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 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).
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.

**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>
<thead>
<tr>
<th>Wound Types</th>
<th>Numbers of Images</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Diabetic Foot Ulcers</td>
<td>121</td>
</tr>
<tr>
<td>2. Pressure Ulcers</td>
<td>13</td>
</tr>
<tr>
<td>3. Vascular Ulcers</td>
<td>1349</td>
</tr>
<tr>
<td>4. Surgical Wounds</td>
<td>156</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>PWAT sub-score</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Size</td>
<td>143</td>
<td>393</td>
<td>537</td>
<td>337</td>
<td>231</td>
</tr>
<tr>
<td>2. Depth</td>
<td>134</td>
<td>135</td>
<td>814</td>
<td>378</td>
<td>178</td>
</tr>
<tr>
<td>3. Necrotic Type</td>
<td>509</td>
<td>472</td>
<td>273</td>
<td>76</td>
<td>309</td>
</tr>
<tr>
<td>4. Necrotic Amount</td>
<td>347</td>
<td>160</td>
<td>150</td>
<td>290</td>
<td>692</td>
</tr>
<tr>
<td>5. Granulation Type</td>
<td>143</td>
<td>319</td>
<td>626</td>
<td>95</td>
<td>456</td>
</tr>
<tr>
<td>6. Granulation Amount</td>
<td>137</td>
<td>77</td>
<td>194</td>
<td>342</td>
<td>889</td>
</tr>
<tr>
<td>7. Edges</td>
<td>142</td>
<td>311</td>
<td>1035</td>
<td>131</td>
<td>20</td>
</tr>
<tr>
<td>8. Skin</td>
<td>435</td>
<td>1061</td>
<td>143</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

**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, but achieved only about 82% in accuracy and F1-score.**
### Medical imaging research assessing PWAT wound attributes

<table>
<thead>
<tr>
<th>Related research</th>
<th>PWAT subscore</th>
<th>machine learning method</th>
<th>task</th>
<th>wound type</th>
<th>dataset size</th>
<th>results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zhao et al. 2019 [37]</td>
<td>3. Nec Type</td>
<td>deep neural network</td>
<td>segment the wound; estimate the size</td>
<td>venous and arterial ulcer</td>
<td>446</td>
<td>estimate wound area in cm² with error of 14%</td>
</tr>
<tr>
<td>Hsu et al. 2017 [22]</td>
<td>6. Gran Amount</td>
<td>clustering method, SVM</td>
<td>detect necrotic tissue</td>
<td>postsurgical wound</td>
<td>42</td>
<td>detection accuracy 95.23%</td>
</tr>
<tr>
<td>Blanco et al. 2020 [23]</td>
<td>Liver lesion CT images</td>
<td>superpixel-driven deep learning approach</td>
<td>segmenting Necrotic and Granulation, and wounded area</td>
<td>arterial and venous ulcers</td>
<td>217</td>
<td>spot wounded tissues with AUC = 0.986</td>
</tr>
<tr>
<td>Nejati et al. 2018 [25]</td>
<td>Deep neural network, SVM</td>
<td>classifying Necrotic and Granulation</td>
<td>chronic wounds</td>
<td>350</td>
<td>tissue classification accuracy 86.4%</td>
<td></td>
</tr>
<tr>
<td>Hsu et al. 2019 [26]</td>
<td>Deep neural network</td>
<td>robust image segmentation, SVM</td>
<td>wound segmentation; detect Necrotic and Granulation</td>
<td>chronic wounds</td>
<td>293</td>
<td>tissue classification accuracy 83.58%</td>
</tr>
<tr>
<td>Matly et al. 2018 [27]</td>
<td>Deep neural network</td>
<td>classifying Necrotic and Granulation</td>
<td>chronic wounds</td>
<td>68</td>
<td>tissue classification accuracy 99%</td>
<td></td>
</tr>
<tr>
<td>Babu et al. 2018 [28]</td>
<td>Naïve bayes and Hoeffding tree</td>
<td>wound segmentation; classifying Necrotic and Granulation</td>
<td>diabetic wound</td>
<td>N. A.</td>
<td>tissue classification accuracy 90.9%</td>
<td></td>
</tr>
</tbody>
</table>

### Wound attributes (all Diabetic Foot Ulcer) medical imaging that explored data augmentation

<table>
<thead>
<tr>
<th>Related research</th>
<th>data augmentation method</th>
<th>machine learning method</th>
<th>task</th>
<th>dataset size</th>
<th>results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bloch et al. 2021 [30]</td>
<td>pix2pixHD</td>
<td>EfficientNets Ensemble</td>
<td>4 classes (Infection and Ischaemia, Infection, Ischaemia, None)</td>
<td>Ulcer foot, augmented to 4341</td>
<td>best macro F1-Score of 60.77%</td>
</tr>
<tr>
<td>Das et al. 2022 [31]</td>
<td>horizontal and vertical flips</td>
<td>CNN-based classification model</td>
<td>DFU vs. normal skin binary classification</td>
<td>292</td>
<td>96.4% accuracy, 95.4% F1 score</td>
</tr>
<tr>
<td>Goyal et al. 2020 [32]</td>
<td>Natural data augmentation</td>
<td>ensemble CNN</td>
<td>2 classes (Ischaemia: Yes, No; Infection: Yes, No) classification</td>
<td>Ischaemia: (235, 1431) augmented; (4935, 4935); Infection: (982, 684) augmented; (2946, 2946)</td>
<td>90% accuracy in ischaemia, 73% accuracy in infection</td>
</tr>
<tr>
<td>Yap et al. 2021 [33]</td>
<td>flip, natural data augmentation, crop, Gaussian noise, rotate shear, scale and adjust contrast</td>
<td>VGG16, ResNet101, InceptionV3, DenseNet121, EfficientNet</td>
<td>4 classes (both Infection and Ischaemia, Infection, Ischaemia, None)</td>
<td>Infection and Ischaemia, None: (2946, 2946)</td>
<td>EfficientNet B0: macro-average Precision, Recall and F1-Score of 0.87, 0.82 and 0.75</td>
</tr>
<tr>
<td>Al-Garaawi et al. 2022 [34]</td>
<td>rotation, flipping, color space augmentation</td>
<td>CNN-based DFU classification method</td>
<td>Part A: 2 classes (healthy and DFU); Part B: 2 classes (Ischaemia: Yes, No; Infection: Yes, No) classification</td>
<td>Part A: 641,1038; Part B: Ischaemia: (4935, 4935); Infection: (2946, 2946)</td>
<td>Ischaemia: 0.995% (AUC), 0.990% (F-measure); Infection: 0.820% (AUC), 0.744% (F-measure)</td>
</tr>
<tr>
<td>Goyal et al. 2018 [35]</td>
<td>rotation, contrast enhancement, color space, random scaling, flipping</td>
<td>a novel CNN model</td>
<td>2 classes (healthy skin and DFU)</td>
<td>641 healthy, 1038 DFU</td>
<td>accuracy 0.925, F-measure 0.939, AUC 0.962</td>
</tr>
</tbody>
</table>

### Non-wounded medical imaging tasks that explored data augmentation

<table>
<thead>
<tr>
<th>Related research</th>
<th>data augmentation method</th>
<th>machine learning method</th>
<th>task</th>
<th>dataset size</th>
<th>results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frid-Adar et al. 2018 [36]</td>
<td>Deep Convolutional GAN (DCGAN)</td>
<td>CNN model</td>
<td>liver lesion CT images classification</td>
<td>liver lesions (CT) images: 182, augmented: 5000</td>
<td>improvement of 7% using GAN with accuracy of 85.7%</td>
</tr>
<tr>
<td>Zhao et al. 2019 [37]</td>
<td>learned spatial and appearance transform models</td>
<td>deep fully convolutional neural networks</td>
<td>one-shot segmentation of brain (MRI) scans</td>
<td>101 brain scan, synthesis of 10,000 different labeled examples</td>
<td>Dice score of 0.815 (0.123)</td>
</tr>
<tr>
<td>Ghorbani et al. 2020 [38]</td>
<td>pix2pix (GAN) based</td>
<td>MobileNet CNN model</td>
<td>skin lesions classification</td>
<td>49920 images, 20000 synthetic images</td>
<td>about 50% accuracy, 14% improvement in F1 score in classes with fewer examples</td>
</tr>
<tr>
<td>Pang et al. 2021 [40]</td>
<td>semi-supervised GAN: TripleGAN</td>
<td>Inception-V3 model</td>
<td>Breast Ultrasound Mass Classification</td>
<td>1447 ultrasound images, augmented to 4341</td>
<td>accuracy 90.41%, sensitivity 87.94%, specificity 85.86%</td>
</tr>
<tr>
<td>Guan et al. 2022 [41]</td>
<td>texture-constrained multichannel progressive generative adversarial network (TMP-GAN)</td>
<td>faster-RCNN</td>
<td>lesion detection of mammography data set and pancreatic tumors</td>
<td>CBIS-DDMS: 1318 pancreatic tumor dataset: 1066; synthetic images: 1, 4, and 9 times the actual images</td>
<td>precision, recall, F1-score CBIS-DDMS: improves 2.5%/2.7%/2.77%, to 86.28%/85.89%/86.08% pancreatic tumor: improves 2.4%/2.06%/2.36%, to 86.28%/85.89%/86.08%</td>
</tr>
</tbody>
</table>

**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 sub-scores.

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
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 $1600 \times 1200$ and $3648 \times 2736$ and they were resized to a dimension of $640 \times 640$, 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 PW AT 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 PW AT 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 PMG mechanism and semi-supervised learning 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:

$$V^l = H^l_{conv}(F^l)$$  \hspace{1cm} (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:

$$y^l = H^l_{class}(V^l)$$  \hspace{1cm} (2)

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

$$V^{concat} = concat(V^{L-S+1}, ..., V^{L-1}, V^L)$$  \hspace{1cm} (3)

It is then input into a classifier:

$$y^{concat} = H^{concat}_{class}(V^{concat})$$  \hspace{1cm} (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

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

and

$$L_{CE}(y^{concat}, y) = -\sum_{i=1}^{m} y_i^{concat} \times \log(y_i^{concat})$$  \hspace{1cm} (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:

$$n = 2^{L-l+1}$$  \hspace{1cm} (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:

$$C_1 = \arg \max (y^{concat})$$  \hspace{1cm} (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 $C_2$ is:

$$C_2 = \arg \max (\sum_{l=L-S+1}^{L} y^l + y^{concat})$$  \hspace{1cm} (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 \( \theta \) 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:

\[
s_L = b(x_i \in S_L) \quad (10)
\]

\[
s_U = b(x_j \in S_U) \quad (11)
\]

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:

\[
z_i = f_\theta(s_L) \quad (12)
\]

\[
z_j = f_\theta(s_U) \quad (13)
\]

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

\[
L_{\text{labeled}}(y_i, z_i) = -\frac{1}{|S_L|} \sum_{i \in S_L} \sum_{k \in K} y_{ik} \log(z_{ik}) \quad (14)
\]

\[
L_{\text{unlabeled}}(y_j, z_j) = -\frac{1}{|S_U|} \sum_{j \in S_U} \sum_{i \in T} y_{ji} \log(z_{ji}) \quad (15)
\]

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

\[
L_{\text{final}} = L_{\text{labeled}}(y_i, z_i) + \omega L_{\text{unlabeled}}(y_j, z_j) \quad (16)
\]

The parameter \( \theta \) will be updated in backpropagation after minimizing the final loss function \( L_{\text{final}} \). The unsupervised cross-entropy loss \( L_{\text{unlabeled}}(y_j, z_j) \) 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 \( \omega \) 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 \( \omega \) between 0.5 to 1.2, which revealed that the model performed best when \( \omega \) 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 sub-scores 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 sub-score 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 sub-scores.

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 sub-score 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 sub-scores 6 times with 3 different EfficientNet B0 based architectures.
3) Experiment 2, comparison of various deep learning architectures and data augmentation methods on PWAT sub-score 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 sub-score. 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 sub-score 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 sub-score 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 sub-score 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 sub-score 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 sub-scores, it is time-consuming to run these many experiments. On the other hand, one PWAT sub-score can indicate the general performance of different architectures on the WoundNet dataset. Therefore, testing different architectures and methods on the most important sub-score 2. Depth enabled discovery of the best architectures and methods for estimating PWAT sub-scores 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 \( i \)th PWAT sub-score in these two figures. For example, train1 is the training accuracy for the sub-score 1. Size. The training and testing accuracies converged and stabilized after about 35 to 45 epochs. The differences between

(c) Examples of synthesis wound images from GANs method

Fig. 4. Training and testing accuracy trajectory and synthesis wound image examples from GANs method
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 sub-scores 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 sub-scores 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 sub-scores 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 sub-scores 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 sub-scores compared to other 3 deep learning architectures. The standard deviation of the testing accuracy, weighted F1 scores, Sensitivity and Specificity for all 8 PWAT sub-scores 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 sub-scores’ 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 sub-scores 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

<table>
<thead>
<tr>
<th>PWAT subscore</th>
<th>Semi-Supervised PMG EfficientNet</th>
<th>PMG EfficientNet</th>
<th>Semi-Supervised EfficientNet</th>
<th>EfficientNet</th>
<th>Patch Attention DenseNet</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Size</td>
<td>0.8753 ± 0.0071</td>
<td>0.8609 ± 0.0177</td>
<td>0.8384 ± 0.0163</td>
<td>0.8068 ± 0.0146</td>
<td>0.8098 ± 0.0132</td>
</tr>
<tr>
<td>2. Depth</td>
<td>0.9039 ± 0.0232</td>
<td>0.8916 ± 0.0071</td>
<td>0.8651 ± 0.0103</td>
<td>0.8267 ± 0.0237</td>
<td>0.8281 ± 0.0185</td>
</tr>
<tr>
<td>3. Nec Type</td>
<td>0.8793 ± 0.0035</td>
<td>0.8753 ± 0.0071</td>
<td>0.8569 ± 0.0158</td>
<td>0.8114 ± 0.0176</td>
<td>0.8244 ± 0.0090</td>
</tr>
<tr>
<td>4. Nec Amount</td>
<td>0.8875 ± 0.0187</td>
<td>0.8569 ± 0.0035</td>
<td>0.8517 ± 0.0180</td>
<td>0.8114 ± 0.0183</td>
<td>0.8098 ± 0.0309</td>
</tr>
<tr>
<td>5. Gran Type</td>
<td>0.8916 ± 0.0094</td>
<td>0.8793 ± 0.0071</td>
<td>0.8722 ± 0.0264</td>
<td>0.8129 ± 0.0097</td>
<td>0.8269 ± 0.0153</td>
</tr>
<tr>
<td>6. Gran Amount</td>
<td>0.8998 ± 0.0248</td>
<td>0.8896 ± 0.0123</td>
<td>0.8671 ± 0.0200</td>
<td>0.8466 ± 0.0371</td>
<td>0.8330 ± 0.0235</td>
</tr>
<tr>
<td>7. Edges</td>
<td>0.9018 ± 0.0123</td>
<td>0.8834 ± 0.0061</td>
<td>0.8731 ± 0.0205</td>
<td>0.8497 ± 0.0146</td>
<td>0.8256 ± 0.0093</td>
</tr>
<tr>
<td>8. Skin</td>
<td>0.9018 ± 0.0061</td>
<td>0.8916 ± 0.0128</td>
<td>0.8497 ± 0.0075</td>
<td>0.8328 ± 0.0153</td>
<td>0.8208 ± 0.0222</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>PWAT subscore</th>
<th>F1 score</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Size</td>
<td>0.8755 ± 0.0067</td>
<td>0.8674 ± 0.0065</td>
<td>0.9673 ± 0.0022</td>
</tr>
<tr>
<td>2. Depth</td>
<td>0.9017 ± 0.0238</td>
<td>0.8656 ± 0.0250</td>
<td>0.9666 ± 0.0057</td>
</tr>
<tr>
<td>3. Nec Type</td>
<td>0.8767 ± 0.0052</td>
<td>0.8047 ± 0.0406</td>
<td>0.9679 ± 0.0006</td>
</tr>
<tr>
<td>4. Nec Amount</td>
<td>0.8862 ± 0.0270</td>
<td>0.8539 ± 0.0253</td>
<td>0.9679 ± 0.0063</td>
</tr>
<tr>
<td>5. Gran Type</td>
<td>0.8915 ± 0.0197</td>
<td>0.8789 ± 0.0139</td>
<td>0.9695 ± 0.0003</td>
</tr>
<tr>
<td>6. Gran Amount</td>
<td>0.8998 ± 0.0247</td>
<td>0.8662 ± 0.0414</td>
<td>0.9765 ± 0.0008</td>
</tr>
<tr>
<td>7. Edges</td>
<td>0.9006 ± 0.0088</td>
<td>0.8838 ± 0.0352</td>
<td>0.9607 ± 0.0054</td>
</tr>
<tr>
<td>8. Skin</td>
<td>0.9013 ± 0.0052</td>
<td>0.8542 ± 0.0447</td>
<td>0.9302 ± 0.0056</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>PWAT subscore</th>
<th>F1 score</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Size</td>
<td>0.8607 ± 0.0172</td>
<td>0.8593 ± 0.0264</td>
<td>0.9626 ± 0.0053</td>
</tr>
<tr>
<td>2. Depth</td>
<td>0.8887 ± 0.0084</td>
<td>0.8535 ± 0.0235</td>
<td>0.9643 ± 0.0028</td>
</tr>
<tr>
<td>3. Nec Type</td>
<td>0.8706 ± 0.0074</td>
<td>0.7998 ± 0.0096</td>
<td>0.9671 ± 0.0023</td>
</tr>
<tr>
<td>4. Nec Amount</td>
<td>0.8547 ± 0.0038</td>
<td>0.8400 ± 0.0361</td>
<td>0.9600 ± 0.0016</td>
</tr>
<tr>
<td>5. Gran Type</td>
<td>0.8795 ± 0.0078</td>
<td>0.8795 ± 0.0024</td>
<td>0.9664 ± 0.0021</td>
</tr>
<tr>
<td>6. Gran Amount</td>
<td>0.8889 ± 0.0111</td>
<td>0.8453 ± 0.0207</td>
<td>0.9639 ± 0.0054</td>
</tr>
<tr>
<td>7. Edges</td>
<td>0.8826 ± 0.0043</td>
<td>0.8506 ± 0.0133</td>
<td>0.9521 ± 0.0016</td>
</tr>
<tr>
<td>8. Skin</td>
<td>0.8905 ± 0.0124</td>
<td>0.8378 ± 0.0229</td>
<td>0.9269 ± 0.0089</td>
</tr>
</tbody>
</table>

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

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

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

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

**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 sub-score 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 sub-scores 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 sub-scores. Test accuracy and F1 Scores for all 8 PWAT sub-scores from these 3 models were high relatively and the difference between different sub-scores were small from all 3 models.

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

D. Comparison of variations of our proposed architecture and baseline methods with PWAT sub-score 2. Depth

The comparison of results from different deep learning architectures and data augmentation methods we used on the PWAT sub-score 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 min 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 sub-scores and had the mean of testing accuracy: 0.7915, which could be considered as the PWAT sub-score 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-
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 subscores. 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 subscores 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 sub-scores 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 subcore 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


