	wound segmentation; detect Necrotic and Granulation	chronic wounds	293	tissue classification accuracy 83.58%
Maity et al. 2018 [27]		deep neural network	classifying Necrotic and Granulation	chronic wounds	68	tissue classification accuracy 99%
Babu et al. 2018 [28]		Naive bayes and Hoeffding tree	wound segmentation; classifying Necrotic and Granulation	diabetic wound	N. A.	tissue classification accuracy 90.9%
Rajathi et al. 2019 [29]		deep neural network	classifying Necrotic and Granulation	varicose ulcer	1250	tissue classification accuracy 99.55%
Wound (all Diabetic Foot Ulcer) medical imaging that explored data augmentation						
Related research	data augmentation method	machine learning method	task	dataset size		results
Bloch et al. 2021 [30]	pix2pixHD	EfficientNets Ensemble	4 classes (Infection and Ischaemia, Infection, Ischaemia, None)	Infection and Ischaemia, Infection, Ischaemia, None: 621, 2555, 277, 2552		best macro F1-Score of 60.77 %
Das et al. 2022 [31]	horizontal and vertical flips	CNN-based classification model	DFU vs. normal skin binary classification	292 Ulcer foot, 105 healthy foot, augmented to 641 normal. 1038 abnormal		96.4% accuracy, 95.4% F1 score
Goyal et al. 2020 [32]	natural data augmentation	ensemble CNN	2 classes (Ischaemia: Yes, No; Infection: Yes, No) classification	Ischaemia: (235, 1431) augmented: (4935, 4935) Infection: (982, 684) augmented: (2946, 2946)		90% accuracy in ischaemia, 73% accuracy in infection
Yap et al. 2021 [33]	flip, natural data augmentation, crop, Gaussian noise, rotate shear, scale and adjust contrast	VGG16, ResNet101, InceptionV3, DenseNet121, EfficientNet	4 classes (both Infection and Ischaemia, Infection, Ischaemia, None)	Infection and Ischaemia, Infection, Ischaemia, None:621, 2555, 277, 2552		EfficientNet B0: macro-average Precision, Recall and F1-Score of 0.57, 0.62 and 0.55
Al-Garaawi et al. 2022 [34]	rotation, flipping, color space augmentation	CNN-based DFU classification method	Part A: 2 classes (healthy and DFU); Part B: 2 classes (Ischaemia: Yes, No; Infection: Yes, No)	Part A: 641,1038; Part B: Ischaemia: (4935, 4935); Infection: (2946, 2946)		Ischaemia: 0.995% (AUC), 0.990% (F-measure) Infection: 0.820% (AUC), 0.744% (F-measure)
Goyal et al. 2018 [35]	rotation, contrast enhancement, color space, random scaling, flipping	a novel CNN model	2 classes (healthy skin and DFU)	641 healthy, 1038 DFU		accuracy 0.925, F-measure 0.939, AUC 0.962
Non-wound medical imaging tasks that explored data augmentation						
Related research	data augmentation method	machine learning method	task	dataset size		results
Frid-Adar et al. 2018 [36]	Deep Convolutional GAN (DCGAN)	CNN model	liver lesion CT images classification	liver lesions (CT) images: 182, augmented: 5000		improvement of 7% using GAN with accuracy of 85.7%
Zhao et al. 2019 [37]	learned spatial and appearance transform models	deep fully convolutional neural networks	one-shot segmentation of brain (MRI) scans	101 brain scan, synthesis of 10,000 different labeled examples		Dice score of 0.815 (0.123)
Ghorbani et al. 2020 [38]	pix2pix (GAN) based	MobileNet CNN model	skin lesions classification	49920 images, 20000 synthetic images		about 50% accuracy, 14% improvement in F1 score in classes with fewer examples
Pollastri et al. 2020 [39]	DCGAN, LAPGAN	custom CNN	skin lesions segmentation	training set: 1882, 2000 augmented images		jaccard index of 0.789
Pang et al. 2021 [40]	semi-supervised GAN: TripleGAN	Inception-V3 model	Breast Ultrasound Mass Classification	1447 ultrasound images, augmented to 4341		accuracy 90.41%, sensitivity 87.94%, specificity 85.86%
Guan et al. 2022 [41]	texture-constrained multichannel progressive generative adversarial network (TMP-GAN)	faster-RCNN	lesion detection of mammography data set and pancreatic tumors	CBIS-DDMS: 1318 pancreatic tumor dataset: 1066; synthetic images: 1, 4, and 9 times the actual images		precision, recall, F1-score CBIS-DDMS: improves 2.59%/2.70%/2.77%, to 86.28%/85.89%/86.08% pancreatic tumor: improves 2.44%/2.06%/ 2.36%, to 86.28%/85.89%/86.08%

TABLE II. Prior work on assessing PWAT attributes and medical imaging (wound and non-wound) data augmentation

which was not clinically usable. Prior research on various medical imaging problems [36] [40] [41] have demonstrated improvements in their model's performance using synthetic images generated using traditional data augmentation methods and GANs. However, in our experiments, we discovered that synthetic images generated using GANs and traditional data augmentation methods did not improve the performance of wound assessment neural networks model. Consequently, in this paper, we proposed a novel method for leveraging a large, external dataset of unlabeled wound images using semi-supervised learning, which improves PWAT-based wound assessment using neural networks models. A detailed analysis of the limitations of prior work is presented in the supplementary materials.

C. Our approach

Semi-supervised learning is used to augment a small, labeled corpus by leveraging a large unlabeled corpus. In this paper, we propose a semi-supervised learning aiding [42] [43] [44], Progressive Multi-Granularity mechanism [45] based EfficientNet B0 architecture [46], named Semi-Supervised PMG EfficientNet (SS-PMG-EfficientNet), to improve the number and balance of our *WoundNet* dataset and utilize the augmented dataset to improve the accuracy of our wound assessment system. SS-PMG-EfficientNet was a creative integration of the semi-supervised learning method and Progressive Multi-Granularity mechanism with the EfficientNet B0 CNN model. SS-PMG-EfficientNet was trained on our *WoundNet* dataset and used to analyze wound images to assess their healing status. The PMG mechanism was a state-of-the-art fine-grained image classification method with its own data augmentation method designed specifically. The semi-supervised learning method enabled our deep learning wound assessment model to utilize other secondary sources of unlabeled wound image dataset for data augmentation while using our *WoundNet* as a labeled reference dataset. EfficientNet is a state-of-the-art image classification architecture that has achieved good performance on wound image related research. In our research to develop SS-PMG-EfficientNet, simpler variant architectures named PMG EfficientNet and Semi-Supervised EfficientNet were also developed, which were integrated models generated by combining the PMG mechanism and EfficientNet B0, and the Semi-Supervised learning component and EfficientNet B0 respectively. Finally SS-PMG-EfficientNet was developed from integrating the semi-supervised learning method and the PMG mechanism into EfficientNet B0 in order to improve the model's performance as much as possible.

1) Semi-supervised learning:

Semi-supervised learning jointly learns from unlabeled data using an unsupervised loss function as well as from labeled data using traditional supervised loss function [42] [43] [44]. The unsupervised loss from the unlabeled data acts as a regularization term for the labeled data's loss. Typically the labeled dataset and unlabeled dataset are sampled from the same distribution. This approach facilitates sampling from a different yet related distribution. To facilitate semi-supervised learning on wound images, the labeled dataset we utilized was our *WoundNet* with all 8 PWAT sub-scores labeled on all 1639 wound images, in conjunction with a larger unlabeled DFUC (Diabetic Foot Ulcer) 2021 dataset [33] [47] with 9870 wound images. We considered the DCUC 2021 dataset as the unlabeled source dataset because while it had infection and ischaemia ground truth assessments by wound experts, the images contained no PWAT sub-scores labels. As Diabetic Foot Ulcers (DFUs) are common chronic wound types and all chronic wounds types have similar appearance with similar features, it was reasonable to utilize DFU images from DFUC 2021 dataset as unlabeled images for semi-supervised learning in order to improve model performance and reduce overfitting. This innovative method of applying semi-supervised learning method using our labeled *WoundNet* dataset and unlabeled images from the DFUC 2021 dataset can be considered a type of data augmentation. The DFUC 2021 dataset was not directly added to the *WoundNet* dataset for the PWAT classification problem but it provided the CNN model with more Diabetic Foot Ulcer images. Allowing the CNN models to see more highly related images, especially when our own dataset was relatively small, encouraged the model to better learn chronic wound features.

2) Progressive Multi-Granularity mechanism:

The Progressive Multi-Granularity mechanism [45] was proposed in a research that combined part granularity learning and cross-granularity feature fusion to work simultaneously. It had a progressive training strategy that fused features from different granularities and a random jigsaw patch generator that forced the model to learn features at specific granularities. The PMG (Progressive Multi-Granularity) mechanism was built on the assumption that fine-grained discriminative information could be extracted from different visual granularities. The PMG mechanism allowed the model to learn at different granularities while fusing multigranularity features simultaneously, instead of detecting image parts first then fusing them later. This PMG framework started with stable finer granularities first and then coarser granularities so that it could avoid the confusion from large intra-class variations in large regions. However, the progressive training tended to focus on learning multi-granularity information from similar region. This problem was tackled by the jigsaw puzzle generator, which generated different granularity levels at each training step that are input to the model. It forced the model to focus on local patch levels that corresponded to specific granularity level, instead of learning the entire image. The Progressive Multi-Granularity mechanism can be viewed as a combination of modifying the CNN model's architecture and providing this new architecture with specific type of images generated from the jigsaw puzzle generator. Sourced from *WoundNet* images, the jigsaw puzzle generated images could be considered as another type of images for data augmentation, as shown in Fig. 3 (b). Chronic wound images augmented in this specific way improve the learning of information contained in different granularities by the PMG mechanism based CNN model, which improves the model's performance on the PWAT classification problem.

D. Novelty of our work

In order to solve related problems of inadequate labeled data and data imbalance, we explored innovative data augmentation and semi-supervised learning approaches. Specifically, the Progressive Multi-Granularity mechanism [45] and semi-supervised learning method [42] [43] [44] were innovatively integrated into the EfficientNet B0 CNN architecture [46], which improved our model's performance significantly by 7% and achieved almost 90% accuracy and F1 score for all 8 PWAT scores. We also demonstrate that our proposed approach outperforms a comprehensive set of baselines that included Generative Adversarial Networks (GANs), which are widely considered the state-of-the-art for data augmentation. In addition to solving data insufficiency and imbalance issues, our model also comprehensively analyzed wounds based on the comprehensive PWAT rubric.

The fine-grained nature of PWAT sub-score prediction is illustrated in Fig. 1 (b), which shows example images of the subscore Nec Amount with scores 0 to 4 (left to right) based on the PWAT wound grading rubric for necrotic tissue. As shown in Fig. 1 (b), wounds with different Nec Amount scores are quite challenging to distinguish visually, as well as grading other PWAT subscores that are based on the type or amount of a specific type of tissue shown in the wound images. The Supplementary Materials shows a brief description of the fine-grained image classification problem in computer vision.

E. Our contributions

There are three main contributions in this paper:

1) We innovatively adapted a semi-supervised learning method inspired mainly by the rotation degree Self-Supervised Learning [42] and the SESEMI method [43], and partially from FixMatch method [44] to the problem of generating synthetic wound images.

2) We proposed a deep learning framework that innovatively integrated the Progressive Multi-Granularity (PMG) mechanism and the semi-supervised learning method with the EfficientNet B0 neural network to comprehensively predict all 8 PWAT sub-scores, which solved challenging fine-grained image task of recognizing clinically-important grades of wound. The semi-supervised learning method worked as a new way of data augmentation to solve our *WoundNet*'s problem of insufficient and imbalanced data and PMG improved the prediction of PWAT-based wound scores, a fine-grained image classification problem.

3) We performed rigorous evaluations and comparison of our proposed model and its variants. Our results show that our proposed semi-supervised learning aiding Progressive Multi-Granularity mechanism based EfficientNet B0 architecture achieves classification accuracies and F1 scores of almost 90% for fine-grained classification of all 8 PWAT sub-scores with more than 7% improvement of our previous model [15]. Our approach was compared to various state-of-the-art baseline CNN models, data augmentation methods and fine-grained image classification techniques. We also demonstrate that state-of-the-art GAN-based data augmentation methods including pix2pixHD [48] and semi-supervised GANs [49] did not improve PWAT wound image classification performance, an unexpected finding. Facilitated by the proposed research, the performance of our deep learning model for PWAT wound assessment system made our wound assessment system clinically usable.

II. MATERIALS AND METHODS

This section introduces Semi-Supervised PMG EfficientNet (SS-PMG-EfficientNet), our proposed deep learning architecture for estimating PWAT subscores for chronic wounds. Sub-Section II-A describes *WoundNet*, our chronic wound dataset and the secondary DFUC 2021 dataset [33] [47]. Sub-section II-B describes our wound assessment system and the deep learning architecture. Sub-section II-C describes the PMG (Progressive Multi-Granularity) mechanism. Sub-section II-D describes the semi-supervised learning approach. Sub-section II-E describes our rigorous evaluation including definitions of our evaluation metrics and experiments conducted.

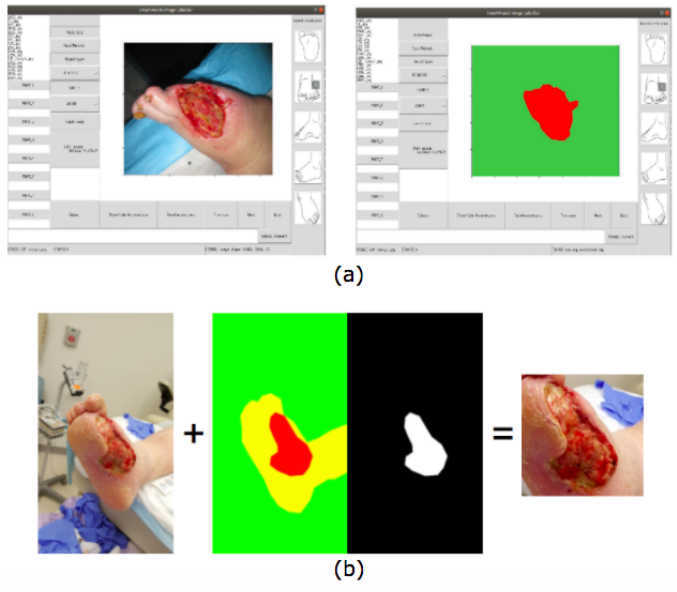
The supplementary materials introduces the detail for data augmentation using pix2pixHD GANs [48] and using semi-supervised GANs [49]. The supplementary materials also includes details of the MMAL ResNet50 architecture [50].

A. WoundNet Dataset, DFUC 2021 dataset and Preprocessing

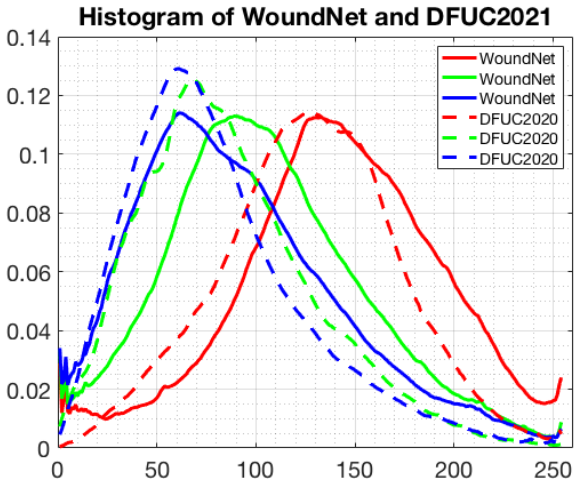
1) WoundNet dataset:

There are 1639 images in our *WoundNet* chronic wound image dataset. 1323 of them were provided by the University of Massachusetts Memorial Medical Center from their archives. 114 of them were captured by our research group using a mechanical wound imaging box that ensured consistent imaging distance, angle and lighting. 202 of them were collected from the Internet using an image search. All images in *WoundNet* were labeled with their 8 PWAT sub-scores calculated based on the PWAT subscore scoring instruction. PWAT sub-scores 1 through 7 were assigned values 0, 1, 2, 3 and 4 and were modeled as 5-class classification problems. PWAT sub-score 8 can only be assigned values 0, 1 and 2 and was modeled as a 3-class classification problem. The number of *WoundNet* images for each of the 8 PWAT sub-scores are summarized in Table Ib, which shows that *WoundNet* has problems of insufficient images corresponding to some PWAT scores and consequently, imbalance in the distribution of PWAT subscores.

Wound image pre-processing: Some of the original *WoundNet* images were poorly captured, which challenged image analyses. For instance, some images had small wounds with large background areas. Some images had large wounds or were mostly occupied by wound and skin area. The *WoundNet* corpus was pre-processed using the following steps in order to make the wound images more consistent and then the *WoundNet* was used for all PWAT sub-scores classification research. First, the wounds and the skins were segmented out of the whole images with our previously developed wound annotation app [51], which is shown in Fig. 2a.a. This segmentation app applied the deep extreme cuts algorithm [52] that ensured consistent, systematic wound image segmentation. The



(a) a: Example of our wound annotation app; b: Example of wound segmentation mask



(b) Histogram showing the percentage of each Red, Green, Blue value in the *WoundNet* and DFUC 2021 datasets

Fig. 2. our wound annotation app, wound segmentation mask and the Histogram of *WoundNet* and DFUC 2021

segmentation mask of the wound image was then utilized as a bounding box to crop the skin and the wound area out from the original wound image. The cropped wound images were resized to a dimension of $512 \times 512 \times 3$. Fig. 2a.b shows an example original wound, its segmentation mask, and cropped image.

2) DFUC 2021 dataset:

The Diabetic Foot Ulcers Grand Challenge (DFUC) 2021 dataset [33] [47] contained DFU images collected from the Lancashire Teaching Hospital with the approval for research from the UK National Health Service (NHS) Research Ethics Committee (REC) (NHS REC reference no. 15/NW/0539). About 3000 images for each class were captured in stable room lighting with a distance of 30-40 cm to the plane of the foot ulcer. The images were acquired by a podiatrist and a diabetic ulcers consultant physician, both with more than 5 years of professional experience, who produced ground truth labels on infection and ischemia status. The size of the original DFU images varies between 1600×1200 and 3648×2736 and they were resized to a dimension of 640×640, which was suitable for deep learning in optimizing performance and minimizing computational costs.

The workload of annotating all images in the DFUC 2021 dataset with all 8 PWAT subscores would have been very large. Consequently, the DFUC 2021 dataset is utilized without PWAT subscores labels, as a secondary dataset, by our model. Therefore, a semi-supervised learning method was applied by our deep learning architecture to utilize both the labeled *WoundNet* dataset and unlabeled DFUC 2021 dataset, which is described in more detail in Sub-section II-D. Using this approach, our proposed deep learning model was trained on our own *WoundNet* dataset as well as the DFUC 2021 dataset so that the model's performance could be further improved. Fig. 2b is a histogram showing the percentage of each Red, Green, Blue value in the *WoundNet* and DFUC 2021 datasets, which demonstrates that the distributions of pixel values for these two datasets are similar.

B. Overview of proposed SS-PMG-EfficientNet wound assessment system

Semi-Supervised PMG EfficientNet (SS-PMG-EfficientNet), our deep learning architecture for accessing all 8 PWAT subscores, is shown in Fig. 3(a). This architecture is composed of 3 main components: the semi-supervised learning component, the PMG (Progressive Multi-Granularity) component and the baseline deep learning model: EfficientNet B0.

The PMG mechanism and semi-supervised learning component are two separate techniques from different prior research and they can both improve the deep learning model's performance. The PMG mechanism was designed specifically for fine-grained image classification problems. It can be built on top of any state-of-the-art baseline CNN models. The PMG mechanism first focuses on discriminative information in local regions then progressively training on higher stages and global structures eventually. The semi-supervised learning method was originally designed for the model to train on both the small labeled subset and large unlabeled subset sampled from the same distribution of images. In our research, this semi-supervised learning method was modified to train the baseline model on our labeled *WoundNet* and DFUC 2021 dataset [33] [47] that is not labeled with PWAT subscores. Detailed descriptions of the PMG mechanism and semi-supervised learning are presented in subsequent Sub-sections.

These two techniques can both improve the baseline model EfficientNet B0's performance, which was proven via extensive evaluation of PMG EfficientNet and Semi-Supervised EfficientNet. The semi-supervised learning component and the PMG component were integrated and assembled innovatively with the EfficientNet B0 model and became SS-PMG-EfficientNet. It was also tested and evaluated extensively to show that SS-

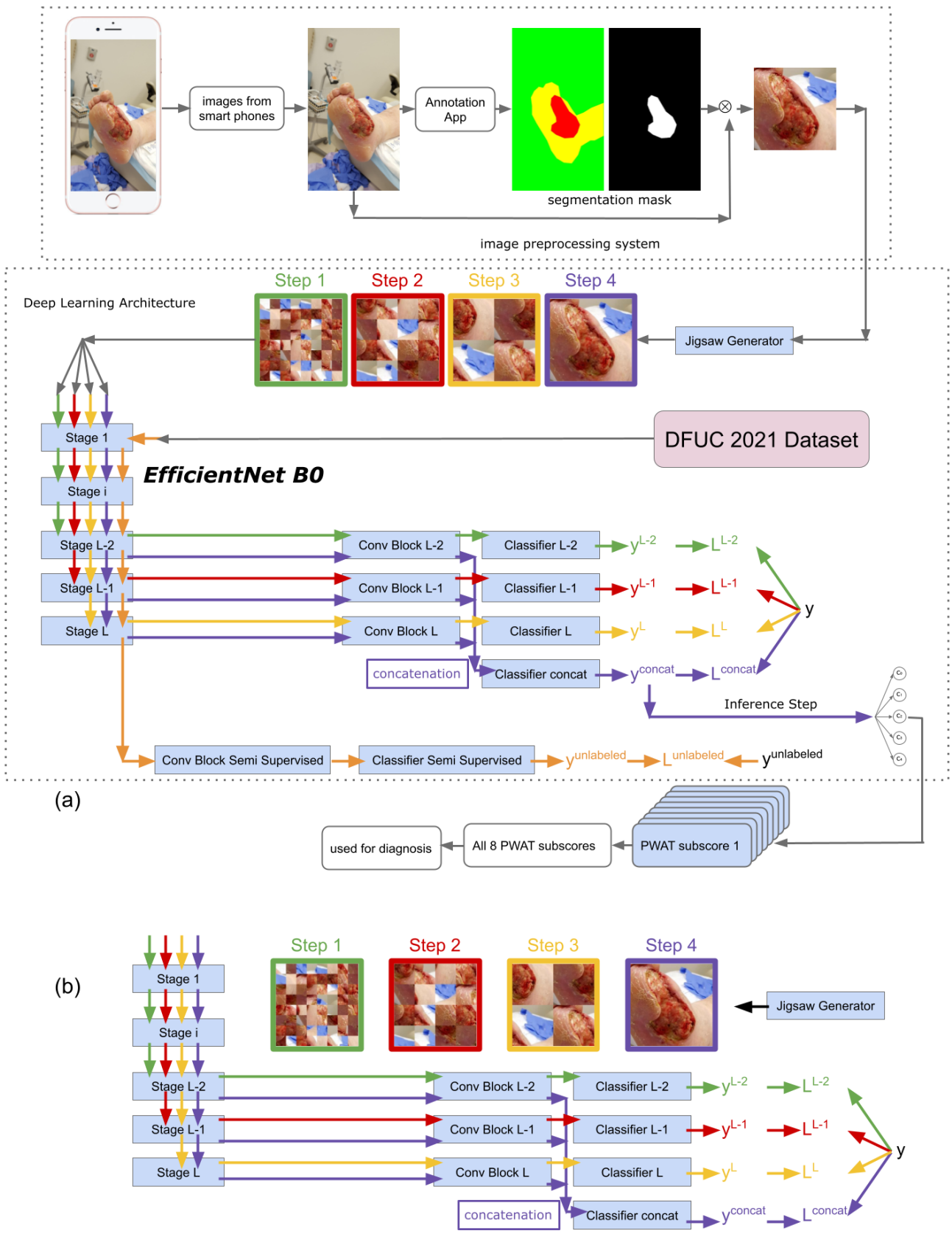


Fig. 3. (a) Our chronic wound image analysis system including annotation app, segmentation and our novel semi-supervised PMG EfficientNet (SS-PMG-EfficientNet); (b) PMG (Progressive Multi-Granularity) mechanism

PMG-EfficientNet outperformed both PMG EfficientNet and Semi-Supervised EfficientNet, which illustrated that the PMG mechanism and semi-supervised methods worked simultaneously in SS-PMG-EfficientNet to boost the model for the best performance.

C. Progressive Multi-Granularity (PMG) mechanism

1) Network Architecture:

The PMG (Progressive Multi-Granularity) mechanism [45] can be implemented as a feature extractor with any state-of-the-art image analysis models such as ResNet [53]. Suppose F is the feature extractor with L stages. Its intermediate stages have output feature-map: $F^l \in \mathbb{R}^{H_l \times W_l \times C_l}$. Here H_l, W_l, C_l are the height, width and number of channels of the feature map at l-th stage, $l = 1,2,...,L$. The next step is to calculate the classification loss on the feature-map from different intermediate stages. The new convolution block H^l_{conv} takes l-th intermediate stage output, F^l, as input. Its output was reduced to a vector representation:

$$V^l = H^l_{conv}(F^l) \tag{1}$$

Then, a classification module H^l_{class} with two fully-connected stage, Batchnorm [54] and Elu [55], calculates the probability distribution for each classes for the l-th stage:

$$y^l = H^l_{class}(V^l) \tag{2}$$

After calculating the last S stages: $l = L, L-1, ..., L-S+1$, the outputs from them are concatenated as:

$$V^{concat} = concat[V^{L-S+1}, ..., V^{L-1}, V^L] \tag{3}$$

It is then input into a classifier:

$$y^{concat} = H^{concat}_{class}(V^{concat}) \tag{4}$$

2) Progressive Training:

In traditional CNN models, training the entire network directly in traditional CNN models means learning all the granularities simultaneously. In progressive training, the low stage is trained first and then new stages are added for training progressively. The PMG mechanism allows the network to first exploit discriminative information from local details such as textures because the low stage has a limited receptive field and representation ability. When the features are gradually input into higher stages, the model can locate discriminative information from local details to global structures.

The outputs from each stage and the output from the concatenated features are input into the cross entropy (CE) L_{CE}. The loss between ground truth label y and prediction probability distribution is calculated as

$$L_{CE}(y^l, y) = -\sum_{i=1}^{m} y^l_i \times log(y^l_i) \tag{5}$$

and

$$L_{CE}(y^{concat}, y) = -\sum_{i=1}^{m} y^{concat}_i \times log(y^{concat}_i) \tag{6}$$

In each training iteration, the data d will be used for $S+1$ times but only to obtain the output for each stage in each time. All parameters used in each stage are updated even though they may already be updated in the previous stages, which helps all stages in the model work together.

3) Jigsaw Puzzle Generator:

The notion of Jigsaw Puzzle is used here to generate input images for different stages of progressive training. It generates different granularity regions so that the model can learn the corresponding granularity level's information which is specific at each training step. The input image $d \in \mathbb{R}^{3 \times W \times H}$ is equally split into $n \times n$ patches with $3 \times \frac{W}{n} \times \frac{H}{n}$ dimensions. The patches are shuffled randomly and merged together into a new image $P(d,n)$ so that the hyper-parameter n controls the patches' granularities.

The correct hyper-parameter n for each stage should guarantee that the patches' size should be smaller than the receptive field at the corresponding stage and the patches' size should increase proportionately as the receptive fields of the stages increase. For the l-th stage, n is chosen as:

$$n = 2^{L-l+1} \tag{7}$$

During training, the jigsaw puzzle generator augments training data batch d to generate several augmented batches $P(d,n)$, which all have the same label y . The batch $P(d,n)$ with $n = 2^{L-l+1}$ is input to the l-th stage which generates the output y^l, then all the parameters used in this process will be updated in this propagation. All the jigsaw generator augmented data batches are input sequentially into the network by $S+1$ steps. The training procedure is shown in Fig. 3(b).

4) Inference:

During inference, the original images are input into the trained model without the jigsaw puzzle generator. To only utilize y^{concat} for prediction, the FC layers for the other three stages are removed and the final result C_1 is:

$$C_1 = argmax(y^{concat}) \tag{8}$$

The prediction from each stage has unique and complementary information from a specific granularity. To obtain a better performance, all outputs are combined together with equal weights and the multi-output combined prediction C2 is:

$$C_2 = argmax(\sum_{l=L-S+1}^{L} y^l + y^{concat}) \tag{9}$$

D. semi-supervised learning

The semi-supervised learning method applied in our research was inspired mainly from the rotation degree Self-Supervised Learning [42] and the SESEMI method [43], and partially from FixMatch method [44]. It is a simple but effective algorithm for semi-supervised image classification via self-supervision. The dataset for the semi-supervised learning method consists of pairs of images and labels $(x,y) \in S_L$ and unlabeled images $x \in S_U$. Usually S_L and S_U are sampled from the same distribution $p(x)$ and S_L is S_U'subset with labels. However, S_L is our *WoundNet* dataset and S_U is the DFUC 2021

dataset [33] [47] in our case, as mentioned in Sub-section II-A. It is possible to sample S_L from $p(x)$ but sample S_U from $q(x)$, a different yet related distribution [56]. This semi-supervised learning method trains a prediction function $f_\theta(x)$ with parameter θ on a combination of S_L and S_U to obtain better model performance than training on S_L alone. During the training process, two batches of data are sampled from the labeled dataset S_L and unlabeled dataset S_U separately in each step:

$$s_L = b(x_i \in S_L) \tag{10}$$

$$s_U = b(x_j \in S_U) \tag{11}$$

Then they are input into the shared baseline model $f_\theta(x)$, which is EfficientNet B0 in our case. The labeled batch s_L and the unlabeled batch s_U are input into $f_\theta(x)$ so that its softmax layer generates prediction vectors from them respectively:

$$z_i = f_\theta(s_L) \tag{12}$$

$$z_j = f_\theta(s_U) \tag{13}$$

The ground truth labels y_i are used for computing the supervised cross-entropy loss $L_{labeled}(y_i, z_i)$. The DFUC 2021 dataset's label is considered as the dataset S_U's label, which is used as the proxy labels y_j to compute the cross-entropy loss $L_{unlabeled}(y_j, z_j)$ for the unsupervised cross-entropy loss.

$$L_{labeled}(y_i, z_i) = -\frac{1}{|s_L|} \sum_{i \in s_L} \sum_{k \in K} y_{ik} log(z_{ik}) \tag{14}$$

$$L_{unlabeled}(y_j, z_j) = -\frac{1}{|s_U|} \sum_{j \in s_U} \sum_{t \in T} y_{jt} log(z_{jt}) \tag{15}$$

The final loss function is defined as the weighted sum of the supervised cross-entropy loss and the unsupervised cross-entropy loss:

$$L_{final} = L_{labeled}(y_i, z_i) + \omega L_{unlabeled}(y_j, z_j) \tag{16}$$

The parameter θ will be updated in backpropagation after minimizing the final loss function L_{final}. The unsupervised cross-entropy loss $L_{unlabeled}(y_j, z_j)$ can be considered as a regularization term in the final loss function and $\omega > 0$ is a regularization hyperparameter that controls the relative contribution of unsupervised learning in the semi-supervised learning process.

E. Evaluation

1) Evaluation Metrics:

Our deep learning architecture was evaluated using metrics: testing accuracy, weighted F1 score, multi-class sensitivity and multi-class specificity.

Synthetic wound images generated by GANs methods were evaluated using the FID score. The Frechet Inception Distance (FID) is a metric that evaluates the quality of synthesis images from generative adversarial networks (GANs) [57]. The FID compares the distribution of synthesis images and the distribution of real images.

The detailed description and the equations of the evaluation metrics are shown in the Supplementary Materials.

2) Baseline models:

ResNet50 [53] and EfficientNet B0 [46] were utilized as baseline models for the comparison of data augmentation methods. Detailed descriptions of ResNet50 and EfficientNet B0 can be found in the Supplementary Materials. The Patch Attention DenseNet deep learning model [15] was also used in the comparison of data augmentation using GANs. ResNet50 and EfficientNet B0 were also used as baseline models for the comparison of wound assessment deep learning architectures.

F. Experiments

1) Hardware, Software and Hyperparameters:

All the experiments were run on the same Ubuntu system desktop with an NVIDIA GTX 1080 Ti GPU. PyTorch was the library used for running the deep learning models. The regularization hyperparameter ω in the semi-supervised learning method mentioned in Sub-section II-D was set to $\omega = 0.8$ after experiments to evaluate different values of ω between 0.5 to 1.2, which revealed that the model performed best when ω was set to values between 0.6 to 1.0. The learning rate for training SS-PMG-EfficientNet was set to 0.0004 after experiments using different values.

2) Experiment 1, comparison of SS-PMG-EfficientNet, the PMG EfficientNet and the Semi-Supervised EfficientNet on all 8 PWAT sub-scores:

The goal of this experiment was to compare our proposed model SS-PMG-EfficientNet to variants PMG EfficientNet (no Semi-Supervised (SS) Learning) and Semi-Supervised EfficientNet (no PMG) on all 8 PWAT subscores and demonstrate the non-trivial contribution of the Semi-Supervised Learning (SS) and PMG to our overall architecture. To facilitate a fair comparison, SS-PMG-EfficientNet, the PMG EfficientNet, the Semi-Supervised EfficientNet and EfficientNet B0 were trained and tested in the same way. 3 training and testing sets were generated for each PWAT subscore with no overlapping images among the 3 testing sets and each of these testing sets contained 10% images of the entire dataset. All four deep learning architectures were trained and tested on these three training and testing sets for all 8 PWAT subscores.

Due to various randomness in the deep learning model, such as the random weight initialization and batch gradient descent during training, the training and testing results can be different and unstable. Each training and testing set was trained 2 times to evaluate whether the models were stable, generating 6 training results for each PWAT subscore for each deep learning architecture.

EfficientNets' run times are very fast and EfficientNet B0 is still very fast even though integrated with PMG mechanism component and semi-supervised learning mechanism. This advantage made it possible to run and test all 8 PWAT subscores 6 times with 3 different EfficientNet B0 based architectures.

3) Experiment 2, comparison of various deep learning architectures and data augmentation methods on PWAT subscore 2. Depth:

The goal of this experiment was to compare our proposed data augmentation method (PMG) and SS-EfficientNet classification architecture to a comprehensive set of baseline methods for augmenting and predicting the Depth PWAT subscore. Due to time and resource constraints, we selected to provide in-depth comparisons on only the depth sub-score but believe that the results for depth are representative of all PWAT sub-scores. All baselines deep learning architectures and data augmentation methods were trained and tested on the PWAT subscore 2. Depth with a test set of 10% of the entire dataset. These deep learning architectures include EfficientNet B0, the Semi-Supervised EfficientNet, the PMG EfficientNet and SS-PMG-EfficientNet, as well as the PMG ResNet50 [45] and MMAL ResNet50 [50]. EfficientNet B0, PMG ResNet50 [45] and MMAL ResNet50 [50] were first trained on the PWAT subscore 2. Depth to show whether these architectures had good performance on the *WoundNet* dataset. The Semi-Supervised EfficientNet, the PMG EfficientNet and the SS-PMG-EfficientNet were also trained on the PWAT subscore 2. Depth to show their performance.

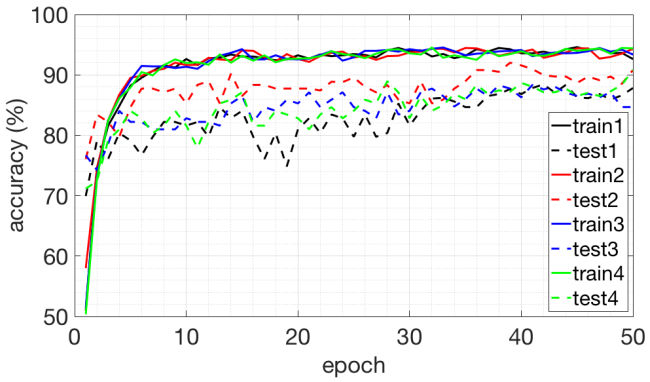
The data augmentation methods include applying pix2pixHD [48] and semi-supervised GANs [49] for the data augmentation of the *WoundNet* dataset. The pix2pixHD model was first trained with the *WoundNet* dataset and it was used for generating synthesis chronic wound images. 2000 synthesis chronic wound images were generated and 1000 of them were labeled with PWAT subscore 2. Depth. The labeled synthesis chronic wound images were added to the training set of the original *WoundNet* dataset and this augmented *WoundNet* dataset was trained with both the Patch Attention DenseNet from our previous work [15] and EfficientNet B0. The semi-supervised GAN model was trained with *WoundNet* as the supervised dataset and DFUC 2021 dataset [33] [47] as the unsupervised dataset.

Although it is more convincing to test and compare different architectures and methods with more PWAT subscores, it is time-consuming to run these many experiments. On the other hand, one PWAT subscore can indicate the general performance of different architectures on the *WoundNet* dataset. Therefore, testing different architectures and methods on the most important subscore 2. Depth enabled discovery of the best architectures and methods for estimating PWAT subscores without loss of generality.

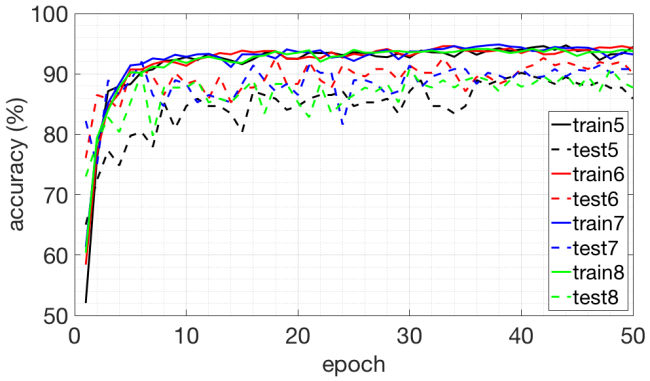
III. RESULTS

A. Training and testing accuracy trajectories

A sample of the trajectories of the training and testing set accuracies for all 8 PWAT sub-scores are shown in Fig. 4a and 4b. The training and testing accuracy trajectories plotted are from the best set results of the 6 results mentioned in II-F2. The number index i represents the ith PWAT sub-score in these two figures. For example, train1 is the training accuracy for the subscore 1. Size. The training and testing accuracies converged and stabilized after about 35 to 45 epochs. The differences between



(a) Training and testing accuracy trajectory (Example 1)



(b) Training and testing accuracy trajectory (Example 1)

(c) Examples of synthesis wound images from GANs method

Fig. 4. Training and testing accuracy trajectory and synthesis wound image examples from GANs method

Fig. 5. Examples of misclassified images for each PWAT sub-score. Row 1 to 8 is PWAT sub-scores 1 to 8.

training and testing accuracies were relatively small indicating that the model did not overfit and generalized well to the test set.

B. Model performance for predicting all 8 PWAT sub-scores

The mean and standard deviation of testing accuracy on applying SS-PMG-EfficientNet, the PMG EfficientNet, the Semi-Supervised EfficientNet, EfficientNet and the Patch Attention DenseNet from our previous work [15] to all 8 PWAT subscores are shown in Table IIIa for comparison. As shown in Table IIIa, the means of testing accuracy of SS-PMG-EfficientNet for all 8 PWAT subscores achieved the best results among all these 5 deep learning architectures for comparison. The PMG EfficientNet achieved the second-best results and the Semi-Supervised EfficientNet achieved the third-best results in terms of the means of testing accuracy for all 8 PWAT sub-scores. The results of EfficientNet B0 and the Patch Attention DenseNet were close to each other generally and EfficientNet B0 had better results in 6. Gran Amount, 7. Edges and 8. Skin while the Patch Attention DenseNet had better results in Nec Type and Gran Type.

Table IIIb, IIIc, IIId and IIIe showed the mean and standard deviation of testing accuracy, weighted F1 scores, Sensitivity and Specificity for all 8 PWAT subscores of SS-PMG-EfficientNet, PMG EfficientNet, Semi-Supervised EfficientNet and the Patch Attention DenseNet.

The means of weighted F1 score of SS-PMG-EfficientNet achieved the best results for all 8 PWAT subscores when compared to the other three deep learning architectures. The means of sensitivity of SS-PMG-EfficientNet also achieved the best results for 7 PWAT subscores except for 5. Gran Type, which was 0.8789 for Semi-Supervised PMG EfficientNet and 0.8795 for PMG EfficientNet. The mean of the Specificity of SS-PMG-EfficientNet also achieved the best results for all 8 PWAT subscores compared to other 3 deep learning architectures. The standard deviation of the testing accuracy, weighted F1 scores, Sensitivity and Specificity for all 8 PWAT subscores of SS-PMG-EfficientNet were relatively small. The Sensitivity for 8. Skin had the largest standard deviation of SS-PMG-EfficientNet with only 0.0447. This demonstrates that SS-PMG-EfficientNet was relatively stable on the *WoundNet* dataset.

Generally, the scores of SS-PMG-EfficientNet were the highest, while the scores of PMG EfficientNet were the second highest and scores of Semi-Supervised EfficientNet were the third highest, which were all higher than the Patch Attention DenseNet model, the previous state-of-the-art for wound grading. There was a 7% improvement between SS-PMG-EfficientNet and Patch Attention DenseNet in testing accuracy, weighted F1 scores and Sensitivity. However, it can be observed that the sensitivity of 3. Nec Type had the lowest scores of SS-PMG-EfficientNet, PMG EfficientNet and Semi-Supervised EfficientNet with 0.8047 as the highest one. It was obviously lower than other PWAT subscores' sensitivity and was only a small improvement from Patch Attention DenseNet with sensitivity in 3. Nec Type: 0.7711. On the other hand, the improvement of specificity between SS-PMG-EfficientNet and Patch Attention DenseNet were relatively small, which was likely because the specificity of Patch Attention DenseNet were already high with little room for improvement. Examples of misclassified images for each PWAT sub-score are shown in Fig. 5.

C. Box plot showing k-fold cross-validation results of all 8 PWAT sub-scores

Figure 6 shows the boxplots of all 8 PWAT subscores from the results of SS-PMG-EfficientNet, the PMG EfficientNet and

(a) Comparison of **Testing Accuracy** from different deep learning architecture for predicting all 8 PWAT sub-scores

	Semi-Supervised PMG EfficientNet		**PMG EfficientNet**		**Semi-Supervised EfficientNet**		**EfficientNet**		**Patch Attention DenseNet**	
PWAT subscore	mean	std	mean	std	mean	std	mean	std	mean	std
1. Size	**0.8753**	**0.0071**	0.8609	0.0177	0.8384	0.0163	0.8068	0.0146	0.8098	0.0132
2. Depth	**0.9039**	0.0232	0.8916	**0.0071**	0.8651	0.0103	0.8267	0.0237	0.8281	0.0185
3. Nec Type	**0.8793**	**0.0035**	0.8753	0.0071	0.8569	0.0158	0.8114	0.0176	0.8244	0.0090
4. Nec Amount	**0.8875**	0.0187	0.8569	**0.0035**	0.8517	0.0180	0.8114	0.0383	0.8098	0.0309
5. Gran Type	**0.8916**	0.0094	0.8793	**0.0071**	0.8722	0.0264	0.8129	0.0097	0.8269	0.0153
6. Gran Amount	**0.8998**	0.0248	0.8896	**0.0123**	0.8671	0.0200	0.8466	0.0371	0.8330	0.0235
7. Edges	**0.9018**	0.0123	0.8834	**0.0061**	0.8731	0.0205	0.8497	0.0146	0.8256	0.0093
8. Skin	**0.9018**	**0.0061**	0.8916	0.0128	0.8497	0.0075	0.8328	0.0153	0.8208	0.0222

(b) Results of **Semi-Supervised PMG EfficientNet** for predicting all 8 PWAT sub-scores

	F1 score		**Sensitivity**		**Specificity**	
PWAT subscore	mean	std	mean	std	mean	std
1. Size	**0.8755**	**0.0067**	**0.8674**	**0.0065**	**0.9673**	**0.0022**
2. Depth	**0.9017**	0.0238	**0.8656**	0.0250	**0.9666**	0.0057
3. Nec Type	**0.8767**	**0.0052**	**0.8047**	0.0406	**0.9679**	**0.0006**
4. Nec Amount	**0.8862**	0.0170	**0.8539**	**0.0253**	**0.9679**	0.0063
5. Gran Type	**0.8915**	0.0093	0.8789	**0.0139**	**0.9695**	0.0025
6. Gran Amount	**0.8998**	0.0247	**0.8662**	0.0414	**0.9705**	0.0080
7. Edges	**0.9006**	0.0088	**0.8838**	0.0352	**0.9607**	0.0054
8. Skin	**0.9013**	**0.0052**	**0.8542**	0.0447	**0.9302**	**0.0056**

(c) Results of **PMG EfficientNet** for predicting all 8 PWAT sub-scores

	F1 score		**Sensitivity**		**Specificity**	
PWAT subscore	mean	std	mean	std	mean	std
1. Size	0.8607	0.0172	0.8593	0.0264	0.9626	0.0053
2. Depth	0.8887	0.0084	0.8535	**0.0235**	0.9643	0.0028
3. Nec Type	0.8706	0.0074	0.7998	**0.0096**	0.9671	0.0023
4. Nec Amount	0.8547	**0.0038**	0.8400	0.0361	0.9600	**0.0016**
5. Gran Type	0.8795	**0.0078**	**0.8795**	0.0242	0.9664	**0.0021**
6. Gran Amount	0.8889	**0.0111**	0.8453	**0.0207**	0.9639	**0.0054**
7. Edges	0.8826	**0.0043**	0.8506	**0.0133**	0.9521	**0.0016**
8. Skin	0.8905	0.0124	0.8378	0.0229	0.9269	0.0089

(d) Results of **Semi-Supervised EfficientNet** for predicting all 8 PWAT sub-scores

	F1 score		**Sensitivity**		**Specificity**	
PWAT subscore	mean	std	mean	std	mean	std
1. Size	0.8423	0.0127	0.8510	0.0081	0.9589	0.0025
2. Depth	0.8649	**0.0070**	0.8278	0.0345	0.9602	**0.0027**
3. Nec Type	0.8525	0.0165	0.7600	0.0144	0.9615	0.0051
4. Nec Amount	0.8555	0.0191	0.8258	0.0482	0.9611	0.0069
5. Gran Type	0.8816	0.0185	0.8691	0.0202	0.9675	0.0044
6. Gran Amount	0.8749	0.0239	0.8305	0.0246	0.9631	0.0074
7. Edges	0.8753	0.0265	0.8030	0.0342	0.9518	0.0066
8. Skin	0.8507	0.0058	0.8104	**0.0049**	0.9000	0.0073

(e) Results of **Patch Attention DenseNet** for predicting all 8 PWAT sub-scores

	F1 score		**Sensitivity**		**Specificity**	
PWAT subscore	mean	std	mean	std	mean	std
1. Size	0.8085	0.0137	0.8154	0.0244	0.9503	0.0034
2. Depth	0.8285	0.0190	0.8156	0.0440	0.9509	0.0056
3. Nec Type	0.8226	0.0096	0.7711	0.0336	0.9533	0.0028
4. Nec Amount	0.8072	0.0307	0.7539	0.0447	0.9498	0.0080
5. Gran Type	0.8263	0.0150	0.8354	0.0306	0.9534	0.0042
6. Gran Amount	0.8324	0.0229	0.8137	0.0227	0.9512	0.0062
7. Edges	0.8269	0.0107	0.8157	0.0617	0.9421	0.0062
8. Skin	0.8189	0.0270	0.7548	0.0771	0.8783	0.0243

TABLE III. Results from the SS-PMG-EfficientNet, the PMG EfficientNet, the Semi-Supervised EfficientNet, EfficientNet and the Patch Attention DenseNet

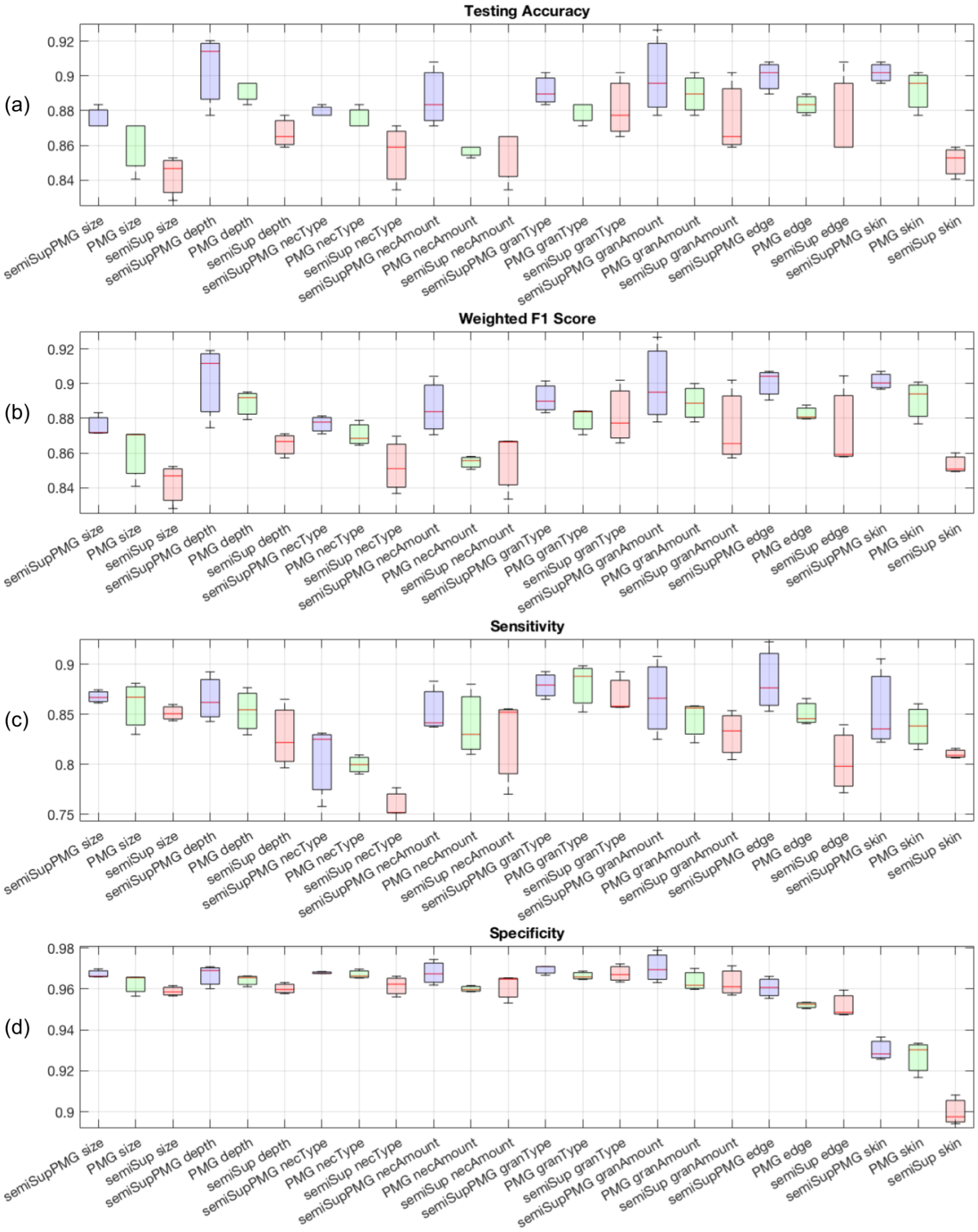


Fig. 6. Boxplots of all 8 PWAT sub-scores from the results of SS-PMG-EfficientNet, the PMG EfficientNet and the Semi-Supervised EfficientNet

the Semi-Supervised EfficientNet. For each PWAT subscore from each deep learning model, there were 6 results, as mentioned in Sub-section II-F and these 6 results were made into the boxplots. There are four sub-boxplots: (a) Testing Accuracy, (b) Weighted F1 Score, (c) Sensitivity and (d) Specificity.

Each of the sub boxplots contains 24 boxes drawn from the results of these 3 deep learning model on all 8 PWAT subscores for that particular metrics. The blue, green and red boxes represent results from SS-PMG-EfficientNet, the PMG EfficientNet and the Semi-Supervised EfficientNet. It can be observed that the results from SS-PMG-EfficientNet generally had the highest average scores, while the results from the PMG EfficientNet and the Semi-Supervised EfficientNet had the second and third highest average scores. On the other hand, the distribution range of results for all 4 metrics from all 3 different models varied on different PWAT subscores. Test accuracy and F1 Scores for all 8 PWAT subscores from these 3 models were high relatively and the difference between different subscores were small from all 3 models.

With regards to sensitivity, the difference between different PWAT subscores from all 3 models were relatively large. The sensitivity for subscore 3 (Nec Type) were lower than those of other subscores' for all 3 models, especially Semi-Supervised EfficientNet, The specificity of all 3 models were relatively high for all PWAT subscores except subscore 8. Skin.

D. Comparison of variations of our proposed architecture and baseline methods with PWAT subscore 2. Depth

The comparison of results from different deep learning architectures and data augmentation methods we used on the PWAT subscore 2. Depth is shown in Table IVa. Semi-Supervised PMG EfficientNet achieved the highest mean of testing accuracy: 0.9039. SS-PMG-EfficientNet, the PMG EfficientNet and the Semi-Supervised EfficientNet were all able to complete model training within 30 minutes with relatively high mean testing accuracy, due to the very fast run time of EfficientNet. PMG ResNet50 achieved higher mean of testing accuracy when comparing to MMAL ResNet50, which came from the result of exploring and testing novel fine-grained image classification mechanism. Consequently, we selected the PMG mechanism when designing and developing our deep learning architecture.

PMG ResNet50 achieved a relatively high testing accuracy mean: 0.8671 but PMG EfficientNet achieved a better mean of testing accuracy: 0.8916, which demonstrated that choosing to integrate the PMG mechanism into EfficientNet B0 was the correct design choice. ResNet50 was the first baseline model we used in the research of estimating PWAT subscores and had the mean of testing accuracy: 0.7915, which could be considered as the PWAT subscore performance baseline and could indicate how much improvement we made with the newly developed deep learning architectures. By researching and studying the *WoundNet* dataset and deep learning related works such as state-of-the-art baseline deep learning models, data augmentation and fine-grained image classification techniques, the newest architecture we designed, Semi-Supervised PMG EfficientNet, showed significant improvement. SS-PMG-

(a) Model comparison for PWAT subscore: 2. Depth.

deep learning architecture	input image	training time	accuracy mean	accuracy std
Our newly proposed method				
Semi-Supervised PMG EfficientNet	224	30 min	**0.9039**	0.0232
PMG EfficientNet	224	25 min	0.8916	**0.0071**
Semi-Supervised EfficientNet	224	25 min	0.8651	0.0103
State-of-the-art deep learning architecture from related work				
PMG ResNet50	448	4 hours	0.8671	0.0128
MMAL ResNet50	448	4 hours	0.8466	0.0184
Our previous used method				
Patch Attention DenseNet (without any kind of data augmentation)	512	3.5 hours	0.8281	0.0185
Bilinear CNN	256	N/A	0.8336	0.0179
Baseline CNN image classification architecture (without any kind of data augmentation)				
ResNet50	224	40 min	0.7915	0.0132
EfficientNet B0	224	**20 min**	0.8267	0.0237
GANs based methods				
data augmenting with pix2pixHD synthesis images				
Patch Attention DenseNet	512	3.5 hours	about 60%	N/A
EfficientNet B0	512	20 min	about 60%	N/A
semi-supervised GAN				
ResNet50 as discriminator	224	60 min	about 50%	N/A
EfficientNet B0 as discriminator	224	30 min	about 50%	N/A

(b) Confusion Matrices1

1. Size		actual class				
		0	1	2	3	4
prediction class	0	11	1	0	0	0
	1	1	46	2	1	0
	2	2	5	37	2	0
	3	0	0	4	29	1
	4	0	0	0	2	19

2. Depth		actual class				
		0	1	2	3	4
prediction class	0	16	0	1	1	0
	1	0	9	1	0	0
	2	0	5	83	4	0
	3	0	0	2	32	0
	4	0	0	0	0	9

3. Nec Type		actual class				
		0	1	2	3	4
prediction class	0	54	6	1	0	2
	1	2	33	0	0	2
	2	0	1	24	2	0
	3	0	0	0	3	0
	4	0	2	1	0	30

4. Nec Amount		actual class				
		0	1	2	3	4
prediction class	0	26	0	0	0	2
	1	1	19	0	0	1
	2	0	0	12	0	0
	3	4	1	1	24	2
	4	2	0	0	5	63

(c) Confusion Matrices2

5. Gran Type		actual class				
		0	1	2	3	4
prediction class	0	17	0	1	0	1
	1	0	31	2	0	0
	2	0	5	59	2	2
	3	0	0	0	7	0
	4	0	0	3	0	33

6. Gran Amount		actual class				
		0	1	2	3	4
prediction class	0	13	0	0	0	3
	1	0	9	1	1	0
	2	0	0	17	3	0
	3	0	0	0	21	1
	4	1	1	0	1	91

7. Edges		actual class				
		0	1	2	3	4
prediction class	0	14	0	0	0	0
	1	1	25	8	0	0
	2	0	4	97	1	0
	3	0	0	2	10	0
	4	0	0	0	0	1

8. Skin		actual class				
		0	1	2	3	4
prediction class	0	43	10	0		
	1	5	90	1		
	2	0	1	13		
	3					
	4					

TABLE IV. Model Comparison for PWAT subscore: 2. Depth and Confusion Matrices

EfficientNet achieved a mean testing accuracy of 0.9039, outperforming the ResNet50 baseline model by more than 10%.

The data augmentation methods using pix2pixHD and semi-supervised GAN decreased the model's performance and resulted in testing accuracy with only about 60%. By examining the method and testing the model multiple times, it was observed that the performance of model with these two data augmentation methods remained the same, which was counter intuitive. The Frechet Inception Distance (FID) between the GAN-synthesized images and the *WoundNet* was 70.6107, which is acceptable and indicates that the quality of the synthesis wound images was acceptable but decreased the model's performance for some reason. It is possible that our PWAT subscore classification is a fine-grained image classification problem and required the images to show high resolution, detailed features of the wounds while the wound features in images augmented from GAN methods are not clear enough. Therefore, we decided not to use GAN based data augmentation method in developing our wound deep learning architecture. Fig. 4c shows some examples of the synthesis chronic wound images.

E. Confusion matrices

Table IVb and Table IVc show a sample of the confusion matrices of the test set results from SS-PMG-EfficientNet for all 8 PWAT sub-scores. Due to space limitations, although there were 3 different results from 3 set of training and testing set, only confusion matrices of the best results are shown here.

The numbers on the diagonal position represent images classified correctly in the confusion matrices. It can be observed that the majority of test images are on the diagonals of the confusion matrices and the misclassified images are mainly distributed beside the diagonal position. The confusion matrices and the small difference between accuracy and F1 scores in our results show that the imbalanced data does not significantly affect our models' performance.

IV. DISCUSSION

The proposed PMG-SS-EfficientNet effectively augmented our small labeled wound dataset and predicted all 8 PWAT sub-scores with clinically usable accuracy In general, improving the machine learning model and providing it with adequate data, especially when the original dataset is small, are two important approaches to improving a machine learning model's performance. Most research on fine-grained image classification problems focused on proposing novel and sophisticated deep learning architectures [58] [59]. These architectures typically had large capacities and performed well when trained and evaluated on large datasets. However, since our *WoundNet* dataset is relatively small, deep learning architectures with large capacity would overfit on our dataset. The original PMG mechanism integrated into ResNet50 was a relatively simple but effective design that did not add too much capacity to the baseline ResNet50 model. Therefore, it enables EfficientNet B0 to better learn fine-grained features in wound images and improve its performance by integrating the PMG mechanism, which also keeps the architecture's capacity relatively small.

State-of-the-art data augmentation methods including GANs did not work well on the PWAT wound classification problem An alternate approach to improve the model's performance is to provide it with more data through data augmentation. However, traditional data augmentation techniques did not significantly improve the model in our chronic wound scoring problem, which prior research has demonstrated [15]. This paper utilized semi-supervised learning with DFUC 2021 [33] [47] as large, secondary dataset and our own *WoundNet* dataset as reference, to facilitate another form of data augmentation and improve the model's performance. This method allowed our model to learn chronic wound features from both our own dataset and the DFUC 2021 dataset. On the other hand, to train the PMG mechanism model to learn the corresponding granularity level's information for different stages, the jigsaw puzzle generator generates different images with different granularity regions from the original dataset image. The original image and the generated images were all input into the PMG mechanism model for training, which could be considered as another form of data augmentation. These two data augmentation methods both helped to improve the model's performance and showed significant improvement after integrating together.

As mentioned in Sub-section III-D, the data augmentation using GANs method, including pix2pixHD and semi-supervised GAN, decreased our model's performance. Although data augmentation using GANs methods can increase model's performance in other medical imaging research [36] [40] [41], it is proved to be unhelpful in our chronic wound scoring problem after multiple times of examining the method and testing the model. In some computer vision tasks, it is possible that models can still benefit from large amount of synthetic images even when they are of low quality with rough shapes and unclear features. However, our chronic wound scoring problem is a fine-grained image classification problem, which is difficult even for human eyes to distinguish the detail. To improve the model's performance on this problem, it requires the wound images to have clear detail and wound features so that the model is able to classify the images based on this subtle information. It can be observed in Fig. 4c that although the synthetic wound images have good quality, their details and wound features, such as textures and colors, are not as good as the original high-resolution wound images.

V. CONCLUSION

Due to challenges with collecting and labeling adequate image data on all wound severities, existing wound datasets frequently are imbalanced and relatively small, which limits the accuracy of deep learning-based wound grading models. The goal of this paper was to augment a small, imbalanced, wound dataset by using semi-supervised learning with a secondary dataset. The augmented wound dataset was then utilized for deep learning-based wound assessment. The primary, labeled wound dataset utilized in the semi-supervised approach was labeled with ground-truth wound assessments based on the comprehensive, clinically-valid wound grading rubric called

PWAT. We proposed a Semi-Supervised PMG EfficientNet deep learning architecture, which estimated all 8 PWAT sub-scores. We applied transfer learning to SS-PMG-EfficientNet model to learn each of the 8 PWAT subscores separately. In rigorous evaluation, the proposed Semi-Supervised PMG EfficientNet architecture performed well on assessing chronic wounds including diabetic foot ulcers, pressure ulcers, vascular ulcers and surgical wounds. Our proposed Semi-Supervised PMG EfficientNet (SS-PMG-EfficientNet) approach estimated all 8 PWAT sub-scores with classification accuracies and F1 scores of about 90% on average, and outperformed a comprehensive list of baseline models and had a 7% improvement over the prior state-of-the-art (without data augmentation). We also demonstrate that synthetic wound image generation using Generative Adversarial Networks (GANs) did not improve wound assessment.

In future work, we plan to systematically investigate the reason why data augmentation using GANs-generated images does not improve our wound assessment model's performance. We will also explore methods to further improve our proposed model's performance for predicting PWAT sub-scores that it did not perform well on, such as subscore 3. Necrotic Type. We will also investigate whether the focal loss can further improve our model's performance on imbalanced data.

ACKNOWLEDGMENT

This project is funded by the National Institutes of Health (NIH) under grant number (1R01EB025801).

REFERENCES

[1] K. Järbrink, G. Ni, H. Sönnergren, A. Schmidtchen, C. Pang, R. Bajpai, and J. Car, "The humanistic and economic burden of chronic wounds: a protocol for a systematic review," *Systematic reviews*, vol. 6, no. 1, pp. 1–7, 2017.

[2] S. R. Nussbaum, M. J. Carter, C. E. Fife, J. DaVanzo, R. Haught, M. Nusgart, and D. Cartwright, "An economic evaluation of the impact, cost, and medicare policy implications of chronic nonhealing wounds," *Value in Health*, vol. 21, no. 1, pp. 27–32, 2018.

[3] L. Gould, P. Abadir, H. Brem, M. Carter, T. Conner-Kerr, J. Davidson, L. DiPietro, V. Falanga, C. Fife, S. Gardner *et al.*, "Chronic wound repair and healing in older adults: current status and future research," *Wound Repair and Regeneration*, vol. 23, no. 1, pp. 1–13, 2015.

[4] C. K. Sen, G. M. Gordillo, S. Roy, R. Kirsner, L. Lambert, T. K. Hunt, F. Gottrup, G. C. Gurtner, and M. T. Longaker, "Human skin wounds: a major and snowballing threat to public health and the economy," *Wound repair and regeneration*, vol. 17, no. 6, pp. 763–771, 2009.

[5] M. A. Fonder, G. S. Lazarus, D. A. Cowan, B. Aronson-Cook, A. R. Kohli, and A. J. Mamelak, "Treating the chronic wound: A practical approach to the care of nonhealing wounds and wound care dressings," *Journal of the American Academy of Dermatology*, vol. 58, no. 2, pp. 185–206, 2008.

[6] S. J. Landis, "Chronic wound infection and antimicrobial use," *Advances in skin & wound care*, vol. 21, no. 11, pp. 531–540, 2008.

[7] H. Nejati, V. Pomponiu, T.-T. Do, Y. Zhou, S. Iravani, and N.-M. Cheung, "Smartphone and mobile image processing for assisted living: health-monitoring apps powered by advanced mobile imaging algorithms," *IEEE Signal Processing Magazine*, vol. 33, no. 4, pp. 30–48, 2016.

[8] E. Sirazitdinova and T. M. Deserno, "System design for 3d wound imaging using low-cost mobile devices," in *Medical Imaging 2017: Imaging Informatics for Healthcare, Research, and Applications*, vol. 10138. International Society for Optics and Photonics, 2017, p. 1013810.

[9] V. N. Shenoy, E. Foster, L. Aalami, B. Majeed, and O. Aalami, "Deepwound: Automated postoperative wound assessment and surgical site surveillance through convolutional neural networks," in *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2018, pp. 1017–1021.

[10] R. Niri, Y. Lucas, S. Treuillet, and H. Douzi, "Smartphone-based thermal imaging system for diabetic foot ulcer assessment," 2019.

[11] L. Wang, P. C. Pedersen, D. Strong, B. Tulu, and E. Agu, "Wound image analysis system for diabetics," in *Medical Imaging 2013: Image Processing*, vol. 8669. International Society for Optics and Photonics, 2013, p. 866924.

[12] L. Wang, P. C. Pedersen, E. Agu, D. M. Strong, and B. Tulu, "Area determination of diabetic foot ulcer images using a cascaded two-stage svm-based classification," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 9, pp. 2098–2109, 2016.

[13] L. Wang, "System designs for diabetic foot ulcer image assessment," 2016.

[14] L. Wang, P. C. Pedersen, E. Agu, D. M. Strong, and B. Tulu, "Boundary determination of foot ulcer images by applying the associative hierarchical random field framework," *Journal of Medical Imaging*, vol. 6, no. 2, p. 024002, 2019.

[15] Z. Liu, E. Agu, P. Pedersen, C. Lindsay, B. Tulu, and D. Strong, "Comprehensive assessment of fine-grained wound images using a patch-based cnn with context-preserving attention," *IEEE open journal of engineering in medicine and biology*, vol. 2, pp. 224–234, 2021.

[16] G. FrykbergRobert *et al.*, "Challenges in the treatment of chronic wounds," *Advances in wound care*, 2015.

[17] P. E. Houghton, C. B. Kincaid, K. E. Campbell, M. Woodbury, and D. Keast, "Photographic assessment of the appearance of chronic pressure and leg ulcers," *Ostomy Wound Management*, vol. 46, no. 4, pp. 20–35, 2000.

[18] P. E. Houghton, C. B. Kincaid, M. Lovell, K. E. Campbell, D. H. Keast, M. G. Woodbury, and K. A. Harris, "Effect of electrical stimulation on chronic leg ulcer size and appearance," *Physical therapy*, vol. 83, no. 1, pp. 17–28, 2003.

[19] N. Thompson, L. Gordey, H. Bowles, N. Parslow, and P. Houghton, "Reliability and validity of the revised photographic wound assessment tool on digital images taken of various types of chronic wounds," *Advances in skin & wound care*, vol. 26, no. 8, pp. 360–373, 2013.

[20] D. Y. Chino, L. C. Scabora, M. T. Cazzolato, A. E. Jorge, C. Traina-Jr, and A. J. Traina, "Segmenting skin ulcers and measuring the wound area using deep convolutional networks," *Computer Methods and Programs in Biomedicine*, vol. 191, p. 105376, 2020.

[21] D. Spinczyk and M. Wideł, "Surface area estimation for application of wound care," *Injury*, vol. 48, no. 3, pp. 653–658, 2017.

[22] J.-T. Hsu, T.-W. Ho, H.-F. Shih, C.-C. Chang, F. Lai, and J.-M. Wu, "Automatic wound infection interpretation for postoperative wound image," in *Eighth International Conference on Graphic and Image Processing (ICGIP 2016)*, vol. 10225. International Society for Optics and Photonics, 2017, p. 1022526.

[23] G. Blanco, A. J. Traina, C. Traina Jr, P. M. Azevedo-Marques, A. E. Jorge, D. de Oliveira, and M. V. Bedo, "A superpixel-driven deep learning approach for the analysis of dermatological wounds," *Computer methods and programs in biomedicine*, vol. 183, p. 105079, 2020.

[24] V. Godeiro, J. S. Neto, B. Carvalho, B. Santana, J. Ferraz, and R. Gama, "Chronic wound tissue classification using convolutional networks and color space reduction," in *2018 IEEE 28th International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 2018, pp. 1–6.

[25] H. Nejati, H. A. Ghazijahani, M. Abdollahzadeh, T. Malekzadeh, N.-M. Cheung, K.-H. Lee, and L.-L. Low, "Fine-grained wound tissue analysis using deep neural network," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 1010–1014.

[26] J.-T. Hsu, Y.-W. Chen, T.-W. Ho, H.-C. Tai, J.-M. Wu, H.-Y. Sun, C.-S. Hung, Y.-C. Zeng, S.-Y. Kuo, and F. Lai, "Chronic wound assessment and infection detection method," *BMC medical informatics and decision making*, vol. 19, no. 1, pp. 1–20, 2019.

[27] M. Maity, D. Dhane, C. Bar, C. Chakraborty, and J. Chatterjee, "Pixel-based supervised tissue classification of chronic wound images with deep autoencoder," in *Advanced Computational and Communication Paradigms*. Springer, 2018, pp. 727–735.

[28] K. Babu, S. Sabut, and D. Nithya, "Efficient detection and classification of diabetic foot ulcer tissue using pso technique," 2018.

[29] V. Rajathi, R. Bhavani, and G. Wiselin Jiji, "Varicose ulcer (c6) wound image tissue classification using multidimensional convolutional neural networks," *The Imaging Science Journal*, vol. 67, no. 7, pp. 374–384, 2019.

[30] L. Bloch, R. Brüngel, and C. M. Friedrich, "Boosting efficientnets ensemble performance via pseudo-labels and synthetic images by pix2pixhd for infection and ischaemia classification in diabetic foot ulcers," in *Diabetic Foot Ulcers Grand Challenge*. Springer, 2021, pp. 30–49.

[31] S. K. Das, P. Roy, and A. K. Mishra, "Dfu_spnet: A stacked parallel convolution layers based cnn to improve diabetic foot ulcer classification," *ICT Express*, vol. 8, no. 2, pp. 271–275, 2022.

[32] M. Goyal, N. D. Reeves, S. Rajbhandari, N. Ahmad, C. Wang, and M. H. Yap, "Recognition of ischaemia and infection in diabetic foot ulcers: Dataset and techniques," *Computers in biology and medicine*, vol. 117, p. 103616, 2020.

[33] M. H. Yap, B. Cassidy, J. M. Pappachan, C. O'Shea, D. Gillespie, and N. D. Reeves, "Analysis towards classification of infection and ischaemia of diabetic foot ulcers," in *EMBS BHI*. IEEE, 2021, pp. 1–4.

[34] N. Al-Garaawi, R. Ebsim, A. F. Alharan, and M. H. Yap, "Diabetic foot ulcer classification using mapped binary patterns and convolutional neural networks," *Computers in biology and medicine*, vol. 140, p. 105055, 2022.

[35] M. Goyal, N. D. Reeves, A. K. Davison, S. Rajbhandari, J. Spragg, and M. H. Yap, "Dfunet: Convolutional neural networks for diabetic foot ulcer classification," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 4, no. 5, pp. 728–739, 2018.

[36] M. Frid-Adar, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, "Synthetic data augmentation using gan for improved liver lesion classification," in *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*. IEEE, 2018, pp. 289–293.

[37] A. Zhao, G. Balakrishnan, F. Durand, J. V. Guttag, and A. V. Dalca, "Data augmentation using learned transformations for one-shot medical image segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 8543–8553.

[38] A. Ghorbani, V. Natarajan, D. Coz, and Y. Liu, "Dermgan: Synthetic generation of clinical skin images with pathology," in *Machine Learning for Health Workshop*. PMLR, 2020, pp. 155–170.

[39] F. Pollastri, F. Bolelli, R. Paredes, and C. Grana, "Augmenting data with gans to segment melanoma skin lesions," *Multimedia Tools and Applications*, vol. 79, no. 21, pp. 15 575–15 592, 2020.

[40] T. Pang, J. H. D. Wong, W. L. Ng, and C. S. Chan, "Semi-supervised gan-based radiomics model for data augmentation in breast ultrasound mass classification," *Computer Methods and Programs in Biomedicine*, vol. 203, p. 106018, 2021.

[41] Q. Guan, Y. Chen, Z. Wei, A. A. Heidari, H. Hu, X.-H. Yang, J. Zheng, Q. Zhou, H. Chen, and F. Chen, "Medical image augmentation for lesion detection using a texture-constrained multichannel progressive gan," *Computers in Biology and Medicine*, vol. 145, p. 105444, 2022.

[42] N. Komodakis and S. Gidaris, "Unsupervised representation learning by predicting image rotations," in *International Conference on Learning Representations (ICLR)*, 2018.

[43] P. V. Tran, "Exploring self-supervised regularization for supervised and semi-supervised learning," *arXiv preprint arXiv:1906.10343*, 2019.

[44] K. Sohn, D. Berthelot, N. Carlini, Z. Zhang, H. Zhang, C. A. Raffel, E. D. Cubuk, A. Kurakin, and C.-L. Li, "Fixmatch: Simplifying semi-supervised learning with consistency and confidence," *Advances in neural information processing systems*, vol. 33, pp. 596–608, 2020.

[45] R. Du, D. Chang, A. K. Bhunia, J. Xie, Z. Ma, Y.-Z. Song, and J. Guo, "Fine-grained visual classification via progressive multi-granularity training of jigsaw patches," in *European Conference on Computer Vision*. Springer, 2020, pp. 153–168.

[46] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*. PMLR, 2019, pp. 6105–6114.

[47] M. H. Yap, C. Kendrick, N. D. Reeves, M. Goyal, J. M. Pappachan, and B. Cassidy, "Development of diabetic foot ulcer datasets: An overview," *Diabetic Foot Ulcers Grand Challenge*, pp. 1–18, 2021.

[48] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional gans," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8798–8807.

[49] C. Li, T. Xu, J. Zhu, and B. Zhang, "Triple generative adversarial nets," *Advances in neural information processing systems*, vol. 30, 2017.

[50] F. Zhang, M. Li, G. Zhai, and Y. Liu, "Multi-branch and multi-scale attention learning for fine-grained visual categorization," in *International Conference on Multimedia Modeling*. Springer, 2021, pp. 136–147.

[51] A. Wagh, "Semantic segmentation of smartphone wound images: Comparative analysis of ahrf and cnn-based approaches wpi technical report wpi-cs-tr-19-02."

[52] K.-K. Maninis, S. Caelles, J. Pont-Tuset, and L. Van Gool, "Deep extreme cut: From extreme points to object segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 616–625.

[53] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[54] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*. PMLR, 2015, pp. 448–456.

[55] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (elus)," *arXiv preprint arXiv:1511.07289*, 2015.

[56] A. Oliver, A. Odena, C. A. Raffel, E. D. Cubuk, and I. Goodfellow, "Realistic evaluation of deep semi-supervised learning algorithms," *Advances in neural information processing systems*, vol. 31, 2018.

[57] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in neural information processing systems*, vol. 30, 2017.

[58] H. Zheng, J. Fu, Z.-J. Zha, and J. Luo, "Looking for the devil in the details: Learning trilinear attention sampling network for fine-grained image recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5012–5021.

[59] P. Shroff, T. Chen, Y. Wei, and Z. Wang, "Focus longer to see better: Recursively refined attention for fine-grained image classification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 868–869.